농학박사학위논문

Bioinformatic studies to identify human genomic features based on structural variants

구조 변이 기반 인간 게놈 특성 규명을 위한 생물정보학 연구

2014년 8월

서울대학교 대학원
농생명공학부 동물생명공학전공
김효영

Bioinformatic studies
to identify human genomic features
based on structural variants

By

HyoYoung Kim

Supervisor: Professor Heebal Kim, Ph. D.

August, 2014

Department of Agricultural Biotechnology

Seoul National University

구조 변이 기반 인간 게놈 특성 규명을 위한 생물정보학 연구

지도교수 김 희 발

이 논문을 농학박사 학위논문으로 제출함

2014 년 8 월

서울대학교 대학원

농생명공학부 동물생명공학전공

김 효 영

김효영의 농학박사 학위논문을 인준함

2014 년 8 월

위 원 장 이 창 규 (인)

부위원장 김 희 발 (인)

위 원 김 경 모 (인)

위 원 윤 숙 희 (인)

위 원 조 서 애 (인)

Abstract

Bioinformatic studies
to identify human genomic features
based on structural variants

HyoYoung Kim
Department of Agricultural Biotechnology
The Graduate School
Seoul National University

Over the past few years, efforts focused on investigating the effects of copy number variations (CNVs) in human disease have been continuing. Genetic differences are attributable in part to large-scale structural variations between individuals. CNV is a form of structural variation as a DNA segment ≥ 1 kb in size when compared to a reference genome. Therefore, CNV was used to identify what associated with susceptibility and resistance to diseases. Genome-wide association studies (GWAS) have been used to investigate novel candidate genes associated with complex traits. Many of studies have been reported the association between SNPs or CNVs and complex diseases. Also, several GWA studies have been applied to a personalized medicine. Data mining provided important insights into the data with complicated and huge quantity. These semantic networks have given researchers knowledgeable information answers to complex questions through integration of the available data. Therefore, this thesis is to identify the genetic variation associated with liver diseases between Koreans, construct biological networks to understand the semantic knowledge about liver functions or ethnic disparities, and develop the visualization tool to explain a biological meaning for CNVs or SNPs.

In chapter 1, the general background of CNV, GWAS, and biological network were summarized. First, for CNV, the general overview, mechanism sources, identification methods, various researches in human, and associations with complex diseases were presented. Second, for GWAS, the general overview, biological background, various methods, result findings, clinical application, and limitations were presented. Third, for biological network, the general overview and biological network systems were presented.

In chapter 2, two parts (KARE1 and KARE2) were constituted as replication studies of GWA (genome-wide association) for hepatic biochemical markers AST or ALT in Korean cohorts. In KARE1, the analysis of CNVs in 8,842 Koreans reveals thirty-nine genes associated with hepatic biochemical markers AST (aspartate aminotransferase) and/or ALT (alanine aminotransferase). I genotyped on Affymetrix Genome-Wide Human 5.0 arrays for all samples and identified 10,162 CNVs using HelixTree software (ver. 7.0). To explain the impact of CNVs on each quantitative trait (AST or ALT), univariate linear regression was performed. As the result, 100 CNVs were significant for AST and 16 were significant for ALT at the significance level of 5%. I identified thirty-nine genes located within the significant CNV regions.

According to the functional annotation by using DAVID tool, the CNV-based genes are likely to be associated with liver diseases. In KARE2, a study of GWA for hepatic biomarkers was investigated in 407 Korean cohorts. Affymetrix Genome-Wide Human 6.0 array was genotyped for all samples and CNVs were identified using HelixTree software. By using univariate linear regression, 32 and 42 CNVs showed significance for AST and ALT, respectively (p-value < 0.05). To replication study of GWA for hepatic biomarker, CNV-based genes between KARE1 (AST-1885, ALT-773) and KARE2 (AST-140, ALT-172) were compared using NetBox software. As a result, nine genes (*CIDEB, DFFA, PSMA3, PSMC5, PSMC6, PSMD12, PSMF1, SDC4*, and *SIAH1*) were overlapped for AST, yet no overlapping genes were found for ALT. Structural variation analysis of CNV-based genes is useful to understand the biological phenotypes or diseases.

In chapter 3, to identify knowledgeable biological meanings for complex big data, two biological networks were constructed on liver functions or ethnic disparities using BioXM software. These semantic networks contained entities (Gene, Disease, Pathway, Chemical, Drug, SNP, CNV, ClinicalTrials, GO, drug, and SomaticMutation) and relationships between two entities (Gene-GO, Gene-Pathway, Gene-Disease, Gene-Chemical, Gene-SNP, Gene-CNV, Gene-SomaticMutation, Pathway-Chemical, Pathway-Chemical, Pathway-Disease, Chemical-Drug, ClinicalTrials-Disease, and ClinicalTrials-Drug). The application of the semantic liver functions network using the KARE2 data are shown in three clusters, including four diseases, one pathway, and seven drugs.

Ethnic disparities network was constructed using the ethnic specific SNP-based genes. By eliminating the overlapped SNPs from HapMap samples, ethnic specific SNPs were identified and the SNP-based genes were mapped to the UCSC RefGene lists (ver. hg18). As a result, ethnic specific 22, 25, and 332 genes were identified in the CEU (USA), JPT (Japan), and YRI (Africa) individuals, respectively. The application of ethnic disparities network showed interesting results in the three categories, including three diseases, one drug, and five pathways. The majority of these findings were consistent with the previous studies that an understanding of genetic variability explained ethnic disparities.

In chapter 4, VCS (Visualization of CNVs or SNPs) tool was constructed to visualize CNVs or SNPs detected in animals such as mammals, vertebrates, insects, and worms. VCS can easily interpret a biological meaning from the numerical value of CNVs or SNPs. The VCS provides six visualization tools: (i) the enrichment of genome contents in CNV region; (ii) the physical distribution of CNV or SNP on chromosomes; (iii) the distribution of log2 ratio of CNVs with criteria of interested; (iv) the number distribution of CNVs or SNPs per binning unit (10 kb, 100 kb, 1Mb, and 10Mb); (v) the homozygosity distribution of SNP genotype on chromosomes; and (vi) cytomap of genes within CNVs or SNPs.

By GWAS analyzing between CNVs and hepatic biochemical markers AST or ALT, a lot of biological meaning associated with liver diseases in

Korean cohorts could be obtained. Also, semantic biological networks for liver functions or ethnic disparities could be obtained knowledgeable findings. Finally, VCS tool could be achieved by interpreting a biological meaning from the numerical value by graphical viewing, and offered more directly insertable tip-top figures in study. Therefore, in this thesis, I analyzed replication study of GWA for hepatic biomarkers AST or ALT (Chapter 2), constructed the semantic biological networks for liver functions or ethnic disparities (Chapter 3), and developed the VCS web-tool to visualize the CNVs or SNPs (Chapter 4).

Key words: Copy number variation, GWAS, Korean, liver, network, single nucleotide polymorphism.

Student number: 2007-30877

Contents

List of Tables

List of Figures

General Introduction

Genetic variations are shown by large-scale structural variants found between individuals. Single nucleotide polymorphisms (SNPs) or copy Number variations (CNVs) are DNA sequence variation compared to a reference genome. While SNP differs in a single nucleotide base, CNV differs about one more Kb in size (Wang et al. 2007b). Several studies reported that CNVs or SNPs are associated with phenotypic variations or genetic diseases (Freeman, et al., 2006; Eichler, et al., 2007). The comprehensive identification of the DNA sequence variations would useful the genetic and functional analysis of genome variations.

Liver is a vital reddish brown organ in human body. This organ has various functions such as detoxification, filtration of harmful substances, and biochemical production for digestion (http://www.mamashealth.com/). Liver has several roles in glycogen storage, red blood cells decomposition, and hormone production (Gitzelmann et al. 1996; Pocai et al. 2006; Zhang and Beynen 2007). Biochemical tests for liver function are usually used to diagnose patients with liver diseases. Aspartate aminotransferase (AST) or alanine aminotransferase (ALT) is used as the most important biochemical markers to detect liver injuries (http://labtestsonline.org/). The AST/ALT ratio is an indicator for evaluating liver damages. AST/ALT < 1.0 indicates moderate liver disease such as nonalcoholic fatty liver disease (NAFLD) and AST/ALT > 1.0

indicates a severe liver disease such as chronic hepatitis or alcoholic fatty liver disease (http://en.wikipedia.org/).

Tremendous efforts have been made to identify ethnic specific SNPs associated with human diseases (Delgado et al. 2002; Picornell et al. 2007). Ethnic disparities were caused by certain genetic, demographic, or socioeconomic factors. These ethnic disparities influence different outcomes in people with certain diseases (Gary et al. 2003). Therefore, genetic disparities cannot be ignored for its plays an important role in determining ethnic disparities.

Biological network is a semantic knowledgeable network into biological big data and help arrive at an adequate interpretation of integrated biological systems (Losko and Heumann 2009). Robust and flexible biological networks enable researchers to ask scientific questions and answers instead of constructing complex biological systems (Mukherjea et al. 2005). The combination of data integration and visualization could provide meaningful insights into heterogeneous data such as gene, chemical, disease, pathway, drug, SNP, or CNV (Shin et al. 2012).

Information of CNVs or SNPs consisting of numerical values is difficult to understand what the number means and how to interpret this value biologically (Popova et al. 2009). Visualization of the data is a graphic statistics and can help interpret biological meanings from the numerical value, even though it is not an additional step necessary for the analysis (Friendly and Denis 2008).

Chapter 1. Literature Review

1.1 Copy Number Variation (CNV)

1.1.1 Overview of CNVs

Copy Number Variations (CNV), a category of structural variation where the DNA of a genome changes, is a topic of interest in the field of genomics as it is a significant source of genetic and phenotypic variation in humans (Henrichsen et al. 2009). CNV includes various forms of DNA structural rearrangements such as deletions, insertions, duplication, inversion, and translocation. For example, the normal section on chromosome has A-B-C-D instead the variation section has A-B-C-C-D as a duplication "C" or A-B-D as a deletion "C" (http://en.wikipedia.org/). Approximately 12% of the human genome is composed of CNV's and their size can range from one Kb to several Mb when compared to a reference genome. (Stankiewicz and Lupski 2010). This form of distinct genetic difference can be used to identify factors associated with susceptibility and resistance to diseases.

1.1.2 Sources of CNVs

CNVs may either be familial inherited or caused by *de novo* copy number mutations. CNVs can be caused by DNA structural rearrangements. Lee et al. (2007) proposed that template switching is the cause of some structural variation (Lee et al. 2007). Low copy repeats (LCRs), segmental repeat sequences, are susceptible to DNA rearrangements. Several studies have reported that differences between copies influence the susceptibility of LCRs (Mills et al. 2011). CNVs influence gene expression, phenotypic variation, and adaptation by disrupting genes and altering gene dosage (Buckland 2003; Nguyen et al. 2006; Repping et al. 2006).

1.1.3 Identification methods of CNVs

CNV's can be identified using techniques such as fluorescent in situ hybridization (FISH), comparative genomic hybridization, array-based comparative genomic hybridization (array-CGH), and SNP genotyping platforms (Korbel et al. 2007). Advances in next-generation sequencing (NGS) technologies have enabled the fine scale discovery of CNVs (Korbel et al. 2007; Mills et al. 2011).

This is significant, as CNVs have been recognized to associate with susceptibility to phenotypic differences and specific diseases (Paudel et al. 2013). For example, in rapidly growing *Escherichia coli* cells, the gene copy number located near the origin region of DNA sequence replications is 4-fold higher than at the termination of DNA replication (Atkinson et al. 2003). Elevation of the copy number of the salivary amylase (*AMY1*) gene can improve the protein expression level in human genome. Higher *AMY1* protein levels can increase the digestion of high-starch foods and buffer against the negative effects on fitness of intestinal disease (Perry et al. 2007).

1.1.4 CNV researches in human

The first discovery of genomic variation among humans was made soon after the Human Genome Project. Sebat et al. (2004) showed that large-scale rearrangements such as copy number polymorphism (CNPs) contribute to the human genomic variations (Sebat et al. 2004). Representational oligonucleotide microarray analysis (ROMA) detected 221 copy number differences (the average length of 465 Kb) comprising 76 unique CNPs among 20 individuals. 70 genes within these CNPs were identified and several genes previously reported to be associated with human diseases (Sebat et al. 2004). Approximately 40% of the genome among unrelated humans typically differ with copy number (Kidd et al. 2008; Zhang et al. 2009). Kluck et al. (2006) observed d*e novo* CNVs between identical twins. The concordance rates for autism in monozygotic twins are 70% in contrast to 5% in dizygotic twins (Klauck 2006).

1.1.5 CNV roles in disease

CNVs have been reported associated with disease susceptibility or resistance. Variation in the dosage of individual genes can lead to profound phenotype differences (Chance et al. 1993). Gene copy number can lead to DNA rearrangements that support growth of cancer cells or cause neurological disorders such as learning disability, Parkinson (Polymeropoulos et al. 1996), Alzheimer (Theuns et al. 2006; Brouwers et al. 2011), Autism (Weiss et al. 2008;

Kumar et al. 2008), Schizophrenia (Stefansson et al. 2008; Stone et al. 2008), Pancreatitis (Sahin-Tóth 2006), and Glomerulonephritis (Iafrate et al. 2004). Deletion of the *COH1* gene causes recessive Cohen syndrome (Sebat et al. 2004). The epidermal growth factor (*EGFR*) gene showed overexpressed copy number in non-small cell lung cancer (NSCLC) (Cappuzzo et al. 2005). In addition, a higher copies of *CCL3L1* were associated with lower influence to HIV susceptibility (Gonzalez et al. 2005) and a low *FCGR3B* copy number was increased susceptibility to systemic lupus erythematosus (SLE) (Aitman et al. 2006). Duplication of 15q11-q13 was found in 1-3% of humans with autism spectrum disorder (ASD) (Cook Jr et al. 1997). Sebat et al. (2007) showed that *de novo* CNVs were identified in 12 out of 118 patients with autism (*P* = 0.0005) (Sebat et al. 2007). However, Craddock et al. (2010) identified several false-positive CNV differences. Although replication analyses confirmed CNVs were associated with complex diseases, common CNVs contribute to the genetic basis in causing disease (Craddock et al. 2010).

Some functional CNVs are favored by positive selection in evolution. Therefore CNVs can be adaptive beneficial in some way (Sabeti et al. 2007; Nguyen et al. 2008). For example, human salivary amylase gene (*AMY1*) showed two diploid copies compared to chimpanzees. It is thought that this is an adaptation on high starch diets that increases the ability to digest and perform starchy foods (Perry et al. 2007). Some CNVs involve genes that influence normal human phenotypes such as triplication of the neuropeptide-Y4 receptor

(*PPYR1*), a gene that is directly involved in the regulation of food intake (Sainsbury et al. 2002).

1.2 Genome-wide association study (GWAS)

1.2.1 Overview of GWAS

Genome-wide association study (GWAS) is a high-throughput examination of common human genetic variants in different individuals to see if any variants are associated with complex traits. GWAS focus on associations between SNPs or CNVs and traits (Hardy and Singleton 2009). GWAS is typically based on comparison the DNA of case-control participants: patients with the disease (case) and disease-free people from the same population (control) (Pearson and Manolio 2008). If one or more alleles of a gene differ in people with a diseases, the SNP is said to be significantly associated with the disease. Unlike methods which test one or specific genetic regions, GWAS detects the entire genome. Therefore, the approach is not ideal for specific candidate-driven studies (Jiang 2013).

First successful GWAS from USA was published in March 2005. Klein et al. (2005) screened 96 patients with age-related macular degeneration (AMD) and identified two SNPs (rs380390 and rs1061170) which had altered allele frequency at the significance 5% level when comparing with healthy 50 controls (Klein et al. 2005). Since then, human GWAS has examined between hundreds or thousands of individuals. Several studies of GWA have often received criticism for omitting the quality control (QC) procedures gives the invalid findings, but modern studies address these problems and concerns.

1.2.2 Background of GWAS

Genomes between any two human may have millions of differences. Sequence differences are single nucleotides as well as copy number variations in the human genome. Any of these variations may lead to alterations traits or phenotypes.

Before the introduction of GWAS, the major method of analysis was through the family investigation of genetic linkage. This method has proved highly useful for associations between gene and disorder (Hamosh et al. 2000). However, for complex human diseases, the results of genetic linkage and specific disease-susceptibility studies has been limited to reproduce (Altmüller et al. 2001). Alternative approach to linkage studies in families was the genetic association studies. This study approach asks if the one or more alleles of a genetic variation is found in individuals with human disease phenotypes. This approach for statistical power could be better than genetic linkage analysis at detecting small gene effects for complex disorders (Risch and Merikangas 1996).

In addition to the several conceptual framework enabled the GWA studies. One was the Biobanks, which are stores of human biological material which time-consuming of collecting sufficient samples and information for biological study (Greely 2007). Another was the International HapMap Project, which had identified a majority of the common variants in the human genome

(Gibbs et al. 2003). The haploblock structures identified by the HapMap project would explain most of the variation.

1.2.3 Methods of GWAS

The most common design of GWAS is classifying individuals as the case-control, healthy group (control) and affected by a disease (case). All samples are genotyped for common SNPs. The number of SNPs vary on the microarray technology (platforms), yet are generally one million or more markers (Bush and Moore 2012). After that, each of these SNPs was analyzed if the allele frequency is significantly differ between case and control groups. Here, the basic unit for effective sizes is the odds ratios (OR). The OR is a measure of association derived from case compared with control. If the OR is higher than 1, the allele frequencies in the case group with a disease risk is greater than in the control group. A significant *p*-value of the odds ratio is generally calculated using a chi-square test. Finding ORs are different from 1 is the goal of the GWA study because this signify the SNP is associated with complex disease (Clarke et al. 2011).

There are several ways in the case-control approach. A common approach to case-control GWAS is the analysis of quantitative traits (e.g. height or biochemical marker concentrations). Calculations are generally done using bioinformatics tools such as PLINK which includes support for genetic-analysis statistics and convenient manner in big dataset (Purcell et al. 2007).

However, association calculation may to accompany several variables which can be potentially confused the results. Gender, age, and area are common example of these variables. Many variations are associated with the geographical patterns (Novembre et al. 2008). Because of this variations leading to potentially confusing result, association studies must consider of the geographical background of participants.

A *p*-value after adjustment were calculated for all variants, and then a common approach is to draw a Manhattan plot. In the GWA study, this plot shows the minus logarithm *p*-values. Therefore the most significant variants will remarkable on the plot. The *p*-value of significance threshold is adjusted for multiple testing and varies by studies, yet generally low *p*-value is considered significant in the tested variants (Bush and Moore 2012).

1.2.4 Results of GWAS

Many of efforts have been create the comprehensive catalog of CNVs or SNPs identified from GWA studies (Hindorff et al. 2009). Up to the recently, thousands of the variants associated with the complex diseases have reported (Johnson and O'Donnell 2009).

The first successful GWA study compared 96 individuals with the age-related macular degeneration (AMD) with 50 healthy controls. AMD is a cause of severe visual impairment. This study identified two significant SNPs (rs380390 and rs1061170) between the case-control groups. The SNPs were located in the complement factor H (*CFH*) gene. Therefore, *CFH* gene can be the susceptibility to AMD. The meaningful findings from these GWAS have revitalized more functional research towards the complex diseases (Haines et al. 2005). Another remarkable GWA study was the Wellcome Trust Case Control Consortium (WTCCC) study (case: 14,000 patients with seven common diseases; control: 3,000 individuals) published in 2007. This study successfully identified many new genes associated with these diseases (Burton et al. 2007). Since these remarkable GWA study, two trends have been created. One trend has been use more larger scale samples for more reliable detection of risk-SNPs (Ioannidis et al. 2009). Another has been towards more concrete phenotypes such as blood lipids (Kathiresan et al. 2008), liver biochemical markers (Kim et al. 2011), or proinsulin (Strawbridge et al. 2011). A key point in the GWAS debates has been that most of risk-SNPs identified by GWA studies have smaller predictive value for complex diseases (Ku et al. 2010). Generally, modest effective size of GWAS tend to be with the median OR is 1.33 per the risk allele (Hindorff et al. 2009).

1.2.5 Clinical application of GWAS

A challenge for GWA study will apply to a way the drug and diagnostic developments (Iadonato and Katze 2009). Several studies have investigated risk-SNPs improving the predictive value for complex diseases (Muehlschlegel et al. 2010). A problem with this approach is the small effective sizes. A small effect only has a small progress of the predictive value accuracy. Therefore, an alternative approach is explain pathophysiology for GWA studies. One of these alternative approaches was identified using the genetic variation associated with hepatitis C virus (HCV) treatment (Ge et al. 2009). For hepatitis C, the GWAS has shown that risk-SNPs near the *IL28B* gene are associated with significant twofold differences in response to the hepatitis C treatment (Ge et al. 2009).

Another aim of elucidating the pathophysiology has investigated the associations between risk variants and the expression of proximal susceptibility genes, the so-called expression as eQTL studies. The reason is that GWA studies for specific-genes improve towards target drug developments (Folkersen et al. 2010). For this reasons, most of GWA studies encompassed comprehensive eQTL analysis (Bown et al. 2011; Consortium 2011).

1.2.6 Limitations of GWAS

GWAS has several important limitations that should be taken into consideration and controlled for through quality control (QC) and study design. There are common issues such as lack of well-defined case and control participants, insufficient sample sizes, correction for multiple testing, population stratification, and many of statistical tests leading to a unexpected potential of false-positive results (Pearson and Manolio 2008). Ignoring these matters has been cited as study with the GWAS methodology problems. For example, a

GWAS investigating 1,055 individuals with long life spans to identify SNP-associated with longevity, was scrutinized due to a discrepancy of the genotyping array type between the case-control groups (Sebastiani et al. 2010). Therefore, the study was recanted.

These issues of GWA studies have suffered the criticism for assumption that genomic variants perform a central role in explaining the disease heritability (Couzin-Frankel 2010). Recently, as the decreasing expenditure of whole genome sequencing, the approach has alternated to GWAS which genotyping array-based.

1.3 Biological network

1.3.1 Overview of semantic network

Semantic networks represent knowledgeable relations between a concept types. Complex Systems are used as a form for representing as computable networks. It is a patterns of directed or undirected graphic notation consisting of edges and connections (Sowa 1991). A semantic network is used when one concept has semantic knowledge related to another. They also consist of arcs and nodes which can be organized into a taxonomic hierarchy.

However, semantic networks difficult handle for massive number of concepts, and they do not identify well-performance. Also, some properties of knowledgeable concepts are not easily represented using a semantic network. There are common examples–the presence of negation, disjunction, or non-taxonomic knowledge.

1.3.2 Biological network

There are several networks in biology such as protein-protein interactions (Mashaghi et al. 2004), gene regulatory (Vaquerizas et al. 2009), metabolic (Proulx et al. 2005), signaling, neuronal (Stephan et al. 2000), between-species interaction (Romanuk et al. 2010), and within-species interaction (Kasper and Voelkl 2009).

Biological network is representing of a large-scale knowledgeable systems. Understanding of principal organizations for biological network can be attain knowledgeable findings between network structure and flexible system (Prill et al. 2005). There are semantic networks software such as BioXM (http://www.biomax.com/), Biograph (http://www.biograph.be/), and Coremine (http://www.coremine.com/). Biograph is a data integration platform for biological information discovery (Liekens et al. 2011) and Coremine Medical is a web resource for seeking health and medicine information (de Leeuw et al. 2012). BioXM enables us to create a customizable knowledge base management for biological large amount and complex data (Maier et al. 2011).

This chapter consists of two parts.

Both all parts were published in *BMB reports*
as a partial fulfillment of HyoYoung Kim's Ph.D program.

Chapter 2. A replication study of GWA between CNVs and hepatic biomarkers AST or ALT in Korean cohorts

2.1 Abstract

Aspartate aminotransferase (AST) and alanine aminotransferase (ALT) are biochemical markers used as indicator for liver diseases and useful for diagnosing patients with liver disease. Copy number variation (CNV) play an important role in determining complex traits and is an emerging area in the study various diseases.

In this study, I performed replication studies of GWA between CNVs and the hepatic biochemical markers AST or ALT in KARE1 (n = 8,842) and KARE2 (n = 407) from population-based cohorts in Korea. I genotyped the genome-wide variations on an Affymetrix Genome-Wide Human 5.0 array in KARE1 and Affymetrix 6.0 in KARE2. CNVs were identified using Helix Tree software. And then, to explain the impact of CNVs on each quantitative trait, univariate linear regression was performed. As the result, in KARE1, 100 CNVs were significant for AST and 16 were significant for ALT (p-value < 0.05 after Bonferroni correction). In KARE2, 32 and 42 CNVs showed significance for AST and ALT, respectively (p-value < 0.05). I compared CNV-based genes between the KARE2 (AST-140, ALT-172) and KARE1 (AST-1885, ALT-773) using NetBox to replication studies. Results showed that nine genes (*CIDEB*, *DFFA*, *PSMA3*, *PSMC5*, *PSMC6*, *PSMD12*, *PSMF1*, *SDC4*, and *SIAH1*) were overlapped for AST, but no overlapped genes were found for ALT. Functional gene classification analysis shown four clusters (proteasome

pathway, Wnt signaling pathway, programmed cell death, and protein binding) using the Visualization and Integrated Discovery (DAVID) tool. Structural variation analysis of CNV-based genes is useful to understand of the biological phenotypes or disease.

2.2 Introduction

The liver with dark reddish brown color is the second largest glandular organ in the human body and is located under the lib on the right side. The organ has many functions, including remove and detoxify harmful substances from blood, storage of glycogen, filtration of harmful substances such as alcohol, and maintenance of normal glucose concentration (http://www.britishlivertrust.org.uk; http://www.liverfoundation.org). The liver also produces urea and the majority of cholesterol in the body (about 80% of the body) (http://www.mamashealth.com/) (Gitzelmann et al. 1996; Pocai et al. 2006; Zhang and Beynen 2007). Biochemical tests for liver function are commonly used to diagnose patients with liver disease (Sattar et al. 2004). Aspartate aminotransferase (AST) and alanine aminotransferase (ALT) are biochemical markers widely used as markers for identify the physical state of liver or diagnosis of hepatic diseases such as fatty liver and alcoholic hepatitis (Bathum et al. 2001; Hanley et al. 2005). ALT is an enzyme mainly found in hepatocytes, and AST is another hepatocellular enzyme. The ratio of serum levels of AST/ALT is used as an indicator for the evaluation of hepatitis patients (Sheth et al. 1998). Typically, an AST/ALT ratio of less than one indicates mild liver disease, such as nonalcoholic fatty liver disease (NAFLD), whereas, an AST/ALT ratio greater than one implies severe liver disease, such as cirrhosis, chronic hepatitis or alcoholic fatty liver disease (Clemenz et al. 2008). Two loci (10q24.2 and 22q13.31) have been identified as influencing the plasma levels of ALT in three European populations (Switzerland: n = 5,636, Italy: n = 1,200, London: n = 879) (Yuan et al. 2008).

Over the past few years, efforts focus on investigate the effects of CNV in human disease have been continuing (Glessner and Hakonarson 2009; Xu et al. 2009; Glessner et al. 2009). Both CNV and SNP were used to identify what associated with susceptibility and resistance to diseases. Genetic differences are shown by large-scale structural variations in different individuals. Differences in copy number contribute to changes in gene expression. Hence, DNA copy number variations (CNVs) contribute to genomic variation between humans (Wang et al. 1998). Copy number variation (CNV) is a form of structural variation as a DNA segment ≥ 1 kb in size when compared to a reference genome assembly. Studies on genetic variation contribute to the understanding of individual phenotypic differences which can be manifested in drug dosage effects and susceptibility to disease (Estivill and Armengol 2007). Many CNVs in the human genome have been identified in various populations (Perry et al. 2008; de Stahl et al. 2008). According to a CNV study from 4 populations with different ancestries in Asia, Africa, and Europe, CNVs accounted for ~12% of the genome in these populations (Redon et al. 2006). CNVs have been shown to comprise 17.7% of the detected variations in gene expression. Consequently, CNVs play an important role in determining complex traits (Stranger et al. 2007; Beckmann et al. 2007).

Genome-wide association studies (GWAS) have been used to investigate novel candidate genes of common diseases. Many studies on the association between CNVs and complex diseases in humans have been reported (Hastings et al. 2009). Recent GWA studies have localized common DNA sequence variants associated with hepatic biomarkers AST or ALR (Kim et al. 2010), and replication of genome-side associations scans revealed common variants nine genes that contribute to liver diseases (Kim et al. 2011). However, association studies between CNV and diseases have been hindered due to incomplete knowledge of CNV detection criterion and lack of a reference CNV. Additionally, although most of the CNVs have been identified in various populations, the results may not directly apply to CNVs of all ethnicities (Yim et al. 2009).

While many studies have examined the biology of liver disease in humans, few have focused on the identification of liver-associated CNVs; moreover, CNVs have not been identified in Koreans. I tried to identify liver-associated CNVs in Koreans and determine their biological significance. Here, I studied the replication studies of GWA based on 8,842 (KARE1) and 407 (KARE2) from population-based cohorts recruited in Korea related to hepatic biomarkers AST or ALT. Through a single-CNV analysis for each liver-related trait using univariate linear regression, I identified 100 with AST and 16 with ALT CNV regions in KARE1 and 32 with AST and 42 with ALT CNV regions in KARE2. I compared CNV-based genes in KARE1 and KARE2. Nine genes were overlapped for AST. This result has functional implications for CNVs associated with liver function. Data obtained from the Korean Genome Association Study of this study provide valuable CNV-related information associated with liver disease.

2.3 Materials and Methods

2.3.1 Study subjects

To the genome-wide association (GWA) studies, the Korea Association Resource 1 (KARE1) and KARE2 project were established in 2007 and 2009, respectively. All study subjects signed an NIH (National Institute of Health)-approved informed-consent forms.

KARE1 data is constituted the urban Ansan (n = 5,020) and rural Ansung (n = 5,018) two population-based Korean cohorts. The participants were aged 40 to 69 (persons born during 1931 – 1963). The genomic DNA were isolated from peripheral blood of healthy participants. In KARE1, I chose 8,842 chips (Ansan = 4,205, Ansung = 4,637) after quality control (QC) of genotyping data with high heterozygosity, high missing genotype call rate, gender inconsistence and individuals with cancer by Cho et al., (Cho et al. 2009). The mean age was 52.2 years. In KARE2, I genotyped 407 unrelated Koreans (men = 154, women = 253). Subject ages ranged from 35 to 80 years (mean 62.13 ± 6.9). For CNV analysis, a 500 ng sample of genomic DNA isolated from the peripheral blood of each participant was measured.

2.3.2 CNV discovery

I assayed the genome-wide variations on an Affymetrix Genome-Wide Human 5.0 array in KARE1 and Affymetrix 6.0 array in KARE2 (Affymetrix, USA).

CEL files containing the intensity-level values were imported into the HelixTree software (ver. 7.0) for the discovery of CNV (Golden Helix Inc., USA) (Lambert 2005). The Helix Tree analysis software reading the intensity data, normalizing on probe intensities against reference sets, and creating normalized log2 ratios. CNVs require a reference genome to be compared with samples. If a reference consists of imported chips run in different labs or using ethnicities, systematic differences represented variability. Therefore, I used the mean intensity value of all chips instead of other ethnic or small samples as a reference to minimize the variability causing chips or systemic differences as much as possible. The copy number analysis module (CNAM) in the HelixTree was used to read the intensity files, normalize intensity values against reference samples, import log2 ratios and segment CNV region. The analysis parameters included a multivariate algorithm, a moving window of 5,000 markers, a maximum of 100 segments/window, a minimum of 1 marker/segment, and a significance level of $p < 0.01$ for pair-wise permutations (n = 1,000). The multivariate algorithm segmented all samples simultaneously, making it possible to perform the CNV association study for all samples.

2.3.3 CNV association study of liver functions

To explain the impact of CNVs on each quantitative trait, I performed univariate linear regression (McMurray et al. 2004). The additive genetic model were corrected for area, age, and gender in KARE1 and corrected age and gender in KARE2.

For continuous variables in KARE1,

$$Y = \beta_0 + \beta_1 CNV + \beta_2 Area + \beta_3 Age + \beta_4 Gender + \varepsilon$$

For continuous variables in KARE2,

$$Y = \beta_0 + \beta_1 CNV + \beta_2 Age + \beta_3 Gender + \varepsilon$$

where β is a coefficients p-vector. For multiple corrections, significance was determined at the level of Bonferroni *p*-value < 0.05 in KARE1 and FDR *p*-value of < 0.05 in KARE2. The log2 ratio of each CNV associated with continuous response variables was analyzed via the following univariate linear regression model. All statistical analyses and parsing were performed using the statistical software R (http://www.r-project.org/; ver. 2.9) and Python software.

2.3.4 Enrichment analysis of CNV-based genes

I assembled the genes whose entire sequences were located within the CNV region associated with the liver phenotypes. The genes were identified using the RefGene (ver. hg18) downloaded from the UCSC genome browser (http://genome.ucsc.edu/; ver. hg18). To the functional analysis of the genes, I adopted two function sets from the Database for Annotation, Visualization and Integrated Discovery tool (http://david.abcc.ncifcrf.gov/) including Gene Ontology (Harris et al. 2004) and KEGG (Kyoto Encyclopedia of Genes and

Genomes) pathway (Kanehisa et al. 2002) (http://david.abcc.ncifcrf.gov/; ver. 6.7 Beta). The GO sets include biological process (BP), molecular process (MF), and cellular component (CC). Diseases associated with genes were obtained using OMIM (http://www.ncbi.nlm.nih.gov/omim/), the Genetic Association Database (http://geneticassociationdb.nih.gov/cgi-bin/index.cgi) and BioGPS (http://biogps.org/#goto=welcome). NetBox software (http://cbio.mskcc.org/tools/netbox.html) (Cerami et al. 2010) was used for replication study of GWA and network modules were visualized using Cytoscape (Shannon et al. 2003). All data were parsed using the Python programming (ver. 2.5).

2.4 Results

2.4.1 Analysis of serum liver enzymes (AST or ALT)

An association studies of CNV with disease susceptibility or dosage effect have become an attractive field since some CNVs were reported to be associated with various types of disease (Glessner and Hakonarson 2009; Glessner et al. 2009; Walters et al. 2010). From the KARE cohorts, I focused on identifying CNVs associated with hepatic biochemical markers AST or ALT in Koreans. In this study, the values of AST and ALT were transformed to 1/(*y*) and 1/square root(*y*) to approximate a normal distribution, respectively. I compared beanplots to show the frequency distributions of the AST or ALT in KARE1 and KARE2 (Figure 2.1). As the results, I did not show differences in distributions between the two populations or between genders. Also, I computed Pearson's correlation coefficients to evaluate whether AST and ALT have a conserved relationship. The results showed that AST has a significant positive correlation with ALT (correlation value of 0.73; p-value < 0.05).

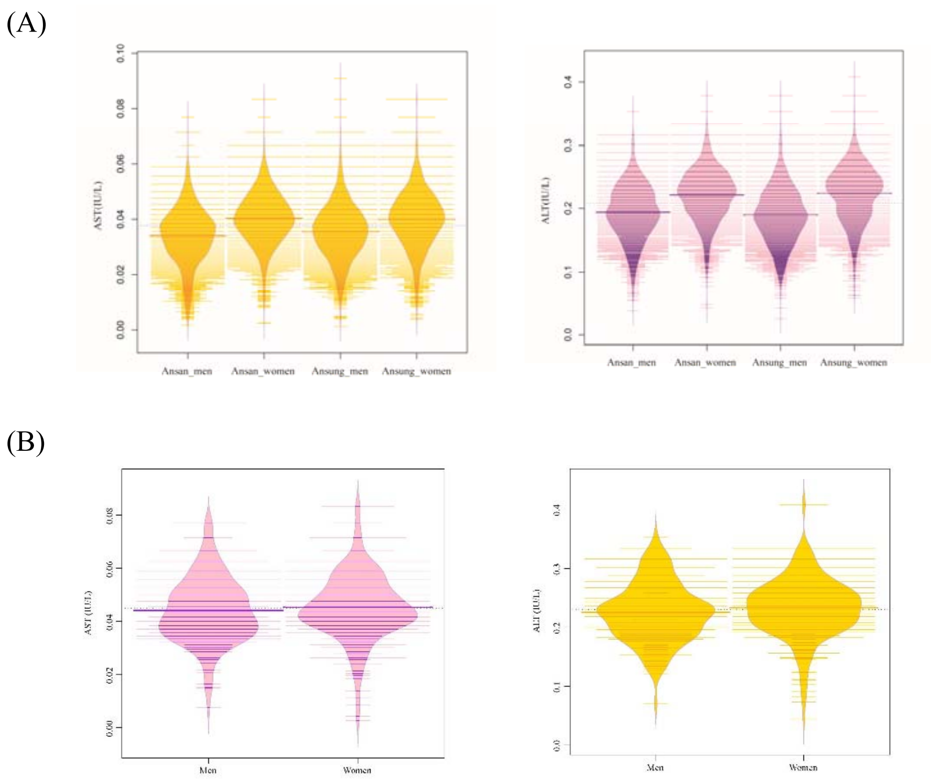


Figure 2.1. Beanplots of the distributions of AST or ALT in KARE1 (A) and KARE2 (B). Thick lines denote the average values of AST or ALT.

2.4.2 Discovery of CNVs

I extracted 10,162 CNVs in KARE1 and 3,046 CNVs in KARE2 using the multivariate segmenting option provided by HelixTree software (Supplementary Figure 2.1). The copy number analysis module (CNAM) is to create normalized log2 ratios. The CNAM module reads the Affymetrix CEL intensity files, normalizes the intensity values against reference samples, and imports the log2 ratios. The CNAM segmenting process is optimized through 1) subdivision of the chromosomal region of markers into a moving window of sub-regions and 2) a permutation algorithm that validates the found cut points. Then, the algorithm detects CNV segment boundaries and can properly delineate segment boundaries with controllable sensitivity and false discovery rate. The multivariate algorithm segments all samples simultaneously. The summary of CNVs is given in Table 2.1.

Table 2.1. Summary of significant CNVs in KARE1 and KARE2 identified in Korean cohorts.

	KARE1	KARE2
Total counts	10,162	3,046
Average size per CNV	727.3	911.0 Kb
Median size (range)	112 Kb (2 - 31,415 Kb)	548 Kb (1 – 24,744 Kb)

2.4.3 Association study between CNVs and hepatic biochemical markers

For each CNV, I analyzed the impact of a single CNV for each quantitative phenotype using univariate linear regression. The linear model for the CNV-trait association study was performed based on continuous value of independent variable (CNV log2 ratio). As the result, the positive β of AST and ALT was 4200 and 5384, and the negative β was 6334 and 5150 in KARE1, respectively. In KARE2, the positive β values of AST and ALT were 1,605 and 1,949, and the negative β values were 1,441 and 1,097, respectively. Univariate linear regression analysis identified significant CNVs 100 loci for AST and 16 loci for ALT in KARE1 and 32 loci for AST and 42 loci for ALT in KARE2 (Figure 2.2; Supplementary Table 2.1). Figure 2.3 shows the genome-wide association signals for AST and ALT on all 22 autosomes in Manhattan plots. The QQ plot displays for AST and ALT results for the GWAS (lambda=1.92 for AST, lambda=1.08 for ALT; Supplementary Figure 2.2).

I found 39 and 228 genes completely located within significant CNV regions for AST or ALT (Supplementary Table 2.2). Table 2.2 summarizes the gene lists, beta-coefficients, and liver-associated phenotypes in KARE1.

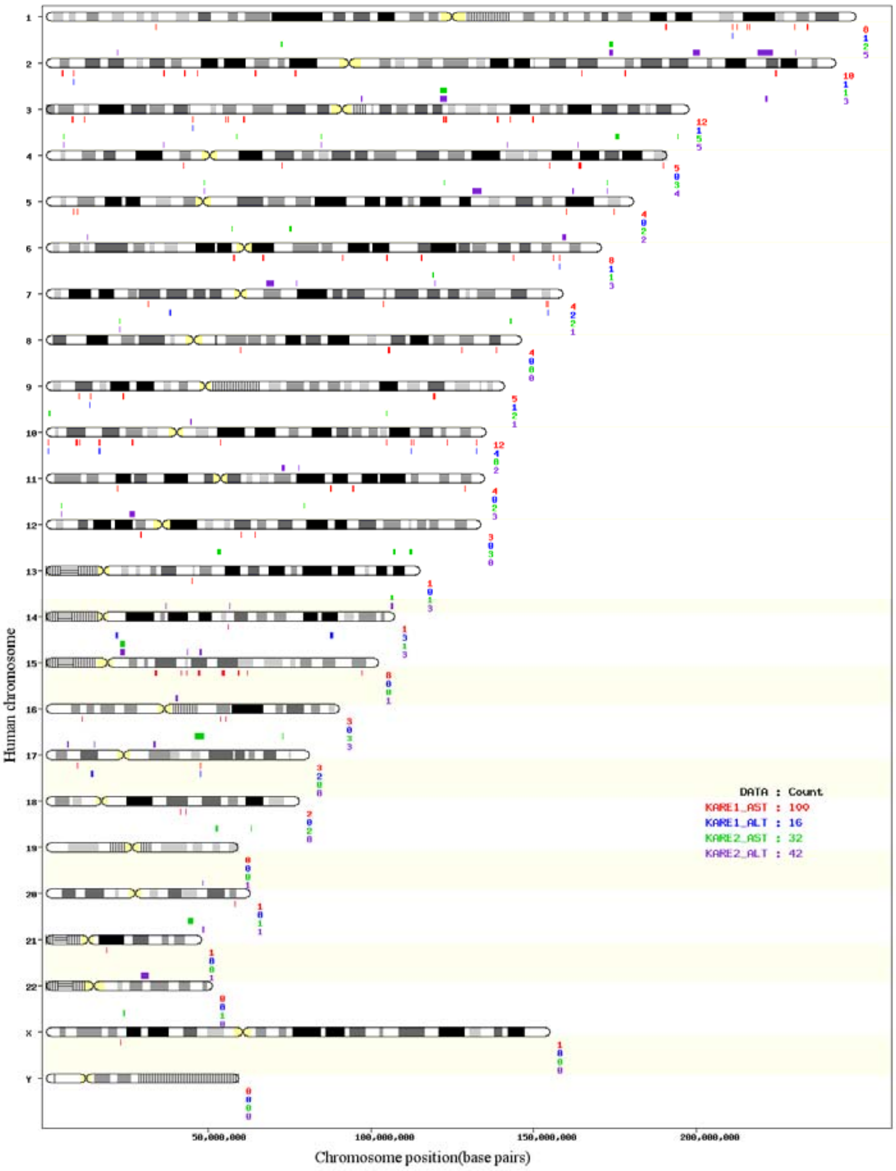


Figure 2.2. Visualization of the physical distribution of significant CNV regions for AST or ALT in KARE1 and KARE2.

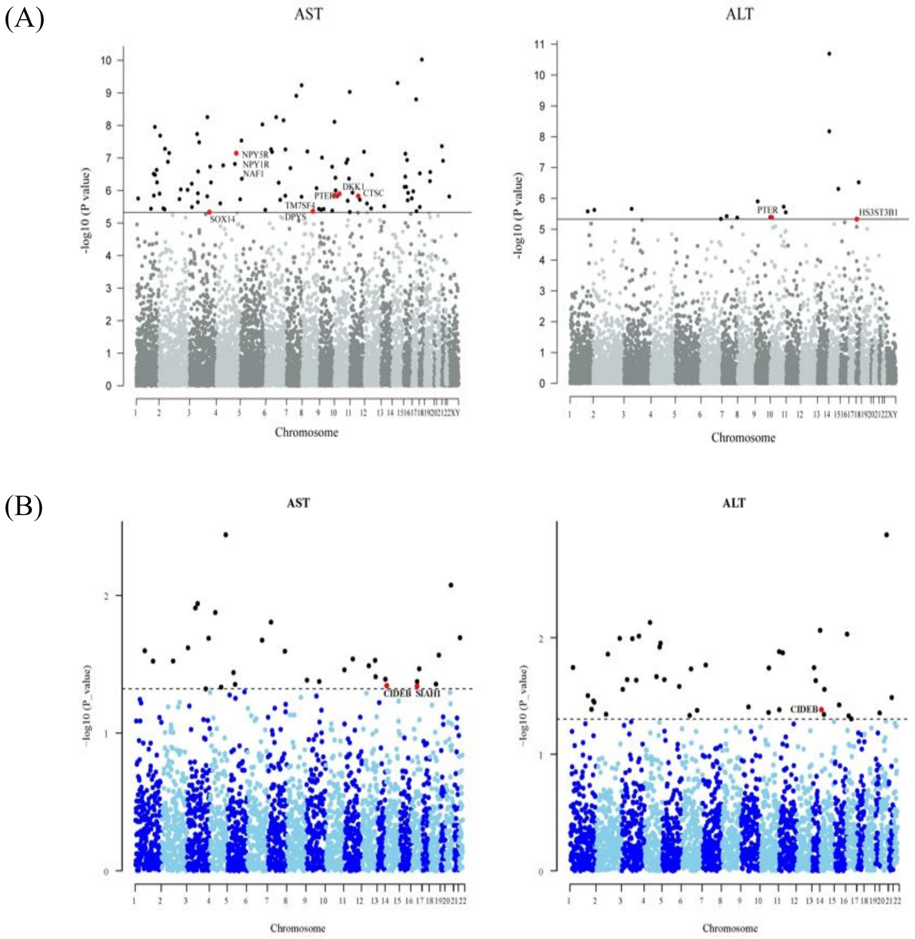


Figure 2.3. Manhattan plot shows the genome-wide association signals between all CNVs and AST or ALT on all 22 autosomes in KARE1 (A) and KARE2 (B). Association was assessed using univariate linear regression adjusted for gender and age. The X axis shows chromosomal locations, and the *p*-value was plotted on the Y axis using a logarithmic scale. The black dotted significant CNVs and the red dotted genes associated with the liver were identified in previously studies.

Table 2.2. Thirty-nine genes associated with serum liver enzymes AST or ALT.

Trait	Gene[a]	Beta-coefficient	*p*-value	Liver-associated phenotype	Literature (year)
AST	*NPY5R*[*]	–0.0126	7.43E-04	Dyslipidemia-caused fatty liver disease	Marceau et al. (2010)
	NPY1R[*]	–0.0126	7.43E-04	Neuropeptide Y receptor activity	GO
	NAF1[*]	–0.0126	7.43E-04	Glycoprotein process	GO
	DKK1[*]	–0.0071	1.38E-02	Wnt signaling inhibitor	Fedi et al. (1999)
	CTSC[*]	–0.0117	1.55E-02	Papillon–Lefèvre syndrome	Almuneef et al. (2003)
	TM7SF4[*]	–0.0126	4.45E-02	Highest expression in liver	Staege et al. (2001)
	DPYS[*]	–0.0126	4.45E-02	Dihydropyrimidinuria	Gennip et al. (1997); Nyhan (2005)
	SOX14[*]	–0.0113	4.80E-02	Lower level in adult liver	Arsic et al. (1998)
	MAP3K7	–0.0108	5.85E-05		
	TKTL2	–0.0126	7.43E-04		
	ZNF280D	–0.0152	1.22E-03		
	TEX9	–0.0152	1.22E-03		
	MNS1	–0.0152	1.22E-03		
	HSPC159	–0.0043	1.38E-03		

	SHC4	–0.0176	8.10E-03		
	COPS2	–0.0176	8.10E-03		
	GALK2	–0.0176	8.10E-03		
	SECISBP2L	–0.0176	8.10E-03		
	CEP152	–0.0176	8.10E-03		
	EID1	–0.0176	8.10E-03		
	SERPINE2	–0.0123	1.00E-02		
	MRPL44	–0.0123	1.00E-02		
	TNP2	–0.0026	1.11E-02		
	PRM2	–0.0026	1.11E-02		
	PRM3	–0.0026	1.11E-02		
	PRM1	–0.0026	1.11E-02		
	ERGIC2	–0.01	2.62E-02		
	FAR2	–0.01	2.62E-02		
	LRP12	–0.0126	1.45E-02		
AST/ALT	_PTER_*	0.0143	1.50E-02	Low expression in liver	Hou et al. (1996)
	_C1QL3_	0.0143	1.50E-02		
ALT	_HS3ST3B1_*	–0.009	4.97E-02	Abundant in liver	Lyon et al. (1994); Shworak et al. (1999)
	KCNK10	0.0142	5.18E-03		
	_ZC3H14_	0.0142	5.18E-03		
	PTPN21	0.0142	5.18E-03		
	SPATA7	0.0142	5.18E-03		

EML5	0.0142	5.18E-03
CDRT15	–0.009	4.97E-02
MGC129 *16*	–0.009	4.97E-02

[a]: There are 39 genes (p-value < 0.05) significantly selected for each trait;

[*]: The 10 genes encompassed GO identified as liver-associated in previous studies are indicated by asterisks ([*]).

2.4.4 Replication study of CNV-based genes associated with AST or ALT

I searched whether some genes were replicated when compared to reported previous study. To replication study of GWA associated with AST or ALT, I compared CNV-based genes between KARE1 (AST: 1,885 and ALT: 773) and KARE2 (AST: 140 and ALT: 172) using the NetBox software. Figure 2.4 shows visualized networks as determined using the Cytoscape, which is a popular software for visualizing complex interaction networks (Shannon et al. 2003). I discovered four large modules, with a network modularity score of 0.004. I identified nine genes (*CIDEB, DFFA, PSMA3, PSMC5, PSMC6, PSMD12, PSMF1, SDC4,* and *SIAH1*) were overlapped for AST (Figure 2.5). Unfortunately, no overlapped gene was found for ALT. Notably, seven genes except for *CIDEB* and *SIAH1*were not included in our gene list, but were identified as linker gene because significantly connected to our input gene list. A total of 8 genes appeared within the network modules, but *SDC4* was not present within the network at the shortest path threshold of 2, and the linker *p*-value cut-off of 0.05.

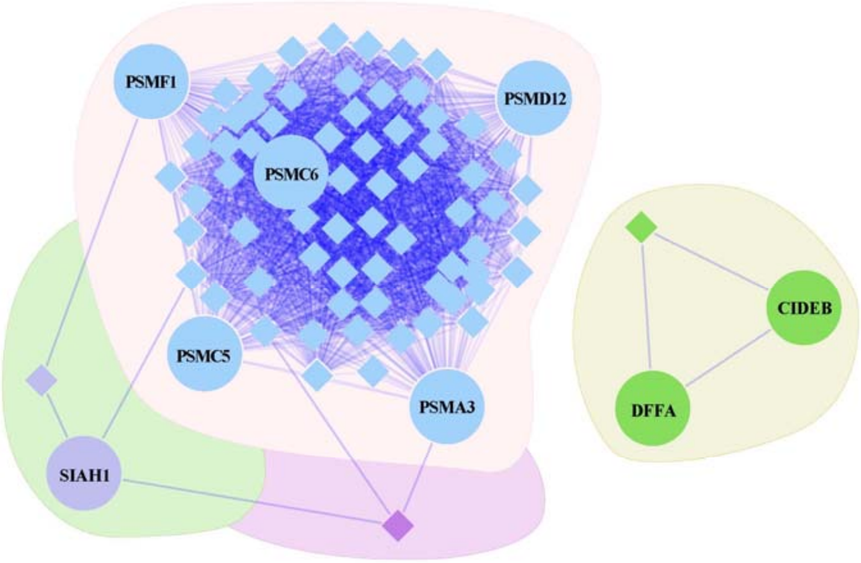


Figure 2.4. The four largest modules were identified with a network modularity score of 0.004. Linker genes, showed as diamond shape, were not included in the original gene list, but are significantly connected with list-altered genes.

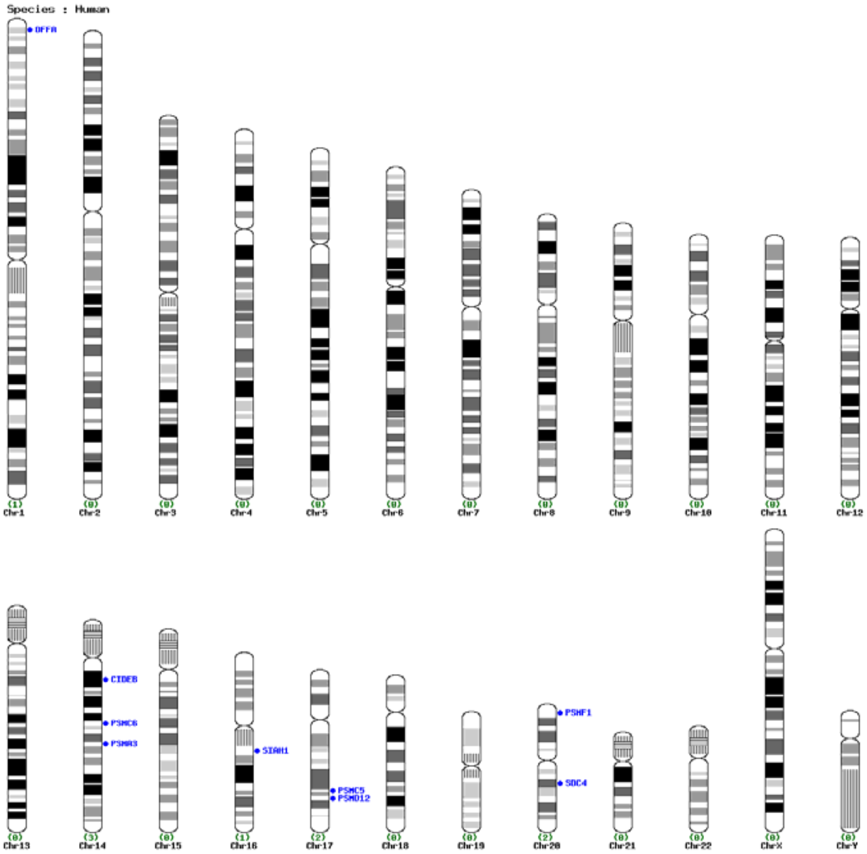
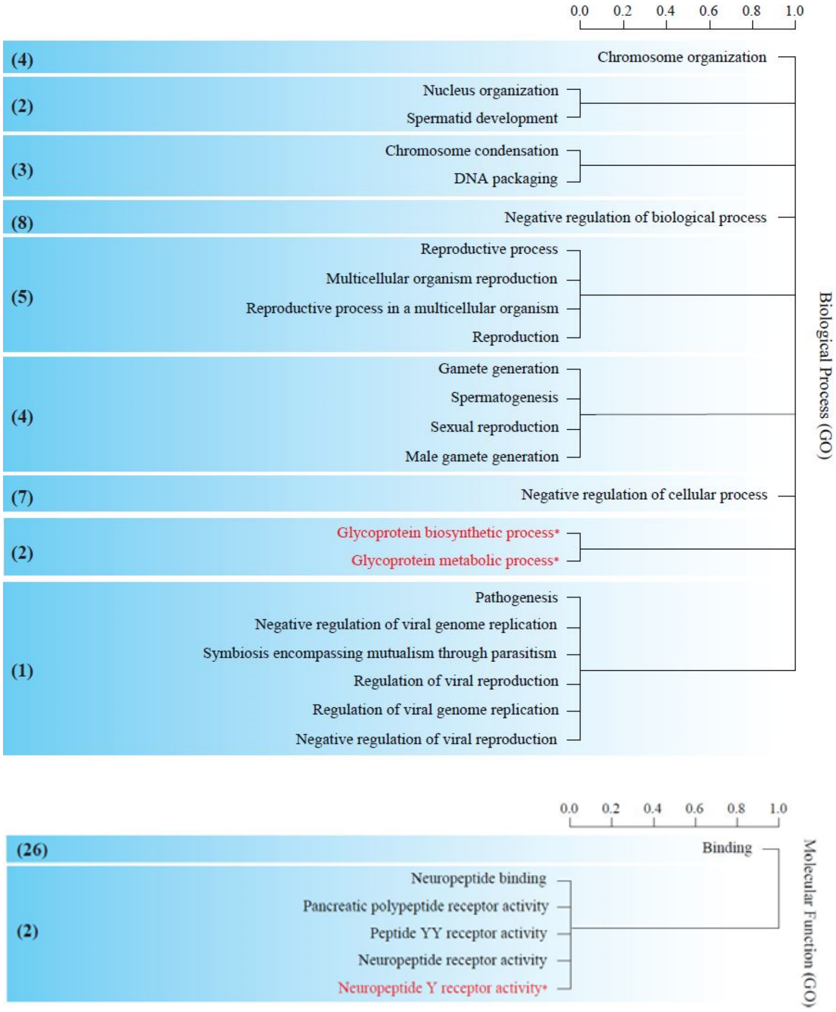


Figure 2.5. CytoMap view of nine genes associate with liver by replication studies of GWA. Green: the total number of genes on each chromosome.

2.4.5 Proteasome pathway is enriched in AST

To probe the functional implications of structural variants, I analyzed the functional annotation of the nine genes included in the CNV by the single linear regression analysis for hepatic biochemical markers using the DAVID tool (Figure 2.6). Among the genes identified, four genes (*PSMF1, PSMC6, PSMD12,* and *PSMA3*) were enriched in the proteasome biochemical pathway (P = 2.20E-07). The proteasome play a role in inhibiting cytokine production by liver cells. A decrease of proteasome activity develops during alcoholic liver injury and leads to inhibition of cell death. Therefore, chronic ethanol consumption suppresses proteasome activity in the liver (Donohue Jr et al. 2007; Donohue Jr 2002). Although not detected the significant enrichment groups in the KEGG pathway, *SIAH1* was found in the Wnt signaling pathway, which plays a role in liver development and regeneration (Armengol et al. 2011). Okabe et al. (2003) found that the expression of *SIAH1* was down-regulated in all hepatoma cells lines examined when compared with normal liver cells by semiquantitative RT-PCR. The decreased expression of *SIAH1* plays an important role in the development of hepatocellular carcinoma (HCC) (Okabe et al. 2003).

(A)


0.0
0.2
0.4
0.6
0.8
1.0
(4)
Chromosome organization
(2)
Nucleus organization
Spermatid development
(3)
Chromosome condensation
DNA packaging
(8)
Negative regulation of biological process
(5)
Reproductive process
Multicellular organism reproduction
Reproductive process in a multicellular organism
Reproduction
(4)
Gamete generation
Spermatogenesis
Sexual reproduction
Male gamete generation
(7)
Negative regulation of cellular process
(2)
Glycoprotein biosynthetic process*
Glycoprotein metabolic process*
(1)
Pathogenesis
Negative regulation of viral genome replication
Symbiosis encompassing mutualism through parasitism
Regulation of viral reproduction
Regulation of viral genome replication
Negative regulation of viral reproduction
Biological Process (GO)
0.0
0.2
0.4
0.6
0.8
1.0
(26)
Binding
(2)
Neuropeptide binding
Pancreatic polypeptide receptor activity
Peptide YY receptor activity
Neuropeptide receptor activity
Neuropeptide Y receptor activity*
Molecular Function (GO)

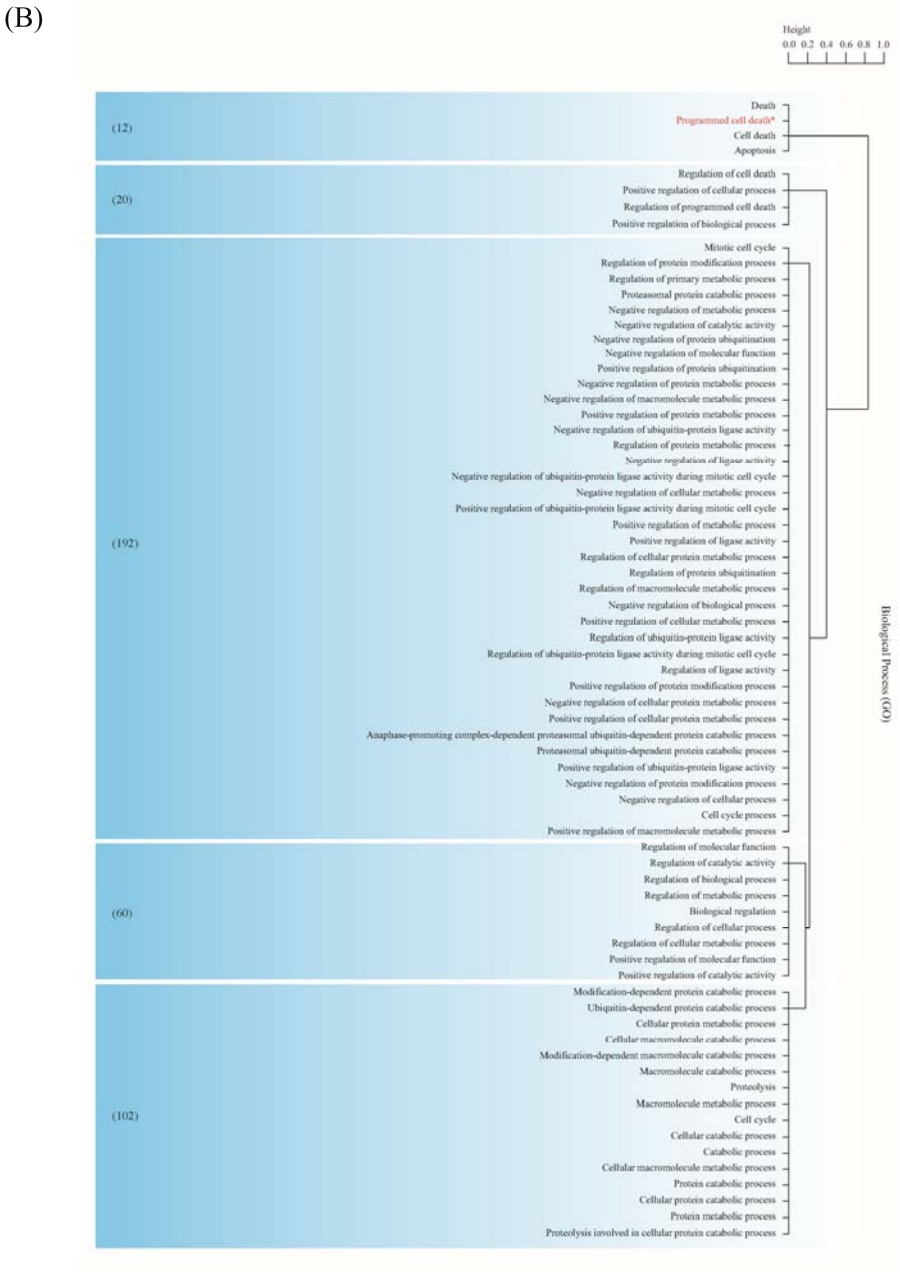

(B)
Height
0.0 0.2 0.4 0.6 0.8 1.0
(12)
(20)
(192)
(60)
(102)
Death
Programmed cell death*
Cell death
Apoptosis
Regulation of cell death
Positive regulation of cellular process
Regulation of programmed cell death
Positive regulation of biological process
Mitotic cell cycle
Regulation of protein modification process
Regulation of primary metabolic process
Proteasomal protein catabolic process
Negative regulation of metabolic process
Negative regulation of catalytic activity
Negative regulation of protein ubiquitination
Negative regulation of molecular function
Positive regulation of protein ubiquitination
Negative regulation of protein metabolic process
Negative regulation of macromolecule metabolic process
Positive regulation of protein metabolic process
Negative regulation of ubiquitin-protein ligase activity
Regulation of protein metabolic process
Negative regulation of ligase activity
Negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle
Negative regulation of cellular metabolic process
Positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle
Positive regulation of metabolic process
Positive regulation of ligase activity
Regulation of cellular protein metabolic process
Regulation of protein ubiquitination
Regulation of macromolecule metabolic process
Negative regulation of biological process
Positive regulation of cellular metabolic process
Regulation of ubiquitin-protein ligase activity
Regulation of ubiquitin-protein ligase activity during mitotic cell cycle
Regulation of ligase activity
Positive regulation of protein modification process
Negative regulation of cellular protein metabolic process
Positive regulation of cellular protein metabolic process
Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
Proteasomal ubiquitin-dependent protein catabolic process
Positive regulation of ubiquitin-protein ligase activity
Negative regulation of protein modification process
Negative regulation of cellular process
Cell cycle process
Positive regulation of macromolecule metabolic process
Regulation of molecular function
Regulation of catalytic activity
Regulation of biological process
Regulation of metabolic process
Biological regulation
Regulation of cellular process
Regulation of cellular metabolic process
Positive regulation of molecular function
Positive regulation of catalytic activity
Modification-dependent protein catabolic process
Ubiquitin-dependent protein catabolic process
Cellular protein metabolic process
Cellular macromolecule catabolic process
Modification-dependent macromolecule catabolic process
Macromolecule catabolic process
Proteolysis
Macromolecule metabolic process
Cell cycle
Cellular catabolic process
Catabolic process
Cellular macromolecule metabolic process
Protein catabolic process
Cellular protein catabolic process
Protein metabolic process
Proteolysis involved in cellular protein catabolic process
Biological Process (GO)

Figure 2.6. Tree views of enriched Gene Ontology (GO) categories. Enriched GO categories are visualized for 39 and 9 genes found within CNVs associated with AST or ALT in KARE1 (A) and KARE2 (B). Numbers in

parentheses at the left denote the gene counts within GO groups. The terms identified as liver-associated in previous studies are indicated by the red asterisks (*).

2.4.6 Programmed cell death and protein binding are enriched in AST

The enriched Gene Ontology terms of biological process included programmed cell death (*DFFA, CIDEB*, and *SIAH1*; P = 0.04; Figure 3.5). Programmed cell death (PCD) plays an important role in liver development (Saad et al. 2009). Inohara et al. (1998) identified *CIDEB*, which is a subunit of the DNA fragment factor (DFF) (Inohara et al. 1998). The *CIDEB* (cell death-inducing DFFA-like effector B) is expressed at high levels and plays an important role as a regulator of lipid metabolism in the liver (Li et al. 2007; Ye et al. 2009; Li 2007). All nine genes demonstrated enriched molecular functions, including protein binding (P = 0.0071). Liver disease can affect protein binding and causes impaired plasma protein binding of azapropazone (Blaschke; Jahnchen et al. 1981). Kojima et al. (1992) isolated *SDC4* from a rat endothelial cell (Kojima et al. 1992). Rioux et al. (2002) identified *SDC4* (Syndecan-4) expressed at high levels in mouse liver tissue by Northern blot analysis (Rioux et al. 2002). The SDC4 gene is comprised of 5 exons, and located in human chromosome 20q12. Table 2.3 shows that in previous studies, all nine genes were reported to be associated with liver function.

Table 2.3. Nine genes identified in the Korean cohorts and previous studies of liver.

Gene	Liver-associated phenotype(s)	Enriched term	References
CIDEB	Programmed cell death	GO_BP	Saad et al. (2009)
	High expression in liver		Li et al. (2007); Ye et al. (2009)
DFFA	Programmed cell death	GO_BP	Saad et al. (2009)
PSMA3	Proteasome	KEGG	Donohue (2002); Donohue et al. (2007)
PSMC6	Proteasome	KEGG	Donohue (2002); Donohue et al. (2007)
	Overexpressed in hepatocytes		Richert et al. (2006)
PSMD12	Proteasome	KEGG	Donohue (2002); Donohue et al. (2007)
	Overexpressed in hepatocytes		Richert et al. (2006)
PSMF1	Proteasome	KEGG	Donohue (2002); Donohue et al. (2007)
SDC4	Abundant in liver		Rioux et al. (2002)
SIAH1	Wnt signaling pathway		Armengol et al. (2011)
	Programmed cell death	GO_BP	Saad et al. (2009)
PSMC5	Protein binding	GO_MF	Jahnchen et al. (1981)

2.5 Discussion

An association study of CNV has been important to understand the effect of variations on complex diseases since some CNVs reported association with disease (Lee et al. 2012). In this study, I extracted 10,162 and 3,046 CNVs associated with hepatic biochemical markers AST or ALT in Korean cohorts. AST or ALT are the most common indicators of liver disease. The median size of CNVs was 112 kb and 547 kb in KARE1 and KARE2, respectively. This result was a little different to distribution size and counts of CNVs compared to a previous CNV study using other samples of the same KARE cohorts (Yim et al. 2009). It seems that because CNVs are not defining but vary criterion and are very diverse depending in technical sources such as platforms or references, and by the different statistical analysis algorithms. Especially, a knowledge on the association study of CNVs and diseases is still incomplete in statistical analysis (Lee et al. 2012). Supplementary Figure 2.3 shows the distribution of the number of CNVs in this study KARE1 and KARE2 compare to previously found CNVs in same Korean populations by Yim et al., (Yim et al. 2010). The average size of per CNV was 727.3 and 911.0 Kb, and the median size of CNVs was 112 and 548 Kb in KARE1 and KARE2, respectively. The high density SNP genotyping arrays have become more popular for copy number variation using a signal intensity measures, however there are limitations to the use of SNP genotyping arrays for CNV detection. Generally, SNPs in these arrays are not uniformly distributed across the genome. For example, SNP array 5.0 does not have quality control (QC) measure while the SNP array 6.0 (Affymetrix) uses MAPD as a QC measure. To overcome these limitations, experimental validation would be the best way to confirm the result. Unfortunately, I would not be able to do the experimental validation. Instead of a validation, I used a simple alternative to decreasing variability to increase the quality of CNV calls from a chip. First, chips selected for CNV analysis were from a QC genotyping reference of the same samples by Cho et al, (2009) (Cho et al. 2009) in which samples with a high missing genotype call rate, high heterozygosity, gender inconsistencies, and those obtained from individuals who had developed any kind of cancer were excluded, along with related or identical individuals whose computed average pairwise identity-by-state value was higher than that estimated from first-degree relatives of Korean sib-pair samples. Second, I used these all chips as a reference group. If a reference was generated from chips run in another lab, such systematic differences inflated apparent variability. Therefore, using a reference generated from the same batch was a way to reduce chip variability. In addition, instead of using a reference or a small sample of references, referencing all of the samples decreased the variability because it used a global average value as a reference for each CNV call from a chip. Supplementary Figure 2.4 shows the CNV log2 ratio distributions of the 16 significant. Some of the frequency distribution did not fit with normal distribution. However, CNV is independent variable so that normal distribution assumption is not necessary condition for the association study.

To detect for association between single CNV and each adjusted phenotype, genome-wide CNV association studies for AST and ALT have been performed. While univariate linear regression used to identify Single Nucleotide Polymorphism (SNP) (Cooper et al. 2008), little reported that apply single linear regression to discover CNVs. However, dingle linear regression model was fitted to explain the impact of single CNV regions on each quantitative trait.

I identified genes inclusive to CNV regions. Genes fully inclusive to a CNV may be explained liver functions by a regression model. I investigated the functional implications of the genes using the DAVID functional annotation tool (Dennis Jr et al. 2003). The results showed clustering several biochemical pathways and Gene Ontology (GO) annotations relevant to AST or ALT. In KARE1, four genes (*DKK1, DPYS*, *HS3ST3B1*, and *MAP3K7*) were enriched in 10 KEGG pathways, including heparan sulfate biosynthesis, pyrimidine and beta-alanine metabolism, and Wnt signaling. The *HS3ST3B1* gene is involved in the heparan sulfate biosynthesis pathway. The *HS3ST3B1* gene was found to be involved in the biosynthesis of heparin sulfate, which is a polysaccharide complex synthesized in most mammalian cells (Shworak et al. 1999). The *3OST3B* gene shows wide expression of multiple transcripts and is most abundant in the liver (Lyon et al. 1994). *DKK1* and *MAP3K7* were found to be involved in the Wnt signaling pathway, which plays an important role in developing and regenerating the liver (Armengol et al. 2011). *DKK1* expression was down-regulated in fetal liver and inhibits Wnt signaling in mammalian cells (Fedi et al. 1999). Also, glycoprotein biosynthetic and neuropeptide Y receptor activity showed enrichment relevant to AST and ALT. The liver produces the glycoprotein hormone that regulates production of bone marrow platelets (http://review-center.net/metabolism/liver-metabolism-pathways-and-its-disorders/). Several glycoproteins, including fibronectin, hyalurinic, laminin, merosin, nidogen, and tenascin, are expressed in fibrotic livers (Kladney et al. 2000). One such glycoprotein, GP73 (Golgi protein), is up-regulated upon hepatitis viral infection (Block et al. 2005). Enriched molecular functions involving neuropeptide Y receptor activity were shown in Figure 2.6. (A). Neuropeptide Y was identified in human livers where it regulates blood flow and secretion in the liver (Ding et al. 1991; El-Salhy 1999). Using the Genetic Association Database (GAD), I detected one gene associated with liver disease, *NPY5R*. Neuropeptide Y receptor Y5 (*NPY5R*) is known to be associated with dyslipidemia, a fatty liver disease. Marceau et al. (2010) showed dyslipidemia is an important risk factor for fatty liver disease (Marceau et al. 1999). Five genes (*CTSC*, *DPYS*, *HS3ST3B1*, *PRM3,* and *SPATA7*) were shown to be correlated with human disease states using OMIM. The genes annotated using OMIM represent nine disease phenotypes, including dihydropyrimidinuria, Papillon–Lefèvre syndrome (PLS), and Haim–Munk syndrome. *DPYS and CTSC* were found to be associated with dihydropyrimidinuria, a deficiency in dihydropy(Van Gennip et al. 1997) rimidinase (DHP), and PLS phenotypes, respectively. The activity of DHP, which is exclusively expressed in the liver, is characterized by increased excretion of dihydrothymine (Nyhan 2005) and dihydrouracil . Mutations in the cathepsin C (*CTSC*) gene cause Haim-Munk syndrome and PLS, a rare autosomal-recessive disease characterized by juvenile periodontitis. Pyogeneic liver abscesses are well recognized complication of neutrophil dysfunction in PLS(Almuneef et al. 2003). Four genes (*CTSC*, *DPYS, GALK2* and *PTER*) were found to be actively expressed in human liver using *BioGPS* (Wu et al. 2009). This is evident from gene expression patterns produced by the GeneAtlas U133A data sets. Further, *Pter* expression was down-regulated in mouse liver tissue (Hou et al. 1996).

The NetBox software is based on copy number alteration and sequence mutation data, and assembles altered genes. It identifies linker genes, connects all altered genes, and then identifies network modules and calculates network modularity (Cerami et al. 2010; Ding et al. 2010). Although many replication-analysis methods have been reported (Bax et al. 2006), none were appropriate for our gene-based CNV data. For replication study of GWA, I compared CNV-based genes between the current study and KARE1 using NetBox for replication-analysis. Results showed that nine genes (*CIDEB, DFFA, PSMA3, PSMC5, PSMC6, PSMD12, PSMF1, SDC4,* and *SIAH1*) were overlapped for only AST, but none were overlapped for ALT.

Regarding functional implications of the 9 genes, I analyzed functional classification using the DAVID tool. Our gene lists were clustered into functionally related groups. This analysis showed interesting results regarding CNV-based genes associated with liver. For AST trait, I identified one enriched gene cluster. The four genes (*PSMF1, PSMC6, PSMD12,* and *PSMA3*) in this cluster were enriched in the proteasome biochemical pathway, which inhibition cytokine production by liver cells (P = 2.20E-07), and *SIAH1* was shown Wnt signaling pathway, which plays a role in liver development and regeneration (Armengol et al. 2011). The decrease of proteasome activity causes alcoholic liver injury and inhibits liver cell death. Therefore, chronic ethanol consumption suppressed proteasome activity in the liver (Donohue Jr et al. 2007; Donohue Jr 2002). Richert et al. (2006) reported that *PSMC6* (ATPase activity subunit) and *PSMD12* (a non-ATPase subunit) were significantly overexpressed in human hepatocytes (Richert et al. 2006).

The enriched Gene Ontology clusters were programmed cell death (*DFFA, CIDEB* and *SIAH1*; P = 0.04) and protein binding (all 9 genes; P = 0.0071). The *PSMC6* and *PSMD12* genes encode a 403 and 397 amino-acid protein, and are located on chromosome 14q22.1 and 17q24.2, respectively. Okabe et al. (2003) found the expression of *SIAH1* was down-regulated in all hepatoma cells lines. The decreased expression of *SIAH1* plays an important role in the development of hepatocellular carcinoma (Okabe et al. 2003).

In conclusion, I investigated CNVs associated with the liver biomarkers AST and ALT in 407 unrelated Koreans using the Affymetrix Genome-Wide

6.0 array. Four genes (*PSMF1, PSMC6, PSMD12,* and *PSMA3*) are involved in the proteasome biochemical pathway, and *SIAH1* was shown to be active in the Wnt signaling pathway. The 3 genes (*DFFA, CIDEB,* and *SIAH1*) were active in programmed cell death, and all 9 genes showed significant enrichment in protein binding, based on Gene Ontology. The enrichment of these genes suggests susceptibility or resistance mechanisms for liver disease. Analysis of specific traits based on genes with CNVs were influenced by the gene interactions involved in different processes associated with liver. Overall, our CNV-based genes identified in this study will provide a valuable resource for further investigations of liver diseases. Additionally, our results require validation for candidate genes using quantitative PCR (qPCR).

This chapter consists of two parts.

The liver function network part was published in *Molecular and Cellular Toxicology* as a partial fulfillment of HyoYoung Kim's Ph.D program.

The ethnic disparities network part was published in *Genomics & Informatics* as a partial fulfillment of HyoYoung Kim's Ph.D program.

Chapter 3. Biological networks to identify knowledgeable meanings for liver functions or ethnic disparities

3.1 Abstract

A semantic network is needed for in-depth understanding of the impacts of SNPs, because phenotypes are modulated by complex networks including biochemical and physiological pathways. Copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) have been emerging out of the efforts to research about human health, complex diseases, and ethnic disparities.

In this study, I focused on constructing semantic networks for liver functions or ethnic disparities using the knowledge integration BioXM software. Entities for the network represented by “Gene”, “Pathway”, “Disease”, “Chemical”, “Drug”, “ClinicalTrials”, “CNV”, “SNP”, “SomaticMutation”, and relationships between entity-entity were obtained such as “Gene-SNP”, “Gene-Disease”, “Gene-Chemical”, “Gene-Pathway”, “Gene-GO”, “Gene-SNP”, “Gene-CNV”, “Gene-SomaticMutation”, “Pathway-Disease”, “Pathway-Chemical”, “ClinicalTrials-Disease”, “ClinicalTrials-Drug”, “Disease-Chemical”, “Chemical-Drug”, and “Disease-Chemical-Drug” through curation. To evaluate the two biological networks, KARE2 and ethnicity specific SNPs data were applied to liver functions or ethnic disparities networks, respectively. KARE2 data were explained in chapter 2. Ethnic specific SNPs were identified by eliminating overlapped SNPs from the HapMap samples, and the ethnic specific SNPs were mapped to the UCSC RefGene lists (ver. hg18). Application of liver diseases network using KARE2 data was shown three clusters, including four diseases (“Hepatocellular

carcinoma", "Liver neoplasm", "Liver cell adenoma", and "Drug-induced liver injury"), one pathway ("Hepatitis C pathway"), and seven drugs ("Acetaminophen", "Chlormezanone", "Stavudine", "Enflurane", "isoniazid", "Mebendazole", and "Nitisinone"). The semantic findings for ethnic disparities network showed interesting results in the three categories, including three diseases ("AIDS-Associated Nephropathy", "Hypertension", and "Pelvic Infection"), one drug ("Methylphenidate"), and five pathways ("Hemostasis", "Systemic lupus erythematosus", "Prostate cancer", "Hepatitis C virus", and "Rheumatoid arthritis").

I found biological implications for liver functions or ethnic disparities using the semantic networks, and the majority of our findings was consistent with the previous studies that an understanding of genetic variability explained liver functions or ethnic disparities.

3.2 Introduction

Data mining is provide important insights into the data with complicated and huge quantity. Semantic modeling has gained attentions as a powerful tool for organizing and integrating biological big data (McCray and Nelson 1995). Semantic technology is needed to provide the knowledge generation helping to gain an adequate interpretation of integrated biological systems (Losko and Heumann 2009). Theses semantic network have given researcher aids to semantic information answer about complex questions through integration of the available data (Shin et al. 2012). Recent advances in ontology development, like the semantic modeling, are considered to contribute to the next-generation approach by enabling the researcher to actually ask scientific questions instead of constructing complicated databases for scientific questions and answers (Mukherjea et al. 2005). This combination of data integration and visualization could provide important insights into heterogeneous data on millions of genes, chemical compounds, diseases and pathways (Kim et al. 2013; Kim et al. 2014).

To model a semantic network-modeling, the BioXM software is a customizable knowledge management program for scientific big data, and the latest solution is designed to provide meaningful interactions through graphical browsing (Maier et al. 2011). Through an advanced query builder, the knowledge consisting of many different and connected queries is flexibly examined. In this way, models for a research project can be constructed and extended effectively. Many data mining studies and software developments have been advanced various fields, but there are relatively few studies have focused on data mining about liver diseases or ethnicity disparities.

In the past few years, enormous efforts have been made to investigate the role of SNPs or CNVs in health and disease (McCarroll and Altshuler 2007). Differences of copy number between individuals contribute to alter in expression of genes sensitive to a disease susceptibility or dosage effect (Redon et al. 2006). GWA studies of SNPs or CNVs are active and fast studying to discover the genetic basis of common complex diseases such as cancer, cardiovascular disorder, and autism (Zhang et al. 2010). Gerber et al. (2012) identified rs6983267 variant was associated with colorectal cancer at a significant genome-wide level (Gerber et al. 2012). Peters et al. (2012) found eight SNP-based genes (*SMAD7*, *GREM1*, *EIF3H*, *11q23*, *BMP2*, *BMP4*, *CDH1*, and *MYC*) to be associated with colorectal cancer using replication GWA study (Peters et al. 2012).

To investigate the biological knowledgeable findings about liver functions or ethnic disparities, in the current study, I constructed semantic knowledgeable networks. I expect that this semantic modeling-based studies will provide valuable information on CNVs associated with liver functions or ethnic specific SNP-based genes, and strongly affects useful knowledge in liver functions or ethnic disparities.

3.3 Materials and Methods

3.3.1 Semantic networks for liver functions or ethnic disparities datasets

I constructed semantic networks in order to a diverse interactions for human liver functions or ethnic disparities using BioXM Knowledge Management Environment software, which efficiently knowledge manages, such as complex scientific big data (ver. 2.2) (Maier et al. 2011). The BioXM enable to create customizable knowledge base management for biological large amount and complex data. The modeling provides semantic networks with the useful knowledgeable relationship information between participating entities. The semantic networks consisted of entities including "Gene (Davis et al. 2009)", "Pathway (Davis et al. 2009)", "Disease (Davis et al. 2009)", "Chemical (Davis et al. 2009)", "Drug (Wishart et al. 2006)", "SNP (Karolchik et al. 2003)" and "ClinicalTrials (http://www.clinicaltrials.gov)", and relations including "Pathway-Gene", "Disease-Pathway", "Disease-Chemical", "Gene-Disease", "Gene-Chemical", "SNP-Gene", "Chemical-Pathway", "Chemical-Drug", "ClinicalTrials-Disease", and "Drug-ClinicalTrials". Semantic network instantly provides semantic objects as well as the connection information between participating objects. I generalized this complex semantic network of detecting entity and connection for the answer to complex questions. Therefore this semantic integration for liver function data enables us to create new knowledge networks with flexible workflow modeling. Conversion of all data to entity input format was parsed using the Python software and R package.

3.3.2 Study subjects for ethnic disparities network

I downloaded the single nucleotide polymorphisms (SNPs) data from Haplotype Map (HapMap) phase 3 (http://www.hapmap.org) for CEU (Utah residents with Northern and Western European ancestry), JPT (Japanese in Tokyo, Japan), and YRI (Yoruba in Ibadan, Nigeria). I focused on the gene-based SNPs associations in the three ethnicities because ethnicity is a highly heritable polygenic quantitative trait of biomedical importance. Ethnicity-specific SNPs were obtained by eliminating common SNPs.

3.3.3 Enrichment analysis for SNP-based gene associated with ethnic disparities

Ethnic specific SNPs were mapped to genes from the UCSC RefGene (http://genome.ucsc.edu/; ver.hg18). For the mapped genes, gene set enrichment analysis (GSEA) was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) tool (http://david.abcc.ncifcrf.gov/; ver. 6.7) (Dennis Jr et al. 2003) with KEGG pathway and Gene Ontology (GO) terms, including biological process (BP), cellular component (CC), and molecular function (MF) (http://www.geneontology.org/). The *p*-values were calculated for the probability of getting a set of genes within a given GO group.

3.4 Results

3.4.1 Semantic networks for liver functions or ethnic disparities

I constructed two semantic networks in order to analyze the knowledgeable findings for liver functions or ethnic disparities. Overall, network entities were used such as “Gene”, “Pathway”, “Disease”, “Chemical”, “Drug”, “ClinicalTrials”, and “SNP”, and pairwise relationships between entity-entity were curated as “Gene-Pathway”, “Gene-Disease”, “Gene-Chemical”, “Disease-Chemical”, “Disease-Pathway”, “Chemical-Pathway”, “Chemical-Drug”, “SNP-Gene”, “ClinicalTrials-Drug”, and “ClinicalTrials-Disease”. Table 3.1 summaries of the source, information, and roles of the entities and Table 3.2 summaries of entity and relation information such as source DB and records. Gene entity was consisted of information, including Gene ID, NCBI Accession, position, curated integrating from the UCSC Human Genome Browser and the Comparative Toxicogenomics Database (CTD). Entities, including “Pathway”, “Chemical”, and “Disease” were collected from the CTD, which is a public database to promote the understanding of the interaction of genes, chemical compounds, and disease networks in human health. Drug was provided information for name, description, CAS number, indication, pharmacology, mechanism of action, toxicity, biotransformation, and absorption, brands from DrugBank (Wishart et al. 2008), which provides detailed drug action information. SNP was mapped against 1000 Genomes (Overbeek et al. 2005), and CNV was mapped against TCGA (Higgins et al.

2007). After that, to promote understanding about the mechanism of entities, the relations between two entities such as Gene-Pathway, Gene-Disease, Gene-Chemical, Disease-Chemical, Disease-Pathway, and Chemical-Pathway associations were curated against from the CTD. In particular, relation data with unknown interaction such as Chemical-Drug, Gene-SNP, Gene-CNV, Gene-GO, Gene-SomaticMutation, ClinicalTrials-Disease, and ClinicalTrials-Drug was parsed using the Python software. Therefore, curated associations are identified, and users helpful improve understanding about biological mechanisms. Figure 3.1 show that the scheme of semantic data integration model for human liver diseases or ethnic disparities is dynamic and flexible. Hierarchy structure is where the parent can have one child, while in Directed Acyclic Graph (DAG) networks, like BioXM is the parent can be has more than one child. For example, Gene A is associated with whether Chemical B or Pathway C. Also, Gene A is associated with Drug C, because Gene A is a curated interaction with Disease B, and Disease B is a curated association with Drug C.

Table 3.1. Summaries of the source, information, and roles of the entities.

Entity	Source	Information	Roles
Gene Ontology (GO)	www.geneontology.org	Offer the organizational functions and roles of gene(s)	Identify the similarity and/or the functional classification of genes
Disease Ontology (DO)	disease-ontology.org	Offer the classification system about various diseases	Identify similar and/or the functional classification of diseases
Gene	genome.ucsc.edu	Offer the in silico PCR, Blat, and information associated with genome	Identify the location information of human genes (ver. hg 19)
SNP/ CNV	www.1000genomes .org	Offer SNPs and/or full genome sequence of the thousands human	Identify the location and information of SNPs in 1,000 individuals
Cancer	cancergenome.nih.gov	Offer the information about Cancer genomics	Identify the information of CNVs associated with cancers
Chemical/ Pathway/	ctdbase.org	Offer the correlation association of	Identify the correlation association of

Disease		Chemical, Gene, Disease, and Pathway	Chemical, Gene, Disease, and Pathway
Clinical Trials	www.clinicaltrials.gov	Offer the information of drugs and clinical trials used to disease treatment	Identify the correlation association of Disease and Drug
Drug	www.drugbank.ca	Offer the detail information associated with drugs	Identify the correlation association of Drug and Chemical
Somatic Mutation	www.sanger.ac.uk	Offer the information of Somatic mutation	Identify the correlation association between Disease and Gene associated with Somatic mutation

Table 3.2. Summaries of integrated data sets.

Entity	Source DB	Records	Relation	Records
Gene	UCSC Genome Bioinformatics ctd Giving insight into how chemicals affect our health Comparative Toxicogenomics Database	46,354	Gene-Disease,	18,391,755
			Gene-Pathway,	62,057
			Gene-Chemical	308,405
Somatic Mutation	sanger	242,217	Somatic Mutation-Gene	32,695
GO	Gene Ontology Consortium	36855	GO-Gene	185,929
Pathway	ctd Giving insight into how chemicals affect our health Comparative Toxicogenomics Database	362	Pathway-Disease	43,139
			Pathway-Chemical	196.073
Disease	ctd Giving insight into how chemicals affect our health Comparative Toxicogenomics Database	9,647	Disease-Chemical	401,145
Chemical	ctd Giving insight into how chemicals affect our health Comparative Toxicogenomics Database	153,021	Chemical-Drug	1,702
Clinical Trials	ClinicalTrials.gov A service of the U.S. National Institutes of Health	1,273	ClinicalTrials-Disease	1,210
Drug	DrugBank Open Data Drug & Drug Target Database	6,712	Drug-ClinicalTrials	1,419
CNV	1000 Genomes A Deep Catalog of Human Genetic Variation	21,591	CNV-Gene	31,740
SNP	1000 Genomes A Deep Catalog of Human Genetic Variation	154,84	SNP-Gene	37,060
Total		518,032	Total	19,498,452

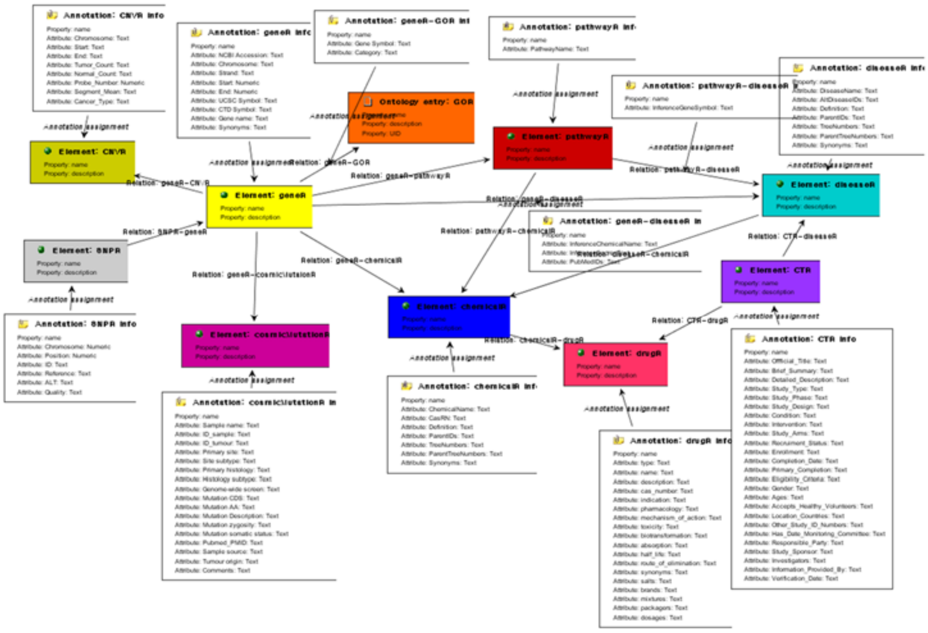

Annotation: CNVR Info
Property: name
Attribute: Chromosome: Text
Attribute: Start: Text
Attribute: End: Text
Attribute: Tumor_Count: Text
Attribute: Normal_Count: Text
Attribute: Probe_Number: Numeric
Attribute: Segment_Mean: Text
Attribute: Cancer_Type: Text
Annotation: geneR Info
Property: name
Attribute: NCBI Accession: Text
Attribute: Chromosome: Text
Attribute: Strand: Text
Attribute: Start: Numeric
Attribute: End: Numeric
Attribute: UCSC Symbol: Text
Attribute: CTD Symbol: Text
Attribute: Gene name: Text
Attribute: Synonyms: Text
Annotation: geneR-GOR Inf
Property: name
Attribute: Gene Symbol: Text
Attribute: Category: Text
Annotation: pathwayR Inf
Property: name
Attribute: PathwayName: Text
Annotation: pathwayR-diseaseR
Property: name
Attribute: InferenceGeneSymbol: Text
Annotation: diseaseR Inf
Property: name
Attribute: DiseaseName: Text
Attribute: AltDiseaseIDs: Text
Attribute: Definition: Text
Attribute: ParentIDs: Text
Attribute: TreeNumbers: Text
Attribute: ParentTreeNumbers: Text
Attribute: Synonyms: Text
Ontology entry: GOR
Property: name
Property: description
Property: UID
Element: CNVR
Property: name
Property: description
Element: geneR
Property: name
Property: description
Element: pathwayR
Property: name
Property: description
Element: diseaseR
Property: name
Property: description
Element: SNPR
Property: name
Property: description
Element: cosmicMutationR
Property: name
Property: description
Element: chemicalR
Property: name
Property: description
Element: CTR
Property: name
Property: description
Element: drugR
Property: name
Property: description
Annotation assignment
Relation: geneR-GOR
Relation: geneR-CNVR
Relation: geneR-pathwayR
Relation: geneR-diseaseR
Relation: pathwayR-diseaseR
Relation: SNPR-geneR
Relation: pathwayR-chemicalR
Relation: geneR-chemicalR
Relation: geneR-cosmicMutationR
Relation: diseaseR-chemicalR
Relation: CTR-diseaseR
Relation: CTR-drugR
Relation: chemicalR-drugR
Annotation: geneR-diseaseR In
Property: name
Attribute: InferenceChemicalName: Text
Attribute: PubMedIDs: Text
Annotation: SNPR Info
Property: name
Attribute: Chromosome: Numeric
Attribute: Position: Numeric
Attribute: ID: Text
Attribute: Reference: Text
Attribute: ALT: Text
Attribute: Quality: Text
Annotation: cosmicMutationR In
Property: name
Attribute: Sample name: Text
Attribute: ID_sample: Text
Attribute: ID_tumour: Text
Attribute: Primary site: Text
Attribute: Site subtype: Text
Attribute: Primary histology: Text
Attribute: Histology subtype: Text
Attribute: Genome-wide screen: Text
Attribute: Mutation CDS: Text
Attribute: Mutation AA: Text
Attribute: Mutation Description: Text
Attribute: Mutation zygosity: Text
Attribute: Mutation somatic status: Text
Attribute: Pubmed_PMID: Text
Attribute: Sample source: Text
Attribute: Tumour origin: Text
Attribute: Comments: Text
Annotation: chemicalR Inf
Property: name
Attribute: ChemicalName: Text
Attribute: CasRN: Text
Attribute: Definition: Text
Attribute: ParentIDs: Text
Attribute: TreeNumbers: Text
Attribute: ParentTreeNumbers: Text
Attribute: Synonyms: Text
Annotation: drugR Info
Property: name
Attribute: type: Text
Attribute: name: Text
Attribute: description: Text
Attribute: cas_number: Text
Attribute: indication: Text
Attribute: pharmacology: Text
Attribute: mechanism_of_action: Text
Attribute: toxicity: Text
Attribute: biotransformation: Text
Attribute: absorption: Text
Attribute: half_life: Text
Attribute: route_of_elimination: Text
Attribute: synonyms: Text
Attribute: salts: Text
Attribute: brands: Text
Attribute: mixtures: Text
Attribute: packagers: Text
Attribute: dosages: Text
Annotation: CTR Info
Property: name
Attribute: Official_Title: Text
Attribute: Brief_Summary: Text
Attribute: Detailed_Description: Text
Attribute: Study_Type: Text
Attribute: Study_Phase: Text
Attribute: Study_Design: Text
Attribute: Condition: Text
Attribute: Intervention: Text
Attribute: Study_Arms: Text
Attribute: Recruitment_Status: Text
Attribute: Enrollment: Text
Attribute: Completion_Date: Text
Attribute: Primary_Completion: Text
Attribute: Eligibility_Criteria: Text
Attribute: Gender: Text
Attribute: Ages: Text
Attribute: Accepts_Healthy_Volunteers: Text
Attribute: Location_Countries: Text
Attribute: Other_Study_ID_Numbers: Text
Attribute: Has_Data_Monitoring_Committee: Text
Attribute: Responsible_Party: Text
Attribute: Study_Sponsor: Text
Attribute: Investigators: Text
Attribute: Information_Provided_By: Text
Attribute: Verification_Date: Text

Figure 3.1. Scheme of semantic data integration model for human liver functions. The color box represents entities with description (white box), such as gene, chemical, pathway, disease, drug, GO, CNV, and SNP. The black arrows indicated the relations between two entities such as Gene-GO, Gene-Pathway, Gene-Disease, Gene-Chemical, Gene-SNP, Gene-CNV, Gene-SomaticMutation, Pathway-Chemical, Pathway-Disease, Chemical-Drug, ClinicalTrials-Disease, ClinicalTrials-Drug.

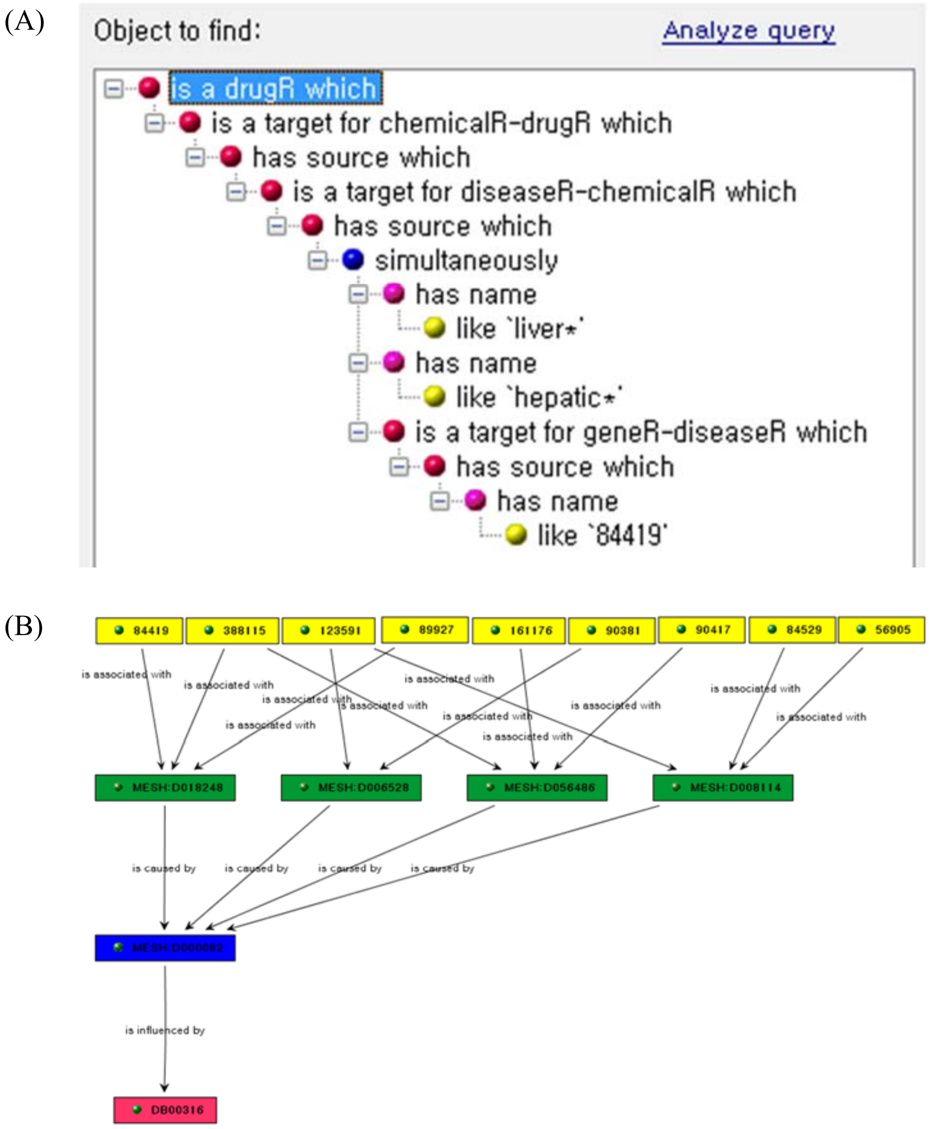
3.4.2 Semantic findings using liver functions network

Using the semantically integrated liver function datasets, I analyzed the functional implications of CNV-based genes for hepatic biomarkers AST or ALT using KARE2 data. As the result, the genes were showed interactions with four diseases (hepatocellular carcinoma, liver neoplasm, liver cell adenoma, and drug-induced liver injury), one pathway (hepatitis C pathway), and seven drugs (acetaminophen, chlormezanone, stavudine, enflurane, isoniazid, mebendazole, and nitisinone).

Hepatocellular Carcinoma is a primary malignant liver neoplasm. It is the six most common cancer and the third leading cause of cancer death in the world (Taniguchi et al. 2002). Liver Neoplasms is an another name for hepatocellular carcinoma (Liu et al. 2011). Liver Cell Adenoma is a hepatocellular benign epithelial tumor (Zucman-Rossi 2011). It occur most often in women who take higher dose of estrogen hormone pills. Since symptoms were generally not observed in patients, most was never detected. When discovered a large adenoma, it is surgically removed (http://www.liverfoundation.org). Drug-Induced Liver Injury (DILI) is known as hepatotoxicity. It caused by drugs agents and reactions, and more than 1000 drugs have been associated with significant hepatic injury (Davern 2012). DILI is classified into intrinsic and idiosyncratic types; intrinsic DILI is dose dependent whereas idiosyncratic DILI is not dose-related (Björnsson and Chalasani 2010).

Supplementary Table 3.1 summaries four diseases and one pathway associated with hepatic biomarkers AST or ALT. Two genes (IRF9 and OAS2) were revealed hepatitis C pathway (KEGG:05160). Hepatitis C is a hepatitis C virus (HCV)-associated liver disease. HCV causes the liver to prevents its functions from working well, and is a main risk factor for hepatocellular carcinoma (Farazi and DePinho 2006). About 25% of people with HCV fully recover within six months, but about 75% of HCV-infected people develop chronic HCV, and chronic HCV can lead to cirrhosis, liver cancer, and liver injury (Kampstra 2008). Most acute or chronic HCV-infected people have no symptoms, but can occur symptoms such as tiredness, dark urine, itchy skin, poor appetite, abdominal pain, muscle soreness, and jaundice (Warrell and Anderson 2014). Treating for acute HCV was recommended rest, drinking large amount of fluids, eating healthy food, and avoiding alcohol. Patients with chronic HCV was treated with taking two oral medicines boceprevir and telaprevir, protease inhibitors that binds to the NS3 active site (Steinkühler et al. 1998).

Drug is very clinically important cause of liver injury, and many drugs have been reported to cause liver injury (Lee 2003). Gene-Drug interaction was established on the semantic integrated human liver disease datasets. Gene A is associated with drug D because gene A has a curated interaction with disease B, and disease B has a curated association with chemical C, and chemical C has a curated association with drug D. By smart query wizards, seven drugs (acetaminophen, chlormezanone, stavudine, enflurane, isoniazid, mebendazole, and nitisinone) were associated with AST and ALT (Figure 3.2; Table 3.3). Acetaminophen (APAP) is metabolized primarily in the liver, and APAP-overdose is the predominent cause of hepatic injury (Davidson and Eastham 1966). Stavudine is an antiviral medication that active against human immunodeficiency virus. It can cause severe or often life-threatening effects on liver, and can increase risk of liver damage while taking it (http://www.drugs.com/mtm/stavudine.html). Isoniazid is an antibiotic, which prevents tuberculous bacteria (http://www.drugs.com/mtm/isoniazid.html). Mebendazole is an anti-worm medication and used for prevents infections of such as pinworm (Kullai Reddy Ulavapalli et al. 2011), whipworm (Miller et al. 1974), roundworm (Lubis 2008), and hookworm (De Clercq et al. 1997). Nitisinone is used to treat hereditary tyrosinemia type 1. It keeps causing harm to liver tissue, and its symptom is liver failure (http://www.drugs.com/mtm/nitisinone.html).



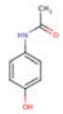
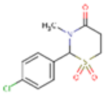
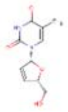
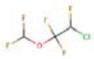
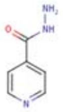
Object to find: Analyze query

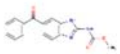
```
is a drugR which
  is a target for chemicalR-drugR which
    has source which
      is a target for diseaseR-chemicalR which
        has source which
          simultaneously
            has name
              like `liver*'
            has name
              like `hepatic*'
            is a target for geneR-diseaseR which
              has source which
                has name
                  like `84419'
```

Figure 3.2. Query wizards: Find drugs associated with liver disease (A). Drug selected using CNV-based genes associated with liver disease (B). For example, Drug DB00316 (red) is influenced by Chemical MESH:D000082 (blue), and MESH:D000082 is caused by diseases MESH:D018248, MESH:D006528,

MESH:D056486, MESH:D008114 (green), and 4 diseases is associated with genes (yellow) such as 84419, 38115, 123591, and 89927.

Table 3.3. Seven drugs related to hepatic biochemical markers ALT or AST.

Drug name	Accession Number	Structure	Chemical Formula	Toxicity
Acetaminophen	DB00316		$C_8H_9NO_2$	Acetaminophen is metabolized primarily in the liver.
Chlormezanone	DB01178		$C_{11}H_{12}ClNO_3S$	Symptoms of overdose include liver damage.
Stavudine	DB00649		$C_{10}H_{12}N_2O_4$	Side effects include severe liver enlargement, inflammation of the liver, and liver failure.
Enflurane	DB00228		$C_3H_2ClF_5O$	Symptoms of chronic overdose include liver dysfunction.
Isoniazid	DB00951		$C_6H_7N_3O$	Adverse reactions include abnormal liver function tests.

Mebendazole	DB00643		$C_{16}H_{13}N_3O_3$	Symptoms of overdose include elevated liver enzymes.
Nitisinone	DB00348		$C_{14}H_{10}F_3NO_5$	Side effects include hepatic and liver failure.

3.4.3 Discovery of ethnic specific SNP-based genes

I identified ethnic specific SNPs by eliminating the overlapped SNPs from the HapMap samples (CEU, JPT, and YRI), and mapped the SNPs positions to the UCSC RefGene lists. As the result, 22, 25, and 332 genes were identified in the CEU, JPT, and YRI individuals, respectively (Figure 3.3; Supplementary Table 3.2). Comparison of the three sets showed that YRI individuals had a biased order of SNP-based genes. This result was a consensus among previous evolutionary findings. CEU and JPT belong to the same cluster, together with Amerindians and Australopapuanr, while YRI belongs to a separate cluster showing the first split between Africans and non-Africans (Nei and Roychoudhury 1993; Prugnolle et al. 2005). African populations subdivided from other sub-Saharan African populations, and a small subset of this population migrated out of Africa in the past 100,000 years. African and non-African populations divided in the past 40,000 years. Phylogenetic analysis of Y chromosomal haplotypes, mtDNA, and autosomes are indicative of the longest history of population subdivision in Africa. Africans are the most ancestral population in human and have fewer sites in linkage disequilibrium (LD), compared with non-African populations (Tishkoff and Williams 2002).

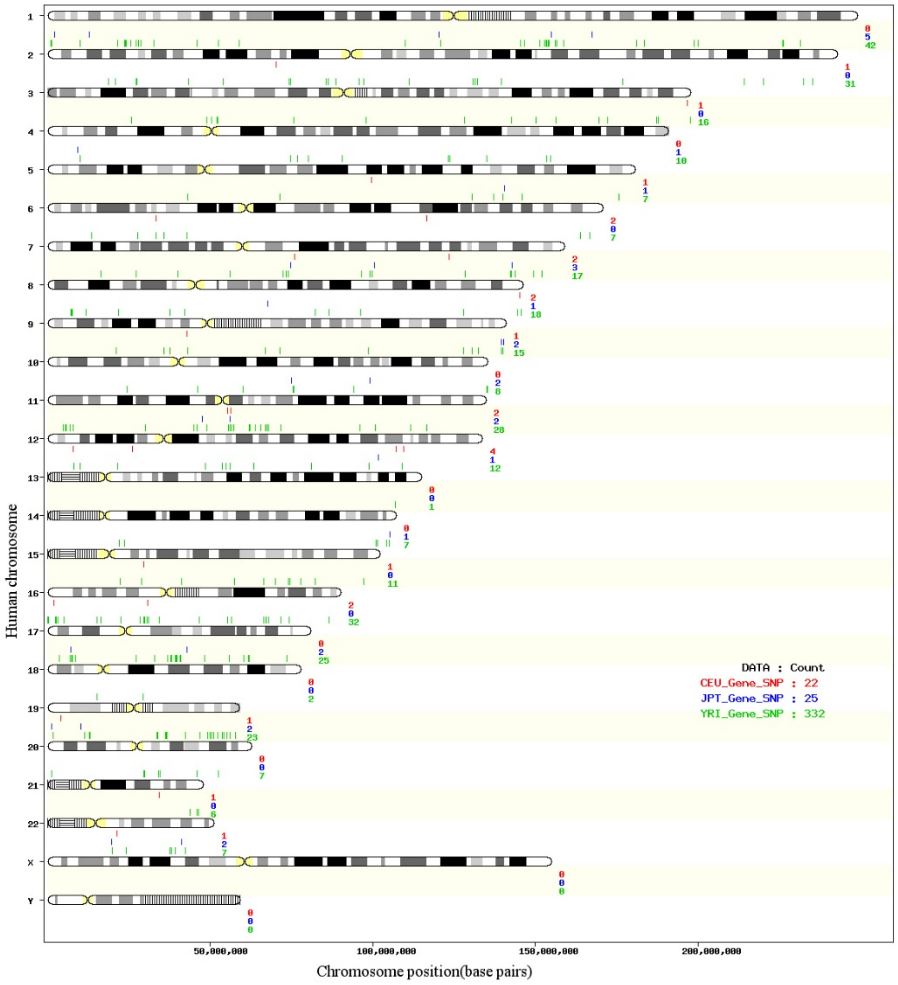


Figure 3.3. Visualization of the physical location for the ethnic specific genes from HapMap samples (CEU: red, JPT: blue, and YRI: green). The horizontal axis is the genomic location and the vertical axis is the number of chromosomes. The colored figure shows a total number of ethnicity-specific SNPs on the chromosome.

To explore the meaningful biological information of structural variations, I analyzed gene set enrichment analysis (GSEA) for the SNP-based genes using the GO categories (biological process (BP), cellular component (CC), and molecular function (MF)) in DAVID tool. The significantly categorized functions (*p*-value < 0.01) of SNP-based genes for YRI are shown as pie charts, but none was significantly enriched for CEU and JPT. Six groups of BP and four groups of MF had with the significant enrichment score have ranges of 1.67~4.85 and 1.9-5.05, respectively (Supplementary Figure 3.1). The top pie chart in biological process presents G-protein coupled receptor protein signaling pathway, including chemotaxis, and defense response to bacterium (Figure 3.4. (A)). In the enriched region, 8% of BP was chemotaxis (GO:0006935) with an enrichment score of 3.88. Chemotaxis contributes to enhancement of disease aggressiveness in African-Americans (Martin et al. 2009). The molecular functions that were significantly enriched were G-protein coupled receptor activity, binding olfactory receptor activity, and transmembrane receptor activity (Figure 3.4. (B)). Enriched functions in cellular components were keratin filament (GO:0045095) with an enrichment score of 5.86, which contained the KRTAP gene family (*KRTAP12-3*, *KRTAP4-11*, *KRT14*, *KRTAP4-4*, *KRTAP9-8*, *KRTAP10-7*, and *KRTAP10-8*). KRTAP family genes that are up-regulated in white hair than in black hair by a microarray analysis. Immunoreactivity for KRTP genes in white hair follicles was increased compared with black hair. Therefore, Choi et al. (2011)

suggested that hair greying hair, a sign of ageing, is associated with hair growth rate (Choi et al. 2011).

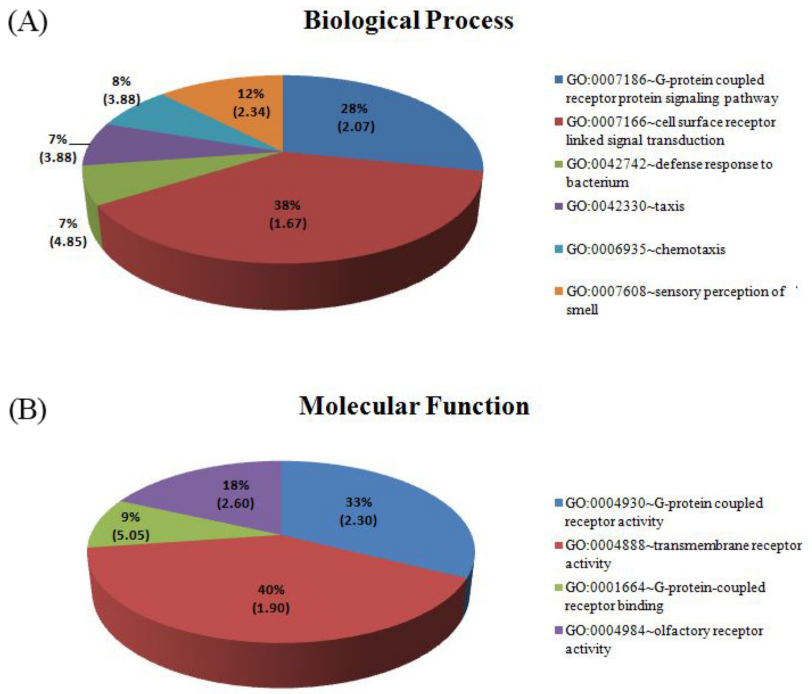


Figure 3.4. Gene Ontology enrichment analysis for YRI-specific SNP-based genes. (A) Biological Process and (B) Molecular Function.

3.4.4 Semantic findings using ethnic disparities

To show the biological knowledgeable diseases or drugs associated with ethnic disparities, I curated "SNP-Gene-Disease-Chemical-Drug" interactions in the ethnic disparities network. Figure 3.5 shows the Venn diagrams of the number of disease, drug, and pathway associated with ethnic disparities (Supplementary Table 3.3). Using these semantic "Gene-Disease" networks, I analyzed the functional implications of ethnic variants. There were 123 diseases associated with ethnic specific SNPs in common populations, 3 in CEU-specific, and 46 in YRI-specific, but JPT had no specified disparity between different ethnic populations (Figure 3.5. (C)).

Table 3.4 summaries of the functions associated with ethnic disparities in previous studies. Three diseases associated with CEU-specific SNPs were shown as Pantom Limb (MESH:D010591), Trochlear Nerve Diseases (MESH:D020432) and Vulvitis (MESH:D014847), while diseases associated with YRI-specific SNPs were observed such as AIDS-associated Nephropathy, hypertension, primary amyloidosis and pelvic infection. By applying the "SNP-Gene-Disease-Chemical-Drug" modeling, 2 and 14 drugs were revealed with CEU-specific and YRI-specific groups, but JPT-specific drugs had no results (Figure 3.5. (B)). Analysis using the semantic modeling for ethnicity-specific SNPs identified 5, 7, and 100 CEU-specific, JPT-specific and YRI-specific biochemical pathways, respectively (Figure 3.5 (A)). In the current study, the pathways shared between all populations were followed by signal transduction (REACT:111102), olfactory transduction (KEGG:04740), and metabolic pathways (KEGG:01100). Theses pathways were reported the common disease-pathway interactions in previous studies.

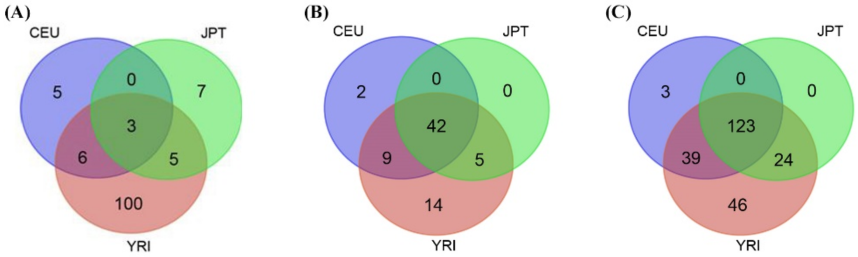


Figure 3.5. The Venn diagrams of ethnic disparities for pathways (A), drugs (B), and diseases (C) between CEU, JPT, and YRI.

Table 3.4. Summaries of the three diseases, one drug, and five pathways associated with the ethnic disparities in previous studies.

	Name	ID	Definition
Disease	AIDS-Associated Nephropathy	MESH:D016263	Renal syndrome in human immunodeficiency virus-infected patients characterized by nephrotic syndrome.
	Hypertension	MESH:D006973	Persistently high systemic arterial blood pressure.
	Pelvic Infection	MESH:D034161	Infection involving the tissues or organs in the pelvic.
Drug	Methylphenidate	DB00422	For use as an integral part of a total treatment program which typically includes other remedial measures for a stabilizing effect in children with a behavioral syndrome characterized by the following inappropriate symptoms.
Pathway	Hemostasis	REACT:604, REACT:82403, REACT:82812, REACT:85674, REACT:89750, REACT:92318	
	Systemic lupus erythematosus	KEGG:05322	A prototypic autoimmune disease characterised by the production of IgG autoantibodies.
	Prostate cancer	KEGG:05215	A major health problem in Western countries.
	Hepatitis C	KEGG:05160	A major cause of chronic liver disease.

Rheumatoid arthritis	KEGG:05323	A chronic autoimmune joint disease where persistent inflammation affects bone remodeling leading to progressive bone destruction.

3.5 Discussion

As reported an important role of structural variations, CNVs and SNPs have become a more attractive field (Lee et al. 2012). Differences of copy number between individuals contribute to alter in expression of genes sensitive to a disease susceptibility or dosage effect (Redon et al. 2006). Liver function test is blood tests to evaluate about patient's liver state (Thapa and Walia 2007). Aspartate aminotransferase (AST) and Alanine aminotransferase (ALT) are an important biochemical markers for evaluating of inflammation degree about liver injury (Ruhl and Everhart 2012). Therefore, I focused on constructing on liver functions or ethnic disparities.

Semantic biological network is an emerging method for comprehensively understanding of the complicated biological processes and spacious networks (Losko and Heumann 2009). The continuous production of increasingly large-scale data in biology field needs for better visualizations of complex and biological big data. I constructed semantic networks for liver functions or ethnic disparities using BioXM Knowledge Management Environment software (http://www.biomax.com). The software efficiently modeled such complex and metadata study, and enables researchers to create knowledgeable networks with flexible workflows for handling big data (Losko et al. 2006). This semantic biological networks provides comprehensive and easy to use resource. Also it enables the retrieval of relationship networks such as Gene-GO, Gene-Pathway, Gene-Disease, Gene-Chemical, Gene-SNP, Gene-

CNV, Gene-SomaticMutation, Pathway-Chemical, Pathway-Disease, Chemical-Drug, ClinicalTrials-Disease, and ClinicalTrials-Drug. The configuration-based approach to semantic integration network is closing the gap between public and experimental data. Recently, two studies of semantic biological networks have been published, which finding molecular signature of chemical l (Shin et al. 2012) and managing toxicogenomic laboratory experiment. This work supports to build such as Gene-Disease-Chemical-Drug relationship.

I investigated gene functional classification about liver functions or ethnic disparities using the semantic networks. For liver function network, the significant results showed the four diseases (hepatocellular carcinoma, liver neoplasm, liver cell adenoma, and drug-induced liver injury), one pathway (hepatitis C pathway), and seven drugs (acetaminophen, chlormezanone, stavudine, enflurane, isoniazid, mebendazole, and nitisinone). Liver Cell Adenoma is a benign neoplasm occurred from liver cell (hepatocytes) (Leese et al. 1988), occur most often in young women (Edmondson et al. 1976). It is important to recognize since it can be advanced a hepatocellular carcinoma (http://www.medicalgeek.com/). Liver Neoplasm is same name for liver (hepatic) cancer, and is an abnormal liver tissue (http://www.rightdiagnosis.com). Hepatocellular carcinoma (HCC) is the most common liver cancer. It occurs most often in men than women (Beasley et al. 1981), and is usually seen in people 50 years of age or older (http://www.nlm.nih.gov). This cancer in Africa and Asia is more common than

North or South America and Europe (Bressac et al. 1991). Drug is very clinically important cause of liver injury. Many drugs have been reported to cause liver (hepatic) injury (Lee 2003). Stavudine is an antiviral medication, which active against human immunodeficiency virus infection (Sommadossi 1995), and isoniazid is an antibiotic, which resists tuberculous bacteria (TB) (Sommadossi 1995), and mebendazole is an anti-worm medication, which used for prevents infections of such as pinworm, round worm, and hookworm (Sommadossi 1995). Nitisinone is used to treat hereditary tyrosinemia type 1 (Santra and Baumann 2008). These drugs keep causing harm to liver tissue and treated to cause liver injury.

Diseases and drugs are very clinically important for understanding ethnic disparities. Many diseases and drugs have been reported to be involved in ethnic disparities, disease susceptibility, drug response, and disposition (May 1994; Dransfield and Bailey 2006). For ethnic disparities network, the significant results reveal three diseases ("AIDS-Associated Nephropathy", "Hypertension", and "Pelvic Infection"), one drug ("Methylphenidate"), and five pathways ("Hemostasis", "Systemic lupus erythematosus", "Prostate cancer", "Hepatitis C virus", and "Rheumatoid arthritis"). AIDS-associated Nephropathy (AIDSAN, MESH:D016263) incidence rates are higher in African-Americans than whites. Although the mortality and morbidity from AIDS infection are reduced, AIDSAN remains a major complication of AIDS infection (http://statgen.ncsu.edu/). Hypertension (MESH:C537095) is a disease threatening the public health in sub-Saharan Africa. In some areas, blacks exhibit higher rates of hypertension than whites. Increased salt intake and obesity are the leading causes of the prevalence of hypertension in Africa (Addo et al. 2007). Pelvic Infection (MESH:D034161) is a kind of inflammatory disease that blacks are more prone to take than other ethnic groups (Eifel et al. 2002).

One drug (Methylphenidate, DB00422) was reported to have ethnic disparities in previously drug studies. The mean dose of methylphenidate is was about 1.5 times higher in the African-American than the Whites (Starr and Kemner 2005), and its use is steadily increasing in South Africa (Truter 2005).

In Hemostasis (REACT:604) associated with cardiovascular diseases, the plasminogen activator inhibitor-1 activity levels of Africans are lower compared to the Caucasians. These negative effects can be seen already at a young age. If addressed in early life, it is possibly adjustable through behavior and optimal dietary changes (Pieters and Vorster 2008). Systemic Lupus Activity Meaure (SLAM) (KEGG:05322) scores were higher in African-Americans (mean = 12.6) and Hispanics (11.0) than in Caucasians (8.5). It caused lack of health insurance, onset of abrupt disease, presence of anti-Ro (SSA) antibody, absence of HLA-DRB, high levels of helplessness, and abnormal illness behaviors. Caucasians lived under less crowded conditions, had less abnormal illness behaviors, and had more education. The results of the regression analyses were showed significant association between higher SLAM scores and higher helplessness, absence of HLA-DRB1*0301, and presence of HLA-DRB*0201 (p-value < 0.01) (Alarcón et al. 1998). Prostate cancer

(KEGG:05215) is a diagnosed male reproductive system cancer. Incidence of prostate cancer in African-Americans men is higher than in the European men (1.6 times). Amundadottir et al. (2006) identified that the chromosomal 8q24 region is most frequently gained in prostate cancers and this gained region has been correlated with aggressive tumors (Amundadottir et al. 2006). Estimated population attributable risk (PAR) is greater in Africans than in European populations. Hepatitis C virus (HCV, KEGG:05160) is a major cause of chronic liver disease in humans. Rates of HCV prevalence in sub-Saharan Africa are the highest in central African (3.0%) compared with the median (2.2%). Conjeevaram et al. (2006) showed that African-Americans with chronic HCV have lower response to interferon-based antiviral therapy than Caucasian Americans (Madhava et al. 2002; Conjeevaram et al. 2006). Rheumatoid arthritis (RA, KEGG:05323) is an autoimmune disease and may affect many organs. The RA prevalence in urban South Africans is similar to in Caucasians (Solomon et al. 1975).

Also, 1 common pathway between all populations was showed. Although ethnicity-specific genes are identified in each population, it is generally observed that genes that are associated with a trait or disease can converge to the same pathway (Fu et al. 2011). Those genes are also supposed to converge to common pathways shared between all populations. Therefore, a pathway-based approach allows us to systematically evaluates multiple polymorphic genes from different populations with respect to pathways as a biological unit (Wang et al. 2007a). Moreover, the pathway-based approach has more capability to detect rare genetic variants with a small effect that do not survived at the stringent significance level (Medina et al. 2009).

I constructed semantic networks for liver functions or ethnic disparities. Functional studies were analyzed with CNV-based genes associated with liver functions or ethnic specific SNP-based genes. These semantic networks showed robust interactions between liver-related to CNVs or ethnic specific SNPs and public data. I expect that the semantic networks are useful for liver functions or ethnic specific SNPs, and the findings will provide prioritization of ethnic specific SNP-based candidate genes. Also, I will constantly develop more robust and flexible algorithms.

This chapter was published in *Asian-Austraiasian Journal of Animal Sciences*.
as a partial fulfillment of HyoYoung Kim's Ph.D program.

Chapter 4. VCS: tool for visualizing Copy number variation and Single nucleotide polymorphism

4.1 Abstract

Copy number variation (CNV) or single nucleotide phlyorphism (SNP) is useful genetic resource to aid in understanding complex phenotypes or susceptibility or resistence to deseasess. Although thousands of CNVs and/or SNPs are currently avaliable in the public databases, they are somewhat difficult to use for analyses without visualization tools. Visualization of CNV and/or SNP can assist to easily interpret a biological meaning from the numerical value of CNV or SNP. Here I developed a web-based tool called VCS (the Visualization of Copy number variation and Single nucleotide polymorphism) to visualize the CNV and/or SNP detected in different animals such as mammals, vertebrates, insects, and worms. The VCS provides six different visualization tools: (ⅰ) the enrichment of genome contents in CNV; (ⅱ) the physical distribution of CNV or SNP on chromosomes; (ⅲ) the distribution of log2 ratio of CNVs with criteria of interested; (ⅳ) number of CNV and SNP per binning unit (10 kb, 100 kb, 1Mb, and 10Mb); (ⅴ) the distribution of homozygosity of SNP genotype on chromosomes; and (ⅵ) cytomap of genes within CNV or SNP region. VCS application is available from http://snugenome.snu.ac.kr/Software/VCS/ and executable examples can be downloaded from the same web site as well. The VCS was implemented as a program written in PHP (ver.5.3), mysql (ver. 5.1.36), and Python (ver.2.5).

VCS use it for free, this tool is user friendly and more offer directly insertable tip-top figures in thesis.

4.2 Introduction

In genomic research, copy number variation (CNV) and single nucleotide polymorphism (SNP) are used to identify the association with complex phenotypes or susceptibility or resistance to diseases (Fanciulli et al. 2007; Yang et al. 2007). CNV encompasses more DNA than SNP and contains entire genes and their regulatory region (Freeman et al. 2006). The type of genetic variant can influence gene dosage other than phenotypic variation, which might cause genetic diseases. A series of studies using CNV and/or SNP were performed to detect the association with different cancer cells or complex diseases (Diskin et al. 2009; Shlien and Malkin 2009). Development of whole genome sequencing projects of different organisms and the current improvement in biotechnologies have contributed to the detection of enormous numbers of SNP and CNV in each species. Thousands of CNV or SNP are currently available in the public databases, but it is not so easy for local researchers to use them for their own analyses. Information regarding CNV and/or SNP in general consists of numerical values which are difficult to understand and to interpret biologically. Visualization of the data may assist researchers to interpret biological meanings from the numerical value, even though it is not a necessary step for the analyses. However, few visualizing software have been reported for CNV and/or SNP. In this study, I developed a web-based visualization tool graphically representing the enrichment of genome contents in CNV, the distribution of CNV and/or SNP on chromosomes, the log2 ratio of fluorescence intensities of CNV, the homozygosity of SNP on chromosomes, and cytomapping of the genes of interest.

4.3 Program overview

I developed a web-based tool called VCS (the Visualization of CNV and SNP) to picture the data of your CNV and/or SNP in the genome. The pictures can help not only to interpret a biological meaning from the numerical value of CNV or SNP but also provide the figures for user's manuscript. VCS tool provides a graphical view of the physical distribution of CNV or SNP on chromosomes. Although several web databases have reported annotated CNV (e.g. Database of Genomic Variants (DGV; http://projects.tcag.ca/variation/), dbSNP 131 (http://www.ncbi.nlm.nih.gov/; (Smigielski et al. 2000)), GWAS CENTRAL (http://www.gwascentral.org/ (Fredman et al. 2002), and SNP and CNV Annotation Database (SCAN; http://www.scandb.org/) (Gamazon et al. 2010) or CNV extraction software (e.g. PennCNV (Wang et al. 2007b), Aroma.Affymetrix (Bengtsson et al. 2008), CRLMM (Scharpf et al. 2010), and Affymetrix Power Tools (Lockstone 2011)), it is often difficult to apply them one's own result. Main features of VCS are as follows:

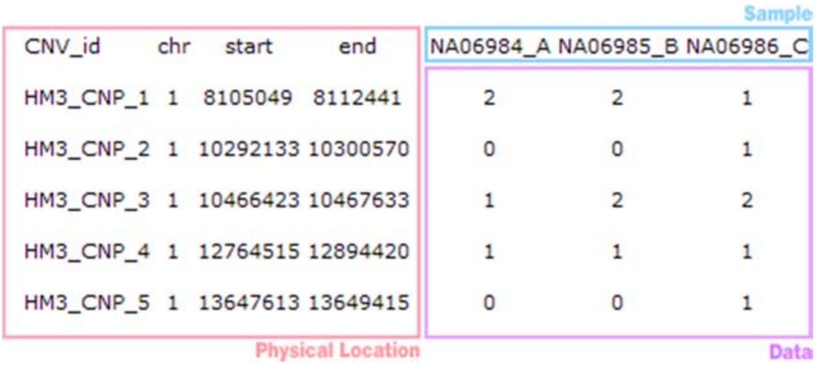
4.3.1 Visualization of the enrichment of genome contents in CNV regions

VCS shows the enrichment genome contents (gene, LINE (long interspersed nuclear element), SINE (short interspersed nuclear element), LTR (long terminal repeat), simple repeat, low complexity, miRNA, tRNA, CpG island,

and Gene Ontology – Biological Process, Molecular Function, and Cellular Component) in region having specific range such as CNV. For cluster analysis, the distance matrix was produced by Hamming distance computation considering deletion and duplication of copy number (Steane 1996). Then the hierarchical cluster and principal component analysis (PCA) were performed using the distance matrix. As the result, user can easily show the nearest clustered samples about the genome content within CNV region.

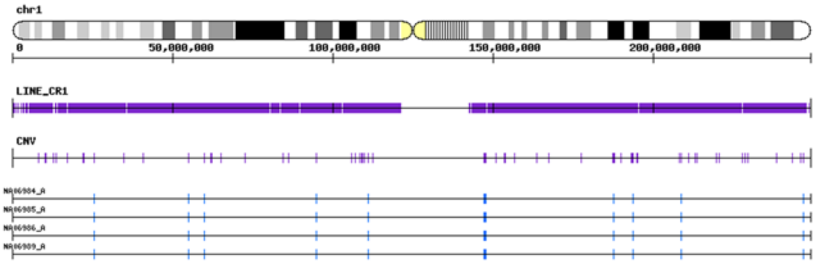
The input file needs matrix format file formed 0, 1, 2, 3, 4, .. (Figure 4.1. (A)). Here, 0 and 1 is deletion and more than 2 is duplication. The figure represents as user-defined such as deletion or insertion. So the user can show the enrichments result figure and table of genome content in a specific region (Figure 4.1. (B), (C)), and show hierarchical clustering (Figure 4.1. (D)) and PCA cluster (Figure 4.1. (E)) among samples. In addition, user can display all the genome contents per sample. If user denotes groups as _A, _B, _C in input file, user can easily and clearly show clusters as editing the grouping image using other graphic tool such as Adobe photoshop or illustrate.

(A)

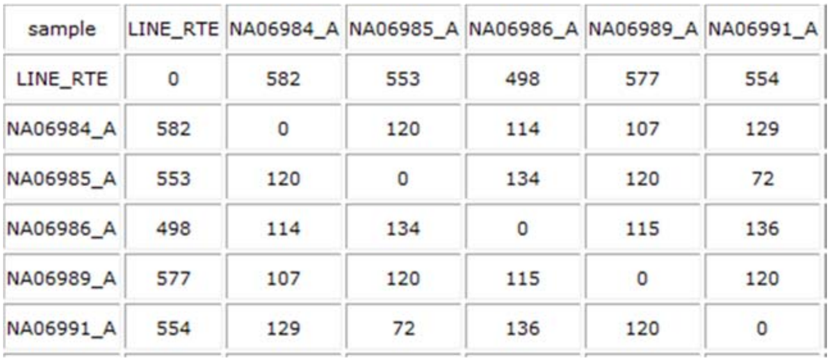

Sample
Physical Location
Data

CNV_id	chr	start	end	NA06984_A	NA06985_B	NA06986_C
HM3_CNP_1	1	8105049	8112441	2	2	1
HM3_CNP_2	1	10292133	10300570	0	0	1
HM3_CNP_3	1	10466423	10467633	1	2	2
HM3_CNP_4	1	12764515	12894420	1	1	1
HM3_CNP_5	1	13647613	13649415	0	0	1

(B)



(C)



sample	LINE_RTE	NA06984_A	NA06985_A	NA06986_A	NA06989_A	NA06991_A
LINE_RTE	0	582	553	498	577	554
NA06984_A	582	0	120	114	107	129
NA06985_A	553	120	0	134	120	72
NA06986_A	498	114	134	0	115	136
NA06989_A	577	107	120	115	0	120
NA06991_A	554	129	72	136	120	0

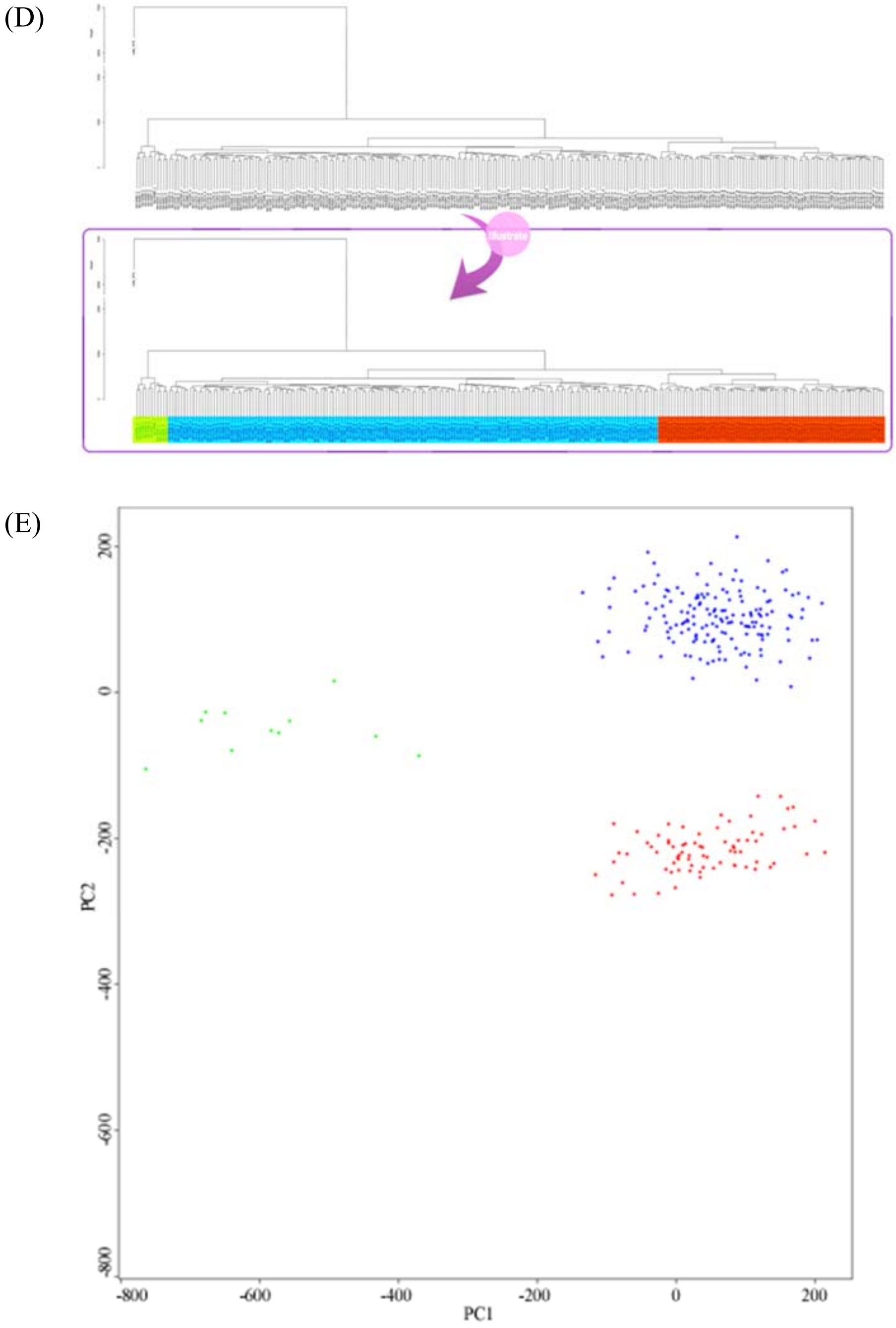
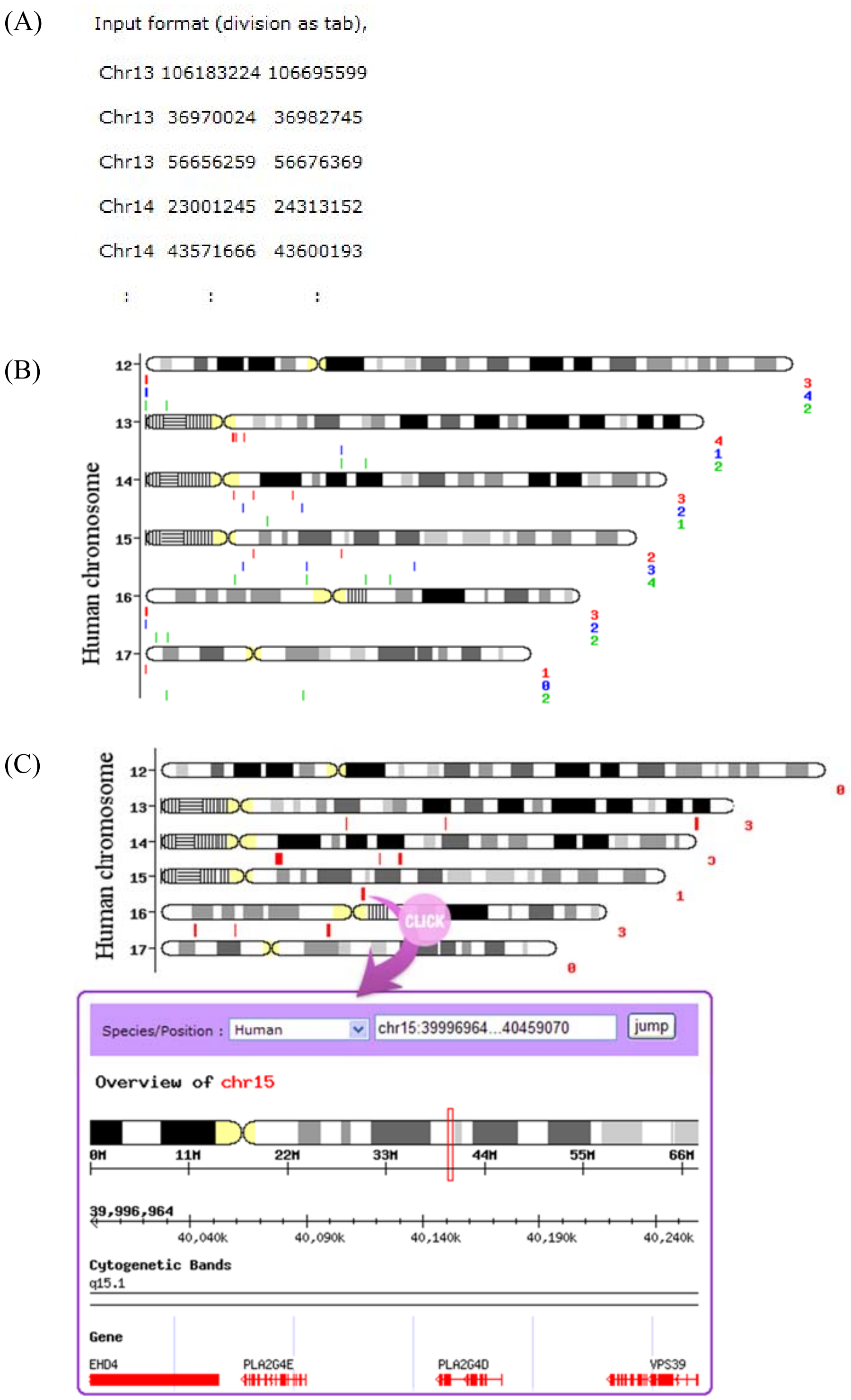

(D)
Illustrate
(E)
PC2
200
0
-200
-400
-600
-800
PC1
-800
-600
-400
-200
0
200

Figure 4.1. Visualization of the enrichment of genome contents in CNV region. (A) Input matrix data with the information of physical location and figure (deletion and insertion) after CNV analysis; Gives the following output is a enrichments result figure (B), a distance matrix (C), a hierarchical clustering (D), and a PCA cluster (E) of genome content in specific region.

4.3.2 Physical distribution visualization

VCS provides a graphical distribution on chromosomes. Any marker contained information of chromosomal position by point (SNP) or specific ranges (CNV, miRNA, and repeat sequence) can be used in this tool. This menu is useful for comparing the physical distribution of your own CNV or SNP. In addition, comparison among samples is available by adding input files up to five (Figure 4.2. (B).

The input file simply needs the information of chromosome number and chromosomal position of either CNV or SNP (Figure 4.2. (A). After your data are loaded on the website, you can obtain the information in detail on the genome where the CNV (SNP) is located by clicking it (Figure 4.2. (C)). User can take a look at the information on genes, and repeat sequences such as SINE, LINE, LTR, and simple repeat around the CNV.



(A)

Input format (division as tab),

```
Chr13 106183224 106695599
Chr13  36970024  36982745
Chr13  56656259  56676369
Chr14  23001245  24313152
Chr14  43571666  43600193
  :        :         :
```

(B)

(C)

Species/Position : Human chr15:39996964...40459070 jump

Overview of chr15

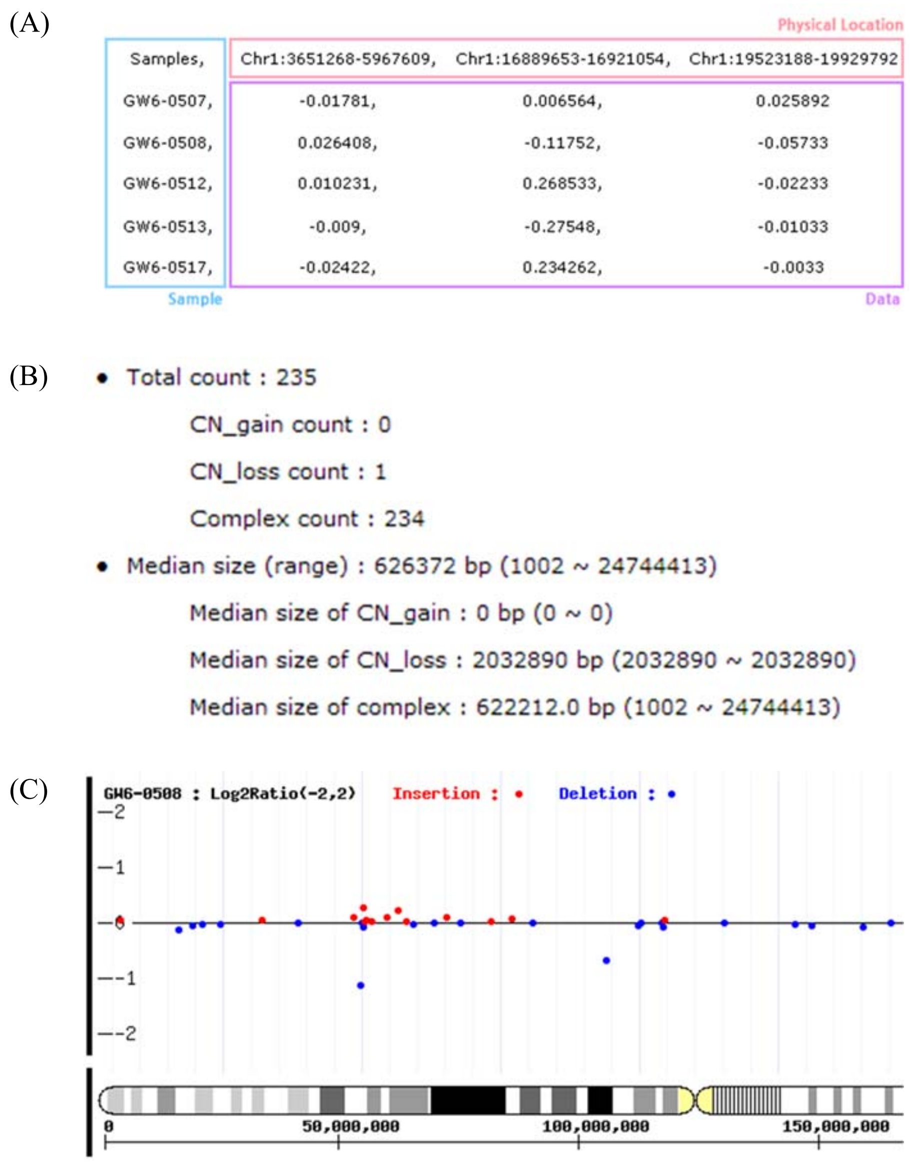
Cytogenetic Bands
q15.1

Figure 4.2. Visualization of the physical distribution for specific position or region. (A) Data with the information of chromosome number and physical location; (B) By clicking the physical location where the CNV (SNP), you can obtain the information in detail on the genome; (C) Comparing among samples is available by adding input files up to five.

4.3.3 Log2 ratio distribution visualization

VCS plots log2 ratio of CNV with insertions and deletions that are more conspicuous. The log2 values are plotted at the middle position of CNV regions across the chromosome. Several web databases represent the whole log2 ratio (e.g. Affymetrix Genotyping Console Browser), but VCS can provide the criteria which is the user-adjustable log2 ratio. So a user can create the view of CNV filtrated by adjusting the criteria with different log2 ratio values for different research purposes. In addition, user can draw a Manhattan plot which easily can define appropriate significance value, and can perform the comparison among samples selected in this menu.

The input file needs matrix format data with the information of physical location and the value of plus (+) or minus (-) such as log2 ratio after CNV analysis ((Figure 4.3. (A)). VCS then gives the following output as user-defined criteria, from which you can obtain total counts and median size of gain (insertion), loss (deletion), and complex (insertion and deletion) (Figure 4.3. (B)). Default of a criteria set up ±0.3 which is widely used in biology research. And you can show distribution of visualized log2 ratio and/or ± values (Figure 4.3. (C)).



Samples,	Chr1:3651268-5967609,	Chr1:16889653-16921054,	Chr1:19523188-19929792
GW6-0507,	-0.01781,	0.006564,	0.025892
GW6-0508,	0.026408,	-0.11752,	-0.05733
GW6-0512,	0.010231,	0.268533,	-0.02233
GW6-0513,	-0.009,	-0.27548,	-0.01033
GW6-0517,	-0.02422,	0.234262,	-0.0033

- Total count : 235

  CN_gain count : 0

  CN_loss count : 1

  Complex count : 234

- Median size (range) : 626372 bp (1002 ~ 24744413)

  Median size of CN_gain : 0 bp (0 ~ 0)

  Median size of CN_loss : 2032890 bp (2032890 ~ 2032890)

  Median size of complex : 622212.0 bp (1002 ~ 24744413)

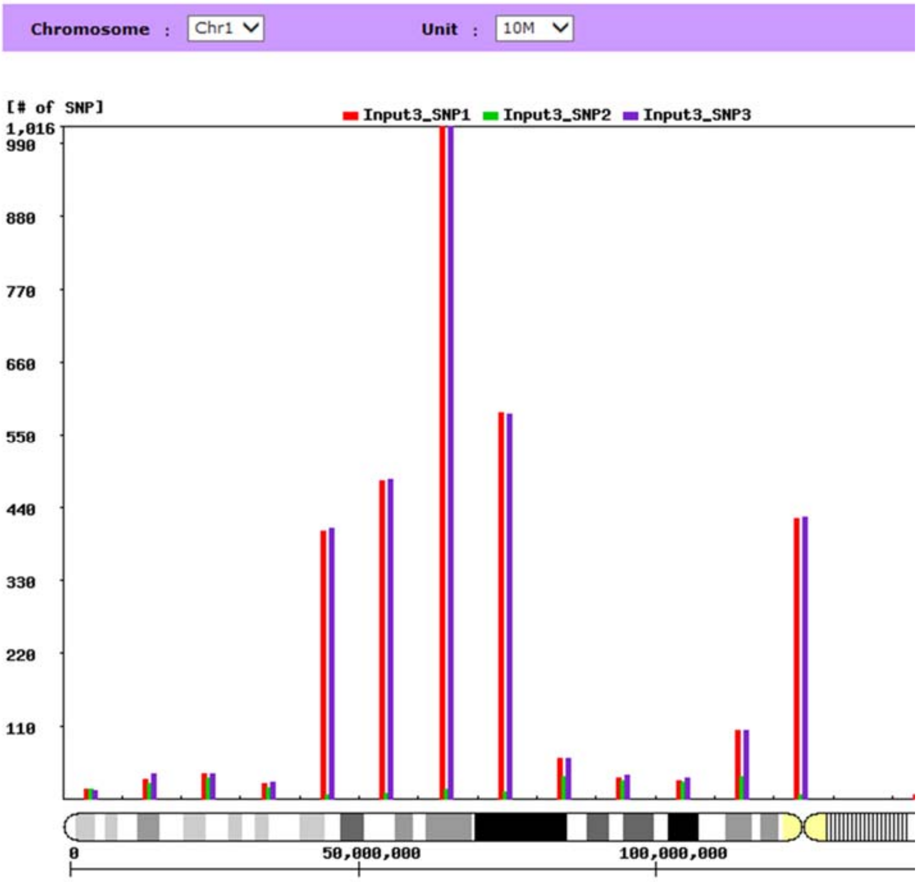
Figure 4.3. Visualization of the distribution of log2 ratio. (A) Input matrix data with the information of physical location and the value of plus (+) or minus (-) such as log2 ratio; (B) Gives the following output 1 is a total counts and median size of gain (insertion), loss (deletion), and complex (insertion and

deletion); (C) Gives the following output 2 is a distribution of visualized log2 ratio and/or ± values, with red (insertion) and blue (deletion) marks.

4.3.4 Variation distribution visualization per binning unit

VCS calculates the number of CNV or SNP per binning units of 10 kb, 100 kb, 1 Mb, and 10 M. The goal of this menu is to look at the number of variants within the certain ranges of physical distances, which allows researchers to take advantage of deciding or selecting the scale of the study area they want to focus on. Also, this menu is useful for comparing the numbers per binning unit by adding more data. The user selects binning unit by simply clicking on the appropriate criteria.

The input file is the same input file with the information of chromosome number and chromosomal position of either CNV or SNP used for the physical location. You can show the visualized distribution per binning unit and decide concentrated study region on genome (Figure 4.4).



Chromosome : Chr1

Unit : 10M

Figure 4.4. Visualization of the distribution of SNP numbers per binning unit 10 kb, 100 kb, 1Mb, and 10 Mb on chromosome.

4.3.5 Homozygosity distribution visualization for SNP genotypes

VCS shows the homozygosity of SNP on chromosomes by using the information of SNP genotypes of samples, chromosomal position of SNP and chromosome number. This menu is useful when comparing homozygosity among samples. VCS calculates homozygosity of all SNP located on an entire chromosome of interest and plots homozygosity of every unit of 100 SNPs along the chromosomes. At the end of the chromosome, the number of SNP is usually less than 100 which is added to the previous unit if the number of SNP is ≤ 50 or is calculated as another unit if it is > 50.

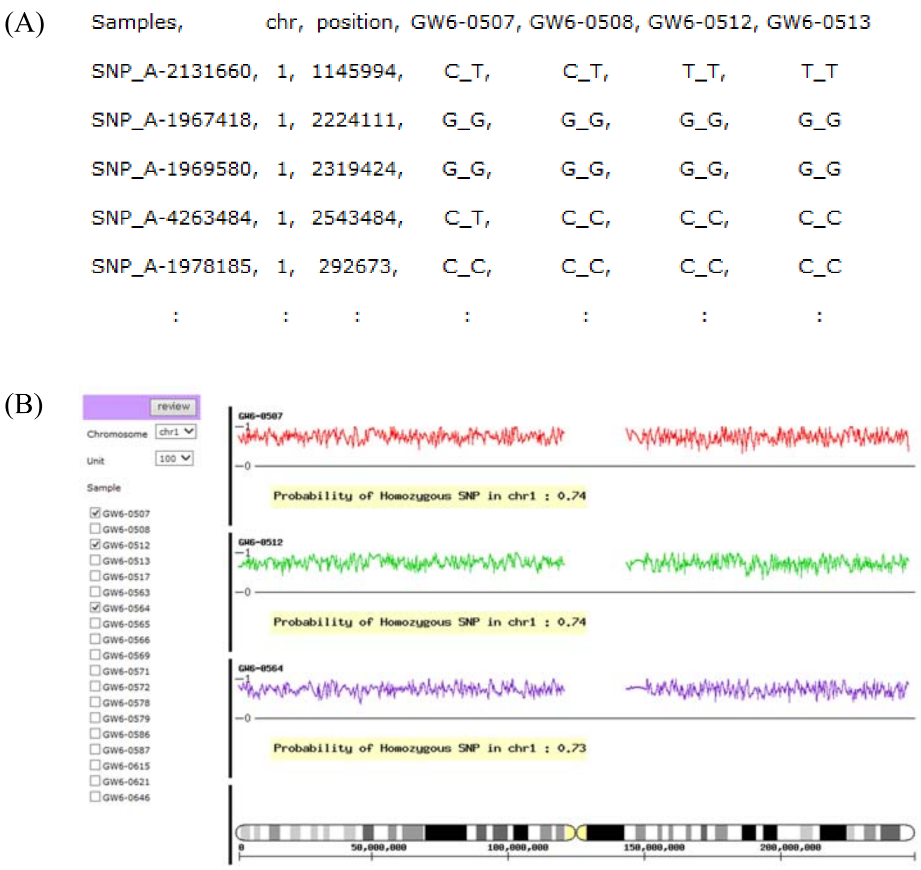
For n-100(k-1) > 50,

$$\frac{1}{k}\left[\frac{1}{100}\sum_{j=1}^{k-1}\sum_{i=1}^{100} y_{ij} + \frac{1}{n-100(k-1)}\sum y_{ik}\right]$$

For n-100(k-1) ≤ 50,

$$\frac{1}{k}\left[\frac{1}{100}\sum_{j=1}^{k-2}\sum_{i=1}^{100} y_{ij} + \frac{1}{100+[n-100(k-1)]}\left\{\sum_{i=1}^{100} y_i\,(k-1) + \sum_{i=1}^{n-100(k-1)} y_{ik}\right\}\right]$$

The input file requires the matrix data with information such as genotypes in SNP analysis (Figure 4.5. (A)). User can then display any area that has a low homozygosity value, and obtain the probability of the homozygous SNP on each chromosome (Figure 4.5. (B)).



(A)

```
Samples,          chr,  position,  GW6-0507, GW6-0508, GW6-0512, GW6-0513
SNP_A-2131660,  1,  1145994,     C_T,        C_T,        T_T,        T_T
SNP_A-1967418,  1,  2224111,     G_G,        G_G,        G_G,        G_G
SNP_A-1969580,  1,  2319424,     G_G,        G_G,        G_G,        G_G
SNP_A-4263484,  1,  2543484,     C_T,        C_C,        C_C,        C_C
SNP_A-1978185,  1,   292673,     C_C,        C_C,        C_C,        C_C
       :             :       :           :            :            :           :
```

(B)

review

Chromosome chr1

Unit 100

Sample

☑ GW6-0507
□ GW6-0508
☑ GW6-0512
□ GW6-0513
□ GW6-0517
□ GW6-0563
☑ GW6-0564
□ GW6-0565
□ GW6-0566
□ GW6-0569
□ GW6-0571
□ GW6-0572
□ GW6-0578
□ GW6-0579
□ GW6-0586
□ GW6-0587
□ GW6-0615
□ GW6-0621
□ GW6-0646

GW6-0507

Probability of Homozygous SNP in chr1 : 0.74

GW6-0512

Probability of Homozygous SNP in chr1 : 0.74

GW6-0564

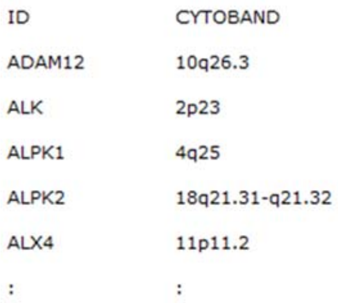
Probability of Homozygous SNP in chr1 : 0.73

Figure 4.5. Visualization of the distribution of homozygous SNP. (A) Input data with the information such as genotypes in SNP analysis; (B) Gives the following output is a distribution and probability of homozygous per number of SNPs. Irregular zig-zagged lines represent the homozygosity value per unit of 100 SNPs.

4.3.6 CytoMap

CytoMap provides the cytomapping figure of your focused-genes (Figure 4.6. (B)). The input file needs only the information of the cytoband of your focused-genes (Figure 4.6. (A)). There are several assembly versions of human genome sequences available in public databases such as NCBI (http://www.ncbi.nlm.nih.gov/) and UCSC (http://genome.ucsc.edu/). However, the physical positions of genes of interest are version-dependant. CytoMap provides a gene map by the cytoband position. This menu is useful for genome-wide view of data.

(A)



ID	CYTOBAND
ADAM12	10q26.3
ALK	2p23
ALPK1	4q25
ALPK2	18q21.31-q21.32
ALX4	11p11.2
:	:

(B)

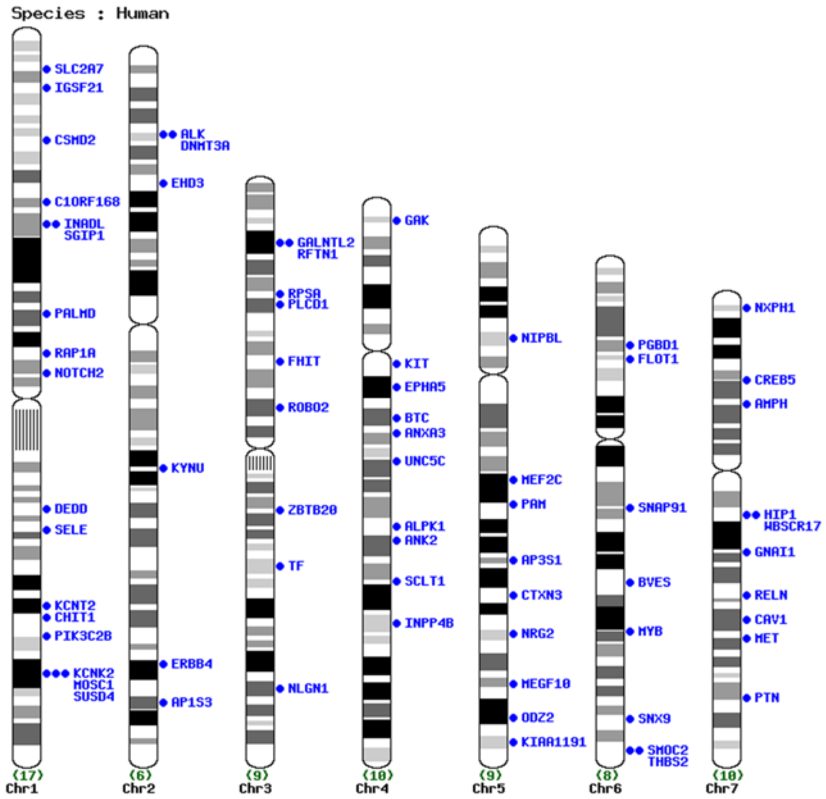


Figure 4.6. Visualization of the CytoMap for genes located in the CNV or SNP subregion. (A) Input data with the information of cytoband; (B) Visualization of the cytomap for focused-genes. Blue and green indicate ID of input file and the total number of IDs, respectively.

4.4 Implementation

VCS is built upon for visualizing data of CNV and/or SNP from local researcher. The VCS was implemented as a program written in PHP (PHP Hypertext Preprocessor; ver. 5.3), mysql (ver. 5.1.36), R (ver. 2.14), and Python (ver. 2.5). All of six menus have a common option to choose a pecies for the analysis. Animal species included in this study are human (hg19), rhesus (rheMac2), mouse (mm9), rat (rn4), dog (canFam2), horse (equCab2), cow (bosTau4), opossum (monDom5), chicken (galGal3), zebrafish (danRer7), D.melanogaster (dm3), and C.elegans (ce6). Genomic information of those species was downloaded from http://genome.ucsc.edu/.

By selecting a species from the pop-up menu, basic genomic information of the species such as total number of chromosomes and sizes of chromosome is set as a default for the analysis. Therefore, a user doesn't need to prepare the information in the input file regardless of any platform such as Affymetrix or Illumina for analysis of either CNV or SNP. The input file only needs the information of chromosomal position or CNV log2 ratio values or SNP genotypes or cytoband after variation analysis. For each menu, input file format take the divided by tabs or comma. For output file, you can select formats: png or bmp. Also user can edit the image using other graphic tool such as photoshop or illustrate.

A researcher who is interested in CNV or SNP can easily access the web site and use it for free without additional steps of downloading and installing it onto their local computer. This tool is user friendly and can be simply used without a thick user’s manual. To development of bioinformatics usages of the data served in VCS, I are continuously developing and updating. I expect to add tool associated with these CNVs and SNPs studies are merged into VCS.

General Discussion

By analyzing CNVs acquired from array-based genotyping, a lot of biological meanings could be obtained through genome-wide association study (GWAS), biological networks, and visualization for structural variations.

In chapter 2, GWA studies enable me to find the genes associated with the hepatic biochemical markers AST or ALT through the CNVs analyses. Many CNVs associated with liver disease have been reported in Caucasians, Africans, Chinese, and Japanese, but they may not properly reflect the CNVs in the genomes of other ethnic groups. Also, univariate linear regression is widely used to identify SNPs, but few studies have reported the statistical method to discover CNVs. Therefore, I used Korean chips as a reference group. Univariate linear regression was performed to examine the impact of single CNV regions for each quantitative trait. By using GWA study, I found that the significant genes associated with AST or ALT in KARE1. Then by the replication studies of GWA, I found the significant nine genes associated with hepatic biochemical markers AST or ALT of Koreans.

In chapter 3 and 4, two biological networks and a visualization tool were constructed. By using the biological semantic networks, I could investigated liver functions or ethnic disparities. The semantic biological networks enable me to create knowledgeable networks with flexible workflows for handling big data. The biological networks provide comprehensive and easy to use resource for human liver functions or ethnic disparities (chapter 3). I could easily interpret biological meanings from the numerical value of CNV or SNP using the visualization tool (chapter 4).

From the Korean cohort data, I could attain useful biological meanings associated with liver diseases, construct knowledegeable biological networks and the visualization tool for variations. The analysis of CNV/SNP-based genes is useful to understand biological phenotypes or diseases, and will provide valuable resources for further investigations of liver diseases. I expect that the biological networks for liver functions or ethnic disparities will provide valuable information and strongly affect useful knowledge. The visualization tool for variants will help interpret biological meanings from the numerical value.

References

Addo, J., L. Smeeth, and D. A. Leon. 2007. Hypertension In Sub-Saharan Africa. *Hypertension* 50 (6):1012-1018.

Aitman, T. J., R. Dong, T. J. Vyse, P. J. Norsworthy, et al. 2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* 439 (7078):851-855.

Alarcón, G. S., J. Roseman, A. A. Bartolucci, A. W. Friedman, et al. 1998. Systemic lupus erythematosus in three ethnic groups: II. Features predictive of disease activity early in its course. *Arthritis and Rheumatism* 41 (7):1173-1180.

Almuneef, M., S. Al Khenaizan, S. Al Ajaji, and A. Al-Anazi. 2003. Pyogenic liver abscess and Papillon-Lefevre syndrome: not a rare association. *Pediatrics* 111 (1):e85.

Altmüller, J., L. J. Palmer, G. Fischer, H. Scherb, et al. 2001. Genomewide scans of complex human diseases: true linkage is hard to find. *The American Journal of Human Genetics* 69 (5):936-950.

Amundadottir, L. T., P. Sulem, J. Gudmundsson, A. Helgason, et al. 2006. A common variant associated with prostate cancer in European and African populations. *Nature Genetics* 38 (6):652-658.

Armengol, C., S. Cairo, M. Fabre, and M. Buendia. 2011. Wnt signaling and hepatocarcinogenesis: the hepatoblastoma model. *The International Journal of Biochemistry and Cell Biology* 43 (2):265-270.

Atkinson, M. R., M. A. Savageau, J. T. Myers, and A. J. Ninfa. 2003. Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. *Cell* 113 (5):597-607.

Bathum, L., H. C. Petersen, J.-U. Rosholm, P. H. Petersen, et al. 2001. Evidence for a substantial genetic influence on biochemical liver function tests: results from a population-based Danish twin study. *Clinical Chemistry* 47 (1):81-87.

Bax, L., L. Yu, N. Ikeda, H. Tsuruta, et al. 2006. Development and validation of MIX: comprehensive free software for meta-analysis of causal research data. *BMC Medical Research Methodology* 6 (1):50.

Beasley, R. P., C. C. Lin, L. Y. Hwang, and C. S. Chien. 1981. Hepatocellular carcinoma and hepatitis B virus: a prospective study of 22 707 men in Taiwan. *The Lancet* 318 (8256):1129-1133.

Beckmann, J., X. Estivill, and S. Antonarakis. 2007. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics* 8 (8):639-646.

Bengtsson, H., K. Simpson, J. Bullard, and K. Hansen. 2008. aroma. affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Department of Statistics, University Of California, Berkeley* 1-79.

Björnsson, E. S., and N. Chalasani. 2010. Drug-Induced Liver Injury. *Practical Gastroenterology and Hepatology: Liver and Biliary Disease*:235-242.

Blaschke, T. Protein binding and kinetics of drugs in liver diseases. *Clinical Pharmacokinetics* 2 (1):32.

Block, T. M., M. A. Comunale, M. Lowman, L. F. Steel, et al. 2005. Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proceedings of The National Academy of Sciences of the United States of America* 102 (3):779-784.

Bown, M. J., G. T. Jones, S. C. Harrison, B. J. Wright, et al. 2011. Abdominal aortic aneurysm is associated with a variant in low-density lipoprotein receptor-related protein 1. *The American Journal of Human Genetics* 89 (5):619-627.

Bressac, B., M. Kew, J. Wands, and M. Ozturk. 1991. Selective G to T mutations of p 53 gene in hepatocellular carcinoma from southern Africa. *Nature* 350 (6317):429-431.

Brouwers, N., C. Van Cauwenberghe, S. Engelborghs, J. Lambert, et al. 2011. Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. *Molecular Psychiatry* 17 (2):223-233.

Buckland, P. R. 2003. Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Annals of Medicine* 35 (5):308-315.

Burton, P. R., D. G. Clayton, L. R. Cardon, N. Craddock, et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447 (7145):661-678.

Bush, W. S., and J. H. Moore. 2012. Genome-Wide Association Studies. *PLoS Computational Biology* 8 (12):e1002822.

Cappuzzo, F., F. R. Hirsch, E. Rossi, S. Bartolini, et al. 2005. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non–small-cell lung cancer. *Journal of the National Cancer Institute* 97 (9):643-655.

Cerami, E., E. Demir, N. Schultz, B. Taylor, et al. 2010. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* 5 (2):e8918.

Chance, P. F., M. K. Alderson, K. A. Leppig, M. W. Lensch, et al. 1993. DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell* 72 (1):143-151.

Cho, Y. S., M. J. Go, Y. J. Kim, J. Y. Heo, et al. 2009. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genetics* 41 (5):527-534.

Choi, H. I., G. I. Choi, E. Kim, Y. Choi, et al. 2011. Hair greying is associated with active hair growth. *British Journal of Dermatology* 165 (6):1183-1189.

Clarke, G. M., C. A. Anderson, F. H. Pettersson, L. R. Cardon, et al. 2011. Basic statistical analysis in genetic case-control studies. *Nature Protocols* 6 (2):121-133.

Clemenz, M., N. Frost, M. Schupp, S. Caron, et al. 2008. Liver-Specific Peroxisome Proliferator-Activated Receptor α Target Gene Regulation by the Angiotensin Type 1 Receptor Blocker Telmisartan. *Diabetes* 57 (5):1405.

Conjeevaram, H. S., M. W. Fried, L. J. Jeffers, N. A. Terrault, et al. 2006. Peginterferon and ribavirin treatment in African American and Caucasian American patients with hepatitis C genotype 1. *Gastroenterology* 131 (2):470-477.

Consortium, C. A. D. G. 2011. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature Genetics* 43 (4):339-344.

Cook Jr, E. H., V. Lindgren, B. L. Leventhal, R. Courchesne, et al. 1997. Autism or atypical autism in maternally but not paternally derived proximal 15q duplication. *American Journal of Human Genetics* 60 (4):928.

Cooper, G., J. Johnson, T. Langaee, H. Feng, et al. 2008. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 112 (4):1022.

Couzin-Frankel, J. 2010. Major heart disease genes prove elusive. *Science* 328 (5983):1220-1221.

Craddock, N., M. E. Hurles, N. Cardin, R. D. Pearson, et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464 (7289):713-720.

Davern, T. J. 2012. Drug-induced liver disease. *Clinics in Liver Disease* 16 (2):231.

Davidson, D., and W. Eastham. 1966. Acute liver necrosis following overdose of paracetamol. *British Medical Journal* 2 (5512):497-499.

Davis, A. P., C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, et al. 2009. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Research* 37 (suppl 1):D786-D792.

De Clercq, D., M. Sacko, J. Behnke, F. Gilbert, et al. 1997. Failure of mebendazole in treatment of human hookworm infections in the southern region of Mali. *American Journal of Tropical Medicine and Hygiene* 57:25-30.

de Leeuw, N., T. Dijkhuizen, J. Y. Hehir-Kwa, N. P. Carter, et al. 2012. Diagnostic interpretation of array data using public databases and internet sources. *Human Mutation* 33 (6):930-940.

de Stahl, T., J. Sandgren, A. Piotrowski, H. Nord, et al. 2008. Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC clone based array. *Human Mutation* 29 (3):398-408.

Delgado, J. C., A. Baena, S. Thim, and A. E. Goldfeld. 2002. Ethnic-specific genetic associations with pulmonary tuberculosis. *Journal of Infectious Diseases* 186 (10):1463-1468.

Dennis Jr, G., B. T. Sherman, D. A. Hosack, J. Yang, et al. 2003. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biolology* 4 (5):P3.

Ding, L., M. Wendl, D. Koboldt, and E. Mardis. 2010. Analysis of Next Generation Genomic Data in Cancer: Accomplishments and Challenges. *Human Molecular Genetics* 1:R9.

Ding, W., M. Fujimura, A. Mori, I. Tooyama, et al. 1991. Light and electron microscopy of neuropeptide Y-containing nerves in human liver, gallbladder, and pancreas. *Gastroenterology* 101 (4):1054-1059.

Diskin, S., C. Hou, J. Glessner, E. Attiyeh, et al. 2009. Copy number variation at 1q21. 1 associated with neuroblastoma. *Nature* 459 (7249):987-991.

Donohue Jr, T. 2002. The ubiquitin proteasome system and its role in ethanol induced disorders. *Addiction Biology* 7 (1):15-28.

Donohue Jr, T., A. Cederbaum, S. French, S. Barve, et al. 2007. Role of the Proteasome in Ethanol Induced Liver Pathology. *Alcoholism: Clinical and Experimental Research* 31 (9):1446-1459.

Dransfield, M. T., and W. C. Bailey. 2006. COPD: racial disparities in susceptibility, treatment, and outcomes. *Clinics in Chest Medicine* 27 (3):463-471.

Edmondson, H. A., B. Henderson, and B. Benton. 1976. Liver-cell adenomas associated with use of oral contraceptives. *New England Journal of Medicine* 294 (9):470-472.

Eifel, P. J., A. Jhingran, D. C. Bodurka, C. Levenback, et al. 2002. Correlation of smoking history and other patient characteristics with major complications of pelvic radiation therapy for cervical cancer. *Journal of Clinical Oncology* 20 (17):3651-3657.

El-Salhy, M. 1999. Neuropeptide levels in murine liver and biliary pathways. *Upsala Journal of Medical Sciences* 105 (3):207-213.

Estivill, X., and L. Armengol. 2007. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics* 3 (10):e190.

Fanciulli, M., P. J. Norsworthy, E. Petretto, R. Dong, et al. 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genetics* 39 (6):721-723.

Farazi, P. A., and R. A. DePinho. 2006. Hepatocellular carcinoma pathogenesis: from genes to environment. *Nature Reviews Cancer* 6 (9):674-687.

Fedi, P., A. Bafico, A. Soria, W. Burgess, et al. 1999. Isolation and biochemical characterization of the human Dkk-1 homologue, a novel inhibitor of mammalian Wnt signaling. *Journal of Biological Chemistry* 274 (27):19465.

Folkersen, L., F. van't Hooft, E. Chernogubova, H. E. Agardh, G et al. 2010. Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. *Circulation: Cardiovascular Genetics* 3 (4):365-373.

Fredman, D., M. Siegfried, Y. Yuan, P. Bork, et al. 2002. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Research* 30 (1):387.

Freeman, J. L., G. H. Perry, L. Feuk, R. Redon, et al. 2006. Copy number variation: new insights in genome diversity. *Genome Research* 16 (8):949-961.

Friendly, M., and D. J. Denis. 2008. Milestones in the history of thematic cartography, statistical graphics, and data visualization. *Seeing Science: Today American Association for the Advancement of Science*.

Fu, J., E. A. M. Festen, and C. Wijmenga. 2011. Multi-ethnic studies in complex traits. *Human Molecular Genetics* 20 (R2):R206-R213.

Gamazon, E. R., W. Zhang, A. Konkashbaev, S. Duan, et al. 2010. SCAN: SNP and copy number annotation. *Bioinformatics* 26 (2):259-262.

Gary, T. L., K. M. V. Narayan, E. W. Gregg, G. L. A. Beckles, et al. 2003. Racial/ethnic differences in the healthcare experience (coverage, utilization, and satisfaction) of US adults with diabetes. *Ethnicity and Disease* 13 (1):47-54.

Ge, D., J. Fellay, A. J. Thompson, J. S. Simon, et al. 2009. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* 461 (7262):399-401.

Gerber, M. M., H. Hampel, N. P. Schulz, S. Fernandez, et al. 2012. Evaluation of allele-specific somatic changes of genome-wide association study susceptibility alleles in human colorectal cancers. *PLoS One* 7 (5):e37672.

Gibbs, R. A., J. W. Belmont, P. Hardenbol, T. D. Willis, et al. 2003. The international HapMap project. *Nature* 426 (6968):789-796.

Gitzelmann, R., M. Spycher, G. Feil, J. Muller, et al. 1996. Liver glycogen synthase deficiency: a rarely diagnosed entity. *European Journal of Pediatrics* 155 (7):561-567.

Glessner, J. T., and H. Hakonarson. 2009. Common variants in polygenic schizophrenia. *Genome Biology* 10 (9):236.

Glessner, J. T., K. Wang, G. Cai, O. Korvatska, et al. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459 (7246):569-573.

Gonzalez, E., H. Kulkarni, H. Bolivar, A. Mangano, et al. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307 (5714):1434-1440.

Greely, H. T. 2007. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annual Reviews Genomics and Human Genetics.* 8:343-364.

Haines, J. L., M. A. Hauser, S. Schmidt, W. K. Scott, et al. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308 (5720):419-421.

Hamosh, A., A. F. Scott, J. Amberger, D. Valle, et al. 2000. Online Mendelian inheritance in man (OMIM). *Human Mutation* 15 (1):57-61.

Hanley, A. J., K. Williams, A. Festa, L. E. Wagenknecht, et al. 2005. Liver Markers and Development of the Metabolic Syndrome The Insulin Resistance Atherosclerosis Study. *Diabetes* 54 (11):3140-3147.

Hardy, J., and A. Singleton. 2009. Genomewide association studies and human disease. *New England Journal of Medicine* 360 (17):1759-1768.

Harris, M., J. Clark, A. Ireland, J. Lomax, et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32 (Database issue):D258.

Hastings, P., J. Lupski, S. Rosenberg, and G. Ira. 2009. Mechanisms of change in gene copy number. *Nature Reviews Genetics* 10 (8):551-564.

Henrichsen, C. N., N. Vinckenbosch, S. Zöllner, E. Chaignat, et al. 2009. Segmental copy number variation shapes tissue transcriptomes. *Nature Genetics* 41 (4):424-429.

Higgins, M. E., M. Claremont, J. E. Major, C. Sander, et al. Lash. 2007. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Research* 35 (suppl 1):D721-D726.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* 106 (23):9362-9367.

Hou, X., R. Maser, B. Magenheimer, and J. Calvet. 1996. A mouse kidney-and liver-expressed cDNA having homology with a prokaryotic parathion hydrolase (phosphotriesterase)-encoding gene: abnormal expression in injured and polycystic kidneys. *Gene* 168 (2):157-163.

Iadonato, S. P., and M. G. Katze. 2009. Genomics: Hepatitis C virus gets personal. *Nature* 461 (7262):357-358.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, et al. 2004. Detection of large-scale variation in the human genome. *Nature Genetics* 36 (9):949-951.

Inohara, N., T. Koseki, S. Chen, X. Wu, et al. 1998. CIDE, a novel family of cell death activators with homology to the 45 kDa subunit of the DNA fragmentation factor. *The EMBO Journal* 17 (9):2526-2533.

Ioannidis, J. P., G. Thomas, and M. J. Daly. 2009. Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics* 10 (5):318-329.

Jahnchen, E., K. Blanck, K. Breuing, H. Gilfrich, et al. 1981. Plasma protein binding of azapropazone in patients with kidney and liver disease. *British Journal of Clinical Pharmacology* 11 (4):361.

Jiang, D. 2013. An Exploration of BMSF Algorithm in Genome-wide Association Mapping. 1-38.

Johnson, A. D., and C. J. O'Donnell. 2009. An open access database of genome-wide association results. *BMC Medical Genetics* 10 (1):6.

Kampstra, P. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. 1-9.

Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Research* 30 (1):42.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, Beanplot: A boxplot alternative for visual comparison of distributions. 2003. The UCSC genome browser database. *Nucleic Acids Research* 31 (1):51-54.

Kasper, C., and B. Voelkl. 2009. A social network analysis of primate groups. *Primates* 50 (4):343-356.

Kathiresan, S., C. J. Willer, G. M. Peloso, S. Demissie, et al. 2008. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics* 41 (1):56-65.

Kidd, J. M., G. M. Cooper, W. F. Donahue, H. S. Hayden, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* 453 (7191):56-64.

Kim, H.-Y., M.-J. Byun, and H. Kim. 2011. A replication study of genome-wide CNV association for hepatic biomarkers identifies nine genes associated with liver function. *Biochemistry and Molecular Biology Reports* 44 (9):578-583.

Kim, H.-Y., J.-H. Park, H. Kim, and B.-C. Kang. 2013. Semantic networks for genome-wide CNV associated with AST and ALT in Korean cohorts. *Molecular and Cellular Toxicology* 9 (2):103-111.

Kim, H., W. G. Yoo, J. Park, H. Kim, et al. 2014. Semantic Modeling for SNPs Associated with Ethnic Disparities in HapMap Samples. *Genomics and Informatics* 12 (1):35-41.

Kim, H. Y., J. Yu, and H. Kim. 2010. Analysis of copy number variation in 8,842 Korean individuals reveals 39 genes associated with hepatic biomarkers AST and ALT. *Biochemistry and Molecular Biology Reports* 43 (8):547-553.

Kladney, R. D., G. A. Bulla, L. Guo, A. L. Mason, et al. 2000. GP73, a novel Golgi-localized protein upregulated by viral infection. *Gene* 249 (1):53-65.

Klauck, S. M. 2006. Genetics of autism spectrum disorder. *European Journal of Human Genetics* 14 (6):714-720.

Klein, R. J., C. Zeiss, E. Y. Chew, J.-Y. Tsai, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308 (5720):385-389.

Kojima, T., N. Shworak, and R. Rosenberg. 1992. Molecular cloning and expression of two distinct cDNA-encoding heparan sulfate proteoglycan core proteins from a rat endothelial cell line. *Journal of Biological Chemistry* 267 (7):4870.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318 (5849):420-426.

Ku, C. S., E. Y. Loy, Y. Pawitan, and K. S. Chia. 2010. The pursuit of genome-wide association studies: where are we now&quest. *Journal of Human Genetics* 55 (4):195-206.

Kullai Reddy Ulavapalli, J., U. R. M. V. R. Pyreddy, and V. Bobbarala. 2011. Indian Journal of Novel Drug Delivery. *Indian Journal of Novel Drug Delivery* 3 (2):134-142.

Kumar, R. A., S. KaraMohamed, J. Sudi, D. F. Conrad, et al. 2008. Recurrent 16p11. 2 microdeletions in autism. *Human Molecular Genetics* 17 (4):628-638.

Lambert, C. 2005. HelixTree® Genetics Analysis Software. Golden Helix. *Inc. http//wwwgoldenhelixcom*.

Lee, B.-Y., D. H. Shin, S. Cho, K.-S. Seo, et al. 2012. Genome-wide analysis of copy number variations reveals that aging processes influence body fat distribution in Korea Associated Resource (KARE) cohorts. *Human Genetics* 131 (11):1795-1804.

Lee, J. A., C. Carvalho, and J. R. Lupski. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131 (7):1235-1247.

Lee, W. M. 2003. Drug-induced hepatotoxicity. *New England Journal of Medicine* 349 (5):474-485.

Leese, T., O. Farges, and H. Bismuth. 1988. Liver cell adenomas. A 12-year surgical experience from a specialist hepato-biliary unit. *Annals of Surgery* 208 (5):558.

Li, J., J. Ye, B. Xue, J. Qi, J. Zhang, et al. 2007. Cideb regulates diet-induced obesity, liver steatosis, and insulin sensitivity by controlling lipogenesis and fatty acid oxidation. *Diabetes* 56 (10):2523.

Li, Z. 2007. To study the physiological role of cideb protein by using cideb knockout mice as a model system. 1-158.

Liekens, A. M., J. De Knijf, W. Daelemans, B. Goethals, et al. 2011. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology* 12 (6):R57.

Liu, K., W. R. Hogan, and R. S. Crowley. 2011. Natural Language Processing methods and systems for biomedical ontology learning. *Journal of Biomedical Informatics* 44 (1):163-179.

Lockstone, H. E. 2011. Exon array data analysis using Affymetrix power tools and R statistical software. *Briefings in Bioinformatics* 12 (6):634-644.

Losko, S., and K. Heumann. 2009. Semantic data integration and knowledge management to represent biological network associations. *Protein Networks and Pathway Analysis* 563:241-258.

Losko, S., K. Wenger, W. Kalus, A. Ramge, et al. 2006. Knowledge networks of biological and medical data: an exhaustive and flexible solution to model life science domains. *Data Integration in the Life Sciences*. 232-239.

Lubis, C. P. 2008. PARASITIC ROUNDWORM. 1-27.

Lyon, M., J. Deakin, and J. Gallagher. 1994. Liver heparan sulfate structure. A novel molecular design. *Journal of Biological Chemistry* 269 (15):11208.

Madhava, V., C. Burgess, and E. Drucker. 2002. Epidemiology of chronic hepatitis C virus infection in sub-Saharan Africa. *The Lancet Infectious Diseases* 2 (5):293-302.

Maier, D., W. Kalus, M. Wolff, S. G. Kalko, et al. 2011. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC Systems Biology* 5 (1):38.

Marceau, P., S. Biron, F. Hould, S. Marceau, et al. 1999. Liver pathology and the metabolic syndrome X in severe obesity. *Journal of Clinical Endocrinology and Metabolism* 84 (5):1513.

Martin, D. N., B. J. Boersma, M. Yi, M. Reimers, et al. 2009. Differences in the tumor microenvironment between African-American and European-American breast cancer patients. *PLoS One* 4 (2):e4531.

Mashaghi, A., A. Ramezanpour, and V. Karimipour. 2004. Investigation of a protein complex network. *The European Physical Journal B-Condensed Matter and Complex Systems* 41 (1):113-121.

May, D. G. 1994. Genetic differences in drug disposition. *The Journal of Clinical Pharmacology* 34 (9):881-897.

McCarroll, S. A., and D. M. Altshuler. 2007. Copy-number variation and association studies of human disease. *Nature Genetics* 39:S37-S42.

McCray, A. T., and S. J. Nelson. 1995. The representation of meaning in the UMLS. *Methods of Information in Medicine* 34 (1-2):193-201.

McMurray, A., P. Pearson, R. Pace, and D. Scott. 2004. *Research: A Commonsense Approach*: Social Science Press.

Medina, I., D. Montaner, N. Bonifaci, M. A. Pujana, et al. 2009. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Research* 37 (suppl 2):W340-W344.

Miller, M. J., I. M. Krupp, M. D. Little, and C. Santos. 1974. Mebendazole. *JAMA: The Journal of the American Medical Association* 230 (10):1412-1414.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470 (7332):59-65.

Muehlschlegel, J. D., K.-Y. Liu, T. E. Perry, A. A. Fox, et al. 2010. Chromosome 9p21 variant predicts mortality after coronary artery bypass graft surgery. *Circulation* 122 (11 suppl 1):S60-S65.

Mukherjea, S., B. Bamba, and P. Kankar. 2005. Information retrieval and knowledge discovery utilizing a biomedical patent semantic Web. *Knowledge and Data Engineering, IEEE Transactions on* 17 (8):1099-1110.

Nei, M., and A. K. Roychoudhury. 1993. Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution* 10 (5):927-943.

Nguyen, D.-Q., C. Webber, and C. P. Ponting. 2006. Bias of selection on human copy-number variants. *PLoS Genetics* 2 (2):e20.

Nguyen, D.-Q., C. P. Webber, J. Hehir-Kwa, R. Pfundt, et al. 2008. Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Research*:gr. 077289.077108.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, et al. 2008. Genes mirror geography within Europe. *Nature* 456 (7218):98-101.

Nyhan, W. L. 2005. Disorders of purine and pyrimidine metabolism. *Molecular Genetics and Metabolism* 86 (1):25-33.

Okabe, H., S. Satoh, Y. Furukawa, T. Kato, et al. 2003. Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. *Cancer Research* 63 (12):3043.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, et al. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* 33 (17):5691-5702.

Paudel, Y., O. Madsen, H.-J. Megens, L. A. Frantz, et al. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* 14 (1):449.

Pearson, T. A., and T. A. Manolio. 2008. How to interpret a genome-wide association study. *JAMA: the Journal of the American Medical Association* 299 (11):1335-1344.

Perry, G., A. Ben-Dor, A. Tsalenko, N. Sampas, et al. 2008. The fine-scale and complex architecture of human copy-number variation. *The American Journal of Human Genetics* 82 (3):685-695.

Perry, G. H., N. J. Dominy, K. G. Claw, A. S. Lee, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics* 39 (10):1256-1260.

Peters, U., C. M. Hutter, L. Hsu, F. R. Schumacher, et al. 2012. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Human Genetics* 131 (2):217-234.

Picornell, Y., L. Mei, K. Taylor, H. Yang, et al. 2007. TNFSF15 is an ethnic-specific IBD gene. *Inflammatory Bowel Diseases* 13 (11):1333-1338.

Pieters, M., and H. H. Vorster. 2008. Nutrition and hemostasis: a focus on urbanization in South Africa. *Molecular Nutrition and Food Research* 52 (1):164-172.

Pocai, A., T. Lam, S. Obici, R. Gutierrez-Juarez, et al. 2006. Restoration of hypothalamic lipid sensing normalizes energy and glucose homeostasis in overfed rats. *Journal of Clinical Investigation* 116 (4):1081-1091.

Polymeropoulos, M. H., J. J. Higgins, L. I. Golbe, W. G. Johnson, et al. 1996. Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. *Science* 274 (5290):1197-1199.

Popova, T., E. Manié, D. Stoppa-Lyonnet, G. Rigaill, et al. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biology* 10 (11):R128-R128.

Prill, R. J., P. A. Iglesias, and A. Levchenko. 2005. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biology* 3 (11):e343.

Proulx, S. R., D. E. Promislow, and P. C. Phillips. 2005. Network thinking in ecology and evolution. *Trends in Ecology and Evolution* 20 (6):345-353.

Prugnolle, F., A. Manica, and F. Balloux. 2005. Geography predicts neutral genetic diversity of human populations. *Current Biology* 15 (5):R159-R160.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81 (3):559-575.

Redon, R., S. Ishikawa, K. Fitch, L. Feuk, et al. 2006. Global variation in copy number in the human genome. *Nature* 444 (7118):444.

Repping, S., S. K. van Daalen, L. G. Brown, C. M. Korver, et al. 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nature Genetics* 38 (4):463-467.

Richert, L., M. Liguori, C. Abadie, B. Heyd, et al. 2006. Gene expression in human hepatocytes in suspension after isolation is similar to the liver of origin, is not affected by hepatocyte cold storage and cryopreservation, but is strongly changed after hepatocyte plating. *Drug Metabolism and Disposition* 34 (5):870.

Rioux, V., R. Landry, and A. Bensadoun. 2002. Sandwich immunoassay for the measurement of murine syndecan-4. *Journal of Lipid Research* 43 (1):167.

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science-AAAS-Weekly Paper Edition* 273 (5281):1516-1517.

Romanuk, T. N., R. J. Vogt, A. Young, C. Tuck, et al. 2010. Maintenance of positive diversity-stability relations along a gradient of environmental stress. *PLoS One* 5 (4):e10378.

Ruhl, C. E., and J. E. Everhart. 2012. Upper limits of normal for alanine aminotransferase activity in the United States population. *Hepatology* 55 (2):447-454.

Saad, A., A. Aziz, I. Yehia, A. El-Ghareeb, et al. 2009. Programmed Cell Death in the Liver of Different Species of Anuran Amphibians During Metamorphosis. *Australian Journal of Basic and Applied Sciences* 3 (4):4644-4655.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449 (7164):913-918.

Sahin-Tóth, M. 2006. Biochemical models of hereditary pancreatitis. *Endocrinology and Metabolism Clinics of North America* 35 (2):303.

Sainsbury, A., C. Schwarzer, M. Couzens, A. Jenkins, et al. 2002. Y4 receptor knockout rescues fertility in ob/ob mice. *Genes and Development* 16 (9):1077-1088.

Santra, S., and U. Baumann. 2008. Experience of nitisinone for the pharmacological treatment of hereditary tyrosinaemia type. *Expert Opinion on Pharmacotherapy* 9 (7):1229-1236.

Sattar, N., O. Scherbakova, I. Ford, D. S. J. O'Reilly, et al. 2004. Elevated alanine aminotransferase predicts new-onset type 2 diabetes independently of classical risk factors, metabolic syndrome, and C-reactive protein in the west of Scotland coronary prevention study. *Diabetes* 53 (11):2855-2860.

Scharpf, R. B., R. Irizarry, W. Ritchie, B. Carvalho, et al. 2010. Using the R package crlmm for genotyping and copy number estimation. *Journal of Statistical Software* 40 (12):1-32.

Sebastiani, P., N. Solovieff, A. Puca, S. W. Hartley, et al. 2010. Genetic signatures of exceptional longevity in humans. *Science* 10:1126.

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* 316 (5823):445-449.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305 (5683):525-528.

Shannon, P., A. Markiel, O. Ozier, N. Baliga, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13 (11):2498.

Sheth, S., S. Flamm, F. Gordon, and S. Chopra. 1998. AST/ALT ratio predicts cirrhosis in patients with chronic hepatitis C virus infection. *The American Journal of Gastroenterology* 93 (1):44-48.

Shin, G.-H., Y.-K. Kang, S.-H. Lee, S. J. Kim, et al. 2012. mRNA-centric semantic modeling for finding molecular signature of trace chemical in human blood. *Molecular and Cellular Toxicology* 8 (1):35-41.

Shlien, A., and D. Malkin. 2009. Copy number variations and cancer. *Genome Medicine* 1 (6):62.

Shworak, N., J. Liu, L. Petros, L. Zhang, et al. 1999. Multiple isoforms of heparan sulfate D-glucosaminyl 3-O-sulfotransferase. *Journal of Biological Chemistry* 274 (8):5170.

Smigielski, E., K. Sirotkin, M. Ward, and S. Sherry. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research* 28 (1):352.

Solomon, L., G. Robin, and H. Valkenburg. 1975. Rheumatoid arthritis in an urban South African Negro population. *Annals of the Rheumatic Diseases* 34 (2):128-135.

Sommadossi, J. P. 1995. Comparison Of Metabolism And In Vitro Antiviral Activity Of Stavudine Versus Other 2 ‘, 3’-Dideoxynucleoside Analogues. *Journal of Infectious Diseases* 171 (Supplement 2):S88.

Sowa, J. F. 1991. Principles of semantic networks.

Stankiewicz, P., and J. R. Lupski. 2010. Structural variation in the human genome and its role in disease. *Annual Review of Medicine* 61:437-455.

Starr, H. L., and J. Kemner. 2005. Multicenter, randomized, open-label study of OROS methylphenidate versus atomoxetine: treatment outcomes in African-American children with ADHD. *Journal of the National Medical Association* 97 (10 Suppl):11S.

Steane, A. M. 1996. Error correcting codes in quantum theory. *Physical Review Letters* 77 (5):793.

Stefansson, H., D. Rujescu, S. Cichon, O. P. Pietiläinen, et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* 455 (7210):232-236.

Steinkühler, C., G. Biasiol, M. Brunetti, A. Urbani, et al. 1998. Product inhibition of the hepatitis C virus NS3 protease. *Biochemistry* 37 (25):8899-8905.

Stephan, K. E., C. C. Hilgetag, G. A. Burns, M. A. O'Neill, et al. 2000. Computational analysis of functional connectivity between areas of primate cerebral cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355 (1393):111-126.

Stone, J. L., M. C. O’Donovan, H. Gurling, G. K. Kirov, et al. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455 (7210):237-241.

Stranger, B., M. Forrest, M. Dunning, C. Ingle, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315 (5813):848.

Strawbridge, R. J., J. Dupuis, I. Prokopenko, A. Barker, et al. 2011. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* 60 (10):2624-2634.

Taniguchi, K., L. R. Roberts, I. N. Aderca, X. Dong, et al. 2002. Mutational spectrum of beta-catenin, AXIN1, and AXIN2 in hepatocellular carcinomas and hepatoblastomas. *Oncogene* 21 (31):4863.

Thapa, B., and A. Walia. 2007. Liver function tests and their interpretation. *Indian Journal of Pediatrics* 74 (7):663-671.

Theuns, J., N. Brouwers, S. Engelborghs, K. Sleegers, et al. 2006. Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease. *The American Journal of Human Genetics* 78 (6):936-946.

Tishkoff, S. A., and S. M. Williams. 2002. Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics* 3 (8):611-621.

Truter, I. 2005. Methylphenidate: Prescribing patterns in a South African primary care patient population. *Journal of Clinical Pharmacy and Therapeutics* 30 (1):59-63.

Van Gennip, A., R. De Abreu, H. Van Lenthe, J. Bakkeren, et al. 1997. Dihydropyrimidinase deficiency: confirmation of the enzyme defect in dihydropyrimidinuria. *Journal of Inherited Metabolic Disease* 20 (3):339-342.

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. 2009. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* 10 (4):252-263.

Walters, R., S. Jacquemont, A. Valsesia, A. De Smith, et al. 2010. A new highly penetrant form of obesity due to deletions on chromosome 16p11. 2. *Nature* 463 (7281):671-675.

Wang, D. G., J.-B. Fan, C.-J. Siao, A. Berno, et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280 (5366):1077-1082.

Wang, K., M. Li, and M. Bucan. 2007a. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics* 81 (6):1278-1283.

Wang, K., M. Li, D. Hadley, R. Liu, et al. 2007b. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research* 17 (11):1665-1674.

Warrell, D., and S. Anderson. 2014. *Expedition Medicine: Revised Edition*: Routledge.

Weiss, L. A., Y. Shen, J. M. Korn, D. E. Arking, et al. 2008. Association between microdeletion and microduplication at 16p11. 2 and autism. *New England Journal of Medicine* 358 (7):667-675.

Wishart, D. S., C. Knox, A. C. Guo, D. Cheng, et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* 36 (suppl 1):D901-D906.

Wishart, D. S., C. Knox, A. C. Guo, S. Shrivastava, et al. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 34 (suppl 1):D668-D672.

Wu, C., C. Orozco, J. Boyer, M. Leglise, et al. 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology* 10 (11):R130.

Xu, B., A. Woodroffe, L. Rodriguez-Murillo, J. L. Roos, et al. 2009. Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proceedings of the National Academy of Sciences* 106 (39):16746-16751.

Yang, Y., E. K. Chung, Y. L. Wu, S. L. Savelli, et al. 2007. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *The American Journal of Human Genetics* 80 (6):1037-1054.

Ye, J., J. Li, Y. Liu, X. Li, et al. 2009. Cideb, an ER-and Lipid Droplet-Associated Protein, aMediates VLDL Lipidation and Maturation byaInteracting with Apolipoprotein B. *Cell Metabolism* 9 (2):177-190.

Yim, S.-H., T.-M. Kim, H.-J. Hu, J.-H. Kim, et al. 2010. Copy number variations in East-Asian population and their evolutionary and functional implications. *Human Molecular Genetics* 19 (6):1001-1008.

Yim, S., T. Kim, H. Hu, J. Kim, et al. 2009. Copy number variations in East-Asian population and their evolutionary and functional implications. *Human Molecular Genetics*.

Yuan, X., D. Waterworth, J. Perry, N. Lim, et al. 2008. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *The American Journal of Human Genetics* 83 (4):520-528.

Zhang, F., W. Gu, M. E. Hurles, and J. R. Lupski. 2009. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* 10:451-481.

Zhang, X., and A. Beynen. 2007. Influence of dietary fish proteins on plasma and liver cholesterol concentrations in rats. *British Journal of Nutrition* 69 (03):767-777.

Zhang, Y., S. De, J. Garner, K. Smith, et al. 2010. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Medical Genomics* 3 (1):1.

Zucman-Rossi, J. 2011. Benign Liver Tumors. *Molecular Pathology of Liver Diseases*:769-775.

Supplementary Materials

Supplementary Table 2.1. Summaries the significant CNVs associated with AST and ALT from KARE1 (A) and KARE2 (B).

(A)

Trait	CNVR	Estimate	t.value	P.value	P.bon*
AST	chr1.190686383.190745497	-0.00562694	-4.63479301	3.62E-06	0.0378
	chr1.211104284.211161180	-0.0041324	-5.12260135	3.08E-07	0.0032
	chr1.212543184.212566377	0.00296302	4.824720614	1.43E-06	0.0149
	chr1.215574462.215683383	-0.00734297	-5.7181841	1.11E-08	0.0001
	chr1.216310320.216321229	-0.00269225	-5.10954504	3.30E-07	0.0034
	chr1.230384738.230395024	-0.00260201	-5.1752873	2.33E-07	0.0024
	chr1.234226764.234251691	-0.00269874	-5.00763274	5.62E-07	0.0058
	chr1.33989245.34023033	0.00350787	4.782238893	1.76E-06	0.0184
	chr2.164793303.164842648	-0.00265892	-4.77319988	1.84E-06	0.0192
	chr2.178202016.178230257	-0.00287572	-4.90963685	9.29E-07	0.0097
	chr2.224474138.224753508	-0.01226929	-4.90421362	9.55E-07	0.0100
	chr2.36375263.36467684	-0.0063179	-4.63885912	3.55E-06	0.0371
	chr2.42724479.42843096	0.00610118	4.625292922	3.79E-06	0.0396
	chr2.46733103.46761131	0.00342362	5.447868676	5.23E-08	0.0005
	chr2.5148726.5223415	-0.00807348	-4.85141196	1.25E-06	0.0130
	chr2.64521312.64568911	-0.0043106	-5.2815113	1.31E-07	0.0013
	chr2.76627489.76884533	-0.00971247	-5.39401339	7.07E-08	0.0007
	chr2.8514325.8536702	-0.0045567	-5.61232065	2.06E-08	0.0002
	chr3.11981746.11996902	-0.00252449	-4.65851932	3.23E-06	0.0337
	chr3.122343006.122543714	-0.00673691	-4.8127667	1.51E-06	0.0158
	chr3.123100746.123177102	-0.0048797	-5.83544121	5.55E-09	5.85E-05
	chr3.138809756.139122947	-0.01129058	-4.58534504	4.59E-06	0.0479
	chr3.142851641.142908108	-0.00415276	-5.21904397	1.84E-07	0.0019
	chr3.149817720.149889489	-0.00529755	-5.00636199	5.65E-07	0.0059
	chr3.45192647.45198645	-0.00314139	-5.63243663	1.83E-08	0.0001
	chr3.55324743.55358528	0.00447925	4.729454681	2.29E-06	0.0238
	chr3.55358671.55371612	-0.00350586	-4.85560177	1.22E-06	0.0127

chr3.56098137.56116883	-0.00292615	-5.15409893	2.60E-07	0.0027
chr3.60832659.61018553	-0.01476049	-5.53016395	3.29E-08	0.0003
chr3.8233174.8360164	-0.00899059	-4.9889908	6.19E-07	0.0064
chr4.154956613.155019392	-0.00801492	-5.25264928	1.53E-07	0.0016
chr4.164012707.164647495	-0.01257508	-5.39398421	7.07E-08	0.0007
chr4.189999425.190017092	-0.0035088	-4.77164008	1.86E-06	0.0194
chr4.42421165.42510046	-0.00615515	-4.71216826	2.49E-06	0.0260
chr4.72731411.72791238	-0.00335105	-5.23447336	1.69E-07	0.0017
chr5.160090803.160146434	-0.0069839	-5.74841313	9.31E-09	9.80E-05
chr5.174836093.174881422	0.002810785	4.617785693	3.93E-06	0.0410
chr5.8703331.8715504	-0.00305482	-5.55122252	2.92E-08	0.0003
chr5.9885874.9895944	0.002650309	5.056296236	4.36E-07	0.0045
chr6.104832382.105058386	-0.01084741	-5.00287341	5.76E-07	0.0060
chr6.115556221.115764663	-0.00679112	-4.76273008	1.94E-06	0.0202
chr6.143855908.143871871	-0.00372814	-5.79708495	6.98E-09	7.35E-05
chr6.156187976.156200183	-0.00332281	-4.82066702	1.45E-06	0.0152
chr6.158049703.158072444	-0.0035938	-5.44010751	5.47E-08	0.0005
chr6.57728086.57936849	-0.00868077	-5.44164878	5.42E-08	0.0005
chr6.66685345.67010174	-0.00796807	-5.40848412	6.52E-08	0.0006
chr6.91257342.91444831	-0.01077119	-5.83537264	5.56E-09	5.85E-05
chr7.103825729.103849313	-0.00431954	-6.08221051	1.23E-09	1.30E-05
chr7.153892749.154018058	0.003747716	4.806661448	1.56E-06	0.0163
chr7.154499812.154511611	-0.0034231	-6.20195279	5.83E-10	6.14E-06
chr7.31487325.31536072	-0.00745008	-5.19927968	2.05E-07	0.0021
chr8.105247995.105727124	-0.01264243	-4.60103467	4.26E-06	0.0444
chr8.128084274.128095592	-0.00309333	-4.92730142	8.49E-07	0.0088
chr8.138607006.138650961	-0.00665006	-4.63199976	3.67E-06	0.0383
chr8.59848863.59867441	-0.00391141	-5.41453008	6.31E-08	0.0006
chr9.10338710.10363923	0.004098743	4.619299717	3.90E-06	0.0407
chr9.119285724.119462020	-0.00955525	-5.21613482	1.87E-07	0.0019
chr9.119632012.119677310	-0.0054962	-4.60643252	4.15E-06	0.0433
chr9.13869083.13915424	-0.0047304	-5.33609823	9.73E-08	0.0010
chr9.23760626.23940769	-0.00884989	-4.62981974	3.71E-06	0.0387
chr10.104845268.104905300	-0.00472968	-5.26769368	1.41E-07	0.0014
chr10.10578227.10631958	0.004086376	4.895347001	9.99E-07	0.0104
chr10.112430512.112501145	-0.00632071	-5.31112879	1.12E-07	0.0011
chr10.113130665.113189385	-0.00654365	-4.74914273	2.07E-06	0.0216

	chr10.123529304.123546416	-0.00316959	-5.05765405	4.33E-07	0.0045
	chr10.132497389.132515518	-0.00306318	-6.12715736	9.33E-10	9.82E-06
	chr10.132566079.132589354	-0.00258077	-4.58581446	4.58E-06	0.0478
	chr10.16371836.16615099	0.01434495	4.824009491	1.43E-06	0.0149
	chr10.26659186.26863275	0.011782458	4.815389409	1.49E-06	0.0156
	chr10.53731444.53843163	-0.00705414	-4.84109261	1.31E-06	0.0137
	chr10.728630.762082	-0.00339361	-5.77961849	7.74E-09	8.15E-05
	chr10.9448261.9709991	-0.00840936	-5.06940222	4.07E-07	0.0042
	chr11.128850539.128920427	-0.00599481	-5.41055885	6.45E-08	0.0006
	chr11.22123259.22182561	-0.00477134	-4.86588082	1.16E-06	0.0121
	chr11.87666857.87892347	-0.01174167	-4.81768894	1.48E-06	0.0154
	chr11.94541819.94585583	-0.00324551	-4.7652776	1.92E-06	0.0200
	chr12.29089080.29470913	-0.01000863	-4.71076811	2.51E-06	0.0261
	chr12.60094390.60212637	-0.00827259	-4.63854106	3.56E-06	0.0371
	chr12.64360488.64443794	-0.0076391	-5.10966343	3.29E-07	0.0034
	chr13.45060186.45080021	-0.00350247	-4.67051672	3.05E-06	0.0318
	chr14.56112557.56137765	-0.00359397	-6.22656391	4.99E-10	5.25E-06
	chr15.33754494.34067560	-0.00925305	-4.94543895	7.74E-07	0.0081
	chr15.41680998.41721532	0.003049095	5.084714883	3.76E-07	0.0039
	chr15.43304465.43329604	-0.00376538	-5.38550437	7.41E-08	0.0007
	chr15.46811866.47497399	-0.01758909	-4.94576833	7.72E-07	0.0080
	chr15.54342294.54871765	-0.01517471	-5.30418739	1.16E-07	0.0012
	chr15.59271321.59333206	-0.00441849	-4.75672187	2.00E-06	0.0208
	chr15.62084153.62100208	-0.00298471	-4.86240187	1.18E-06	0.0123
	chr15.97260662.97313771	-0.00338892	-4.78385393	1.75E-06	0.0182
	chr16.11261998.11292512	-0.00260492	-4.88432404	1.06E-06	0.0110
	chr16.53867341.53882426	0.00338734	6.042676478	1.58E-09	1.66E-05
	chr16.55344608.55440939	0.007241801	4.597144391	4.34E-06	0.0453
	chr17.47587358.47612927	-0.00391375	-6.48345039	9.45E-11	9.95E-07
	chr17.9800824.9830867	-0.00252485	-4.6600614	3.21E-06	0.0335
	chr17.9835846.9859734	0.003782784	5.128942324	2.98E-07	0.0031
	chr18.41464192.41550284	0.006699921	5.023567584	5.17E-07	0.0054
	chr18.43135458.43184988	0.004991118	5.147293955	2.70E-07	0.0028
	chr20.58158885.58179582	-0.00323678	-5.4813767	4.34E-08	0.0004
	chr21.18819915.18895085	-0.00536218	-5.29444057	1.22E-07	0.0012
AST/ ALT	chr1.211104284.211161180	-0.0041324	-5.12260135	3.08E-07	0.0032

		-0.01643616	-4.70023634	2.64E-06	0.0277
	chr2.8514325.8536702	-0.0045567	-5.61232065	2.06E-08	0.0002
		-0.01662266	-4.72171978	2.37E-06	0.0249
	chr3.45192647.45198645	-0.00314139	-5.63243663	1.83E-08	0.0001
		-0.01145903	-4.73835734	2.19E-06	0.0230
	chr6.158049703.158072444	-0.0035938	-5.44010751	5.47E-08	0.0005
		-0.01312749	-4.5830787	4.64E-06	0.0488
	chr7.154499812.154511611	-0.0034231	-6.20195279	5.83E-10	6.14E-06
		-0.01101853	-4.60195973	4.24E-06	0.0446
	chr10.112430512.112501145	-0.00632071	-5.31112879	1.12E-07	0.0011
		-0.02460858	-4.76987178	1.87E-06	0.0197
	chr10.132497389.132515518	-0.00306318	-6.12715736	9.33E-10	9.82E-06
		-0.01016156	-4.68595299	2.83E-06	0.0297
	chr10.16371836.16615099	0.01434495	4.824009491	1.43E-06	0.0149
		0.059351689	4.604946411	4.18E-06	0.0440
	chr10.728630.762082	-0.00339361	-5.77961849	7.74E-09	8.15E-05
		-0.011744	-4.61199984	4.04E-06	0.0425
	chr17.47587358.47612927	-0.00391375	-6.48345039	9.45E-11	9.95E-07
		-0.01342386	-5.12669489	3.01E-07	0.0031
ALT	chr7.38268982.38353956	0.027888439	4.625996932	3.78E-06	0.0397
	chr9.13558063.13688019	-0.04373448	-4.85087036	1.25E-06	0.0131
	chr14.21761956.22004498	0.038342492	6.71252981	2.03E-11	2.14E-07
	chr14.21518495.21697688	0.042803875	5.8046152	6.67E-09	7.03E-05
	chr14.87637635.88374957	0.078309677	5.03327452	4.92E-07	0.0051
	chr17.14014164.14613835	-0.05885667	-4.57958737	4.72E-06	0.0496

[*]: *p*-value after bonferroni correction.

(B)

Trait	CNVR	Estimate	t.value	P.value
AST	chr1.72541492.72583724	-0.00252224	-2.24589801	0.025251678
	chr3.175301024.176121434	-0.04281881	-1.98598239	0.047712383
	chr3.194317766.194360584	0.036765682	2.326108446	0.020508044
	chr3.58683409.58684433	-0.0027403	-2.51312433	0.012356015
	chr4.122501906.122504585	0.004124158	1.998321312	0.046353262
	chr4.48788531.48849514	0.003911	2.485606105	0.01333764
	chr5.57361772.57369278	0.002246951	2.099705328	0.036376027
	chr5.75155872.75479220	0.040766266	2.017344963	0.044321999
	chr6.119014075.119139790	-0.03279574	-2.31403497	0.021168015
	chr7.142929078.143198980	0.010337496	2.242334569	0.025484269
	chr9.104852719.104862217	-0.0031732	-2.0378374	0.042220395
	chr9.1105998.1315179	0.035575854	2.046601394	0.041347115
	chr11.79553671.79645152	0.034996344	2.191484375	0.028991318
	chr12.107007904.107314707	0.038847651	2.182512919	0.029651252
	chr12.111893559.112564198	0.042190581	2.071007765	0.038997279
	chr12.52741396.53668999	0.058676793	2.146295691	0.03244886
	chr16.46036433.47063550	0.08739629	2.004573774	0.04568048
	chr16.47069610.48586106	0.066291865	2.037393956	0.042266629
	chr16.72954547.73014090	-0.00417354	-2.12568147	0.034140381
	chr18.52339709.52856818	-0.05336115	-2.01952796	0.044097107
	chr18.63110881.63118233	0.002246578	2.217343961	0.027159639
	chr20.43763418.45210381	0.110809943	2.647422811	0.008431346
	chr22.24019061.24248709	0.012967419	2.32978956	0.020314153
AST/ALT	chr1.173288426.174429270	0.05531068	2.176667432	0.030085286
		0.221662139	2.160263457	0.031341173
	chr2.121450996.123193472	0.086699014	2.177781201	0.030001613
		0.396768393	2.47249729	0.013829131

	chr3.5510570.5514102	0.002872552	2.264579058	0.024069039
		0.011317978	2.209156225	0.027725623
	chr3.84782471.84784814	0.003647539	2.539271162	0.01148337
		0.014967632	2.581347443	0.01019396
	chr4.172611447.172614496	-0.00385719	-2.92487633	0.00364064
		-0.01360894	-2.54943512	0.011159253
	chr7.22702971.22707984	0.013605183	2.427025625	0.015660687
		0.054183899	2.393449497	0.017147422
	chr11.4926841.4930593	-0.00187396	-2.11866766	0.0347329
		-0.00887023	-2.49015039	0.013172931
	chr13.106183224.106695599	0.055358502	2.053620507	0.040660435
		0.286113872	2.638843555	0.008643068
	chr14.23001245.24313152	0.073052483	2.007816106	0.045333154
		0.300218611	2.044934632	0.04151363
ALT	chr1.199155367.200961788	0.360539082	2.048881351	0.041120852
	chr1.219016303.223441941	0.385485141	2.117289005	0.03484734
	chr1.22191576.22210700	0.020790086	2.375220487	0.018005551
	chr1.230525067.230526526	0.008692195	2.106979393	0.035736815
	chr2.221471681.221823524	0.301410053	2.583529314	0.010130772
	chr2.97121236.97236352	-0.04456678	-2.00746587	0.045367233
	chr3.141976161.142025217	0.070762307	2.280118292	0.023122319
	chr3.163699311.163709641	-0.01467698	-2.59934102	0.009683221
	chr3.36258949.36260022	0.010868991	2.284123045	0.022883656
	chr4.131273681.133796865	0.312804497	2.306821112	0.021571145
	chr4.162104357.162151102	-0.03474751	-2.52280248	0.012026362
	chr5.12868768.12873064	0.0123726	2.283787234	0.022903586
	chr5.158988746.159912305	0.352296656	2.231954329	0.026167009
	chr6.119818455.119820127	-0.01604655	-2.04022396	0.041979427
	chr6.67859000.70188705	0.35631135	1.997015926	0.046495485

chr6.77073751.77084489	0.021686236	2.364831305	0.018511324
chr9.44667843.44795721	-0.02460546	-2.06829006	0.039251753
chr10.72716105.73401893	0.244773694	2.02248002	0.043788446
chr10.77927055.77930579	0.007544342	2.371988077	0.018162767
chr11.25870713.27225055	0.478678528	2.482046133	0.013471597
chr11.4931728.4933471	-0.0248949	-2.04532853	0.041474612
chr13.36970024.36982745	0.016282382	2.37388628	0.01807216
chr13.56656259.56676369	0.017671657	2.276670101	0.02333217
chr14.43571666.43600193	-0.01758893	-2.00596218	0.045531477
chr14.47477520.47968292	0.174713937	2.208825194	0.027751645
chr15.39996964.40459070	0.23001712	2.084990103	0.037702224
chr16.14897352.14967222	-0.02578776	-1.99400826	0.046827937
chr16.33288255.33680554	-0.02147122	-1.96829647	0.049722414
chr16.6838607.7021740	0.131778821	2.612846773	0.009316726
chr19.48394861.48448065	0.007010712	2.018375531	0.044217453
chr20.48228324.48562412	0.307554806	3.236654039	0.001310124
chr21.29429536.31660765	0.30194253	2.144282262	0.03261234

Supplementary Table 2.2. Summaries of the genes whose entire sequences were located within the CNV region associated with hepatic biochemical markers in KARE1 (A) and KARE2 (B).

(A)

Trait	CNV regions	No. Genes	Genes
AST	Chr1:199155367-200961788	26	RNPEP,PTPN7,IGFN1,ELF3,PPP1R12B,UBE2T,SHISA4,SYT2,CACNA1S,TNNT2,TIMM17A,KIF21B,LAD1,NAV1,IPO9,GPR37L1,ARL8A,LGR6,LMOD1,PKP1,PTPRV,CSRP1,PHLDA3,TNNI1,TMEM9,RPS10P7
	Chr1:219016303-223441941	24	HHIPL2,SUSD4,DUSP10,TP53BP2,CNIH4,TLR5,DNAH14,C1orf65,TAF1A,MIA3,HLX,AIDA,NVL,WDR26,CNIH3,FAM177B,DISP1,FBXO28,DEGS1,MOSC1,C1orf58,CAPN2,LOC400804,CAPN8
	Chr1:22191576-22210700	0	
	Chr1:230525067-230526526	0	
	Chr10:72716105-73401893	4	CDH23,SLC29A3,C10orf54,PSAP
	Chr10:77927055-77930579	0	
	Chr11:25870713-27225055	5	MUC15,SLC5A12,FIBIN,BBOX1,ANO3
	Chr11:4931728-4933471	0	
	Chr13:36970024-36982745	0	
	Chr13:56656259-56676369	0	
	Chr14:43571666-43600193	0	
	Chr14:47477520-47968292	0	
	Chr15:39996964-40459070	6	PLA2G4E,VPS39,TMEM87A,PLA2G4D,PLA2G4F,GANC
	Chr16:14897352-14967222	1	NPIP
	Chr16:33288255-33680554	0	
	Chr16:6838607-7021740	0	
	Chr19:48394861-48448065	0	
	Chr2:221471681-221823524	0	
	Chr2:97121236-97236352	0	
	Chr20:48228324-48562412	1	CEBPB
	Chr21:29429536-31660765	38	CLDN17,KRTAP26-1,KRTAP27-1,KRTAP23-1,KRTAP13-2,KRTAP13-4,KRTAP15-1,KRTAP19-2,KRTAP19-3,KRTAP19-4,KRTAP19-5,KRTAP19-6,KRTAP19-7,KRTAP6-3,KRTAP6-2,KRTAP22-1,KRTAP6-1,KRTAP20-1,KRTAP20-

			2,KRTAP21-2,KRTAP21-1,KRTAP8-1,KRTAP11-1,GRIK1,C21orf41,BACH1,KRTAP13-1,KRTAP13-3,KRTAP19-8,KRTAP20-3,KRTAP19-1,KRTAP24-1,CLDN8,KRTAP25-1,KRTAP20-4,KRTAP7-1,NCRNA00110,C21orf109
	Chr3:141976161-142025217	0	
	Chr3:163699311-163709641	0	
	Chr3:36258949-36260022	0	
	Chr4:131273681-133796865	0	
	Chr4:162104357-162151102	0	
	Chr5:12868768-12873064	0	
	Chr5:158988746-159912305	9	CCNJL,TTC1,SLU7,ADRA1B,FABP6,C5orf54,C1QTNF2,PTTG1,PWWP2A
	Chr6:119818455-119820127	0	
	Chr6:67859000-70188705	1	BAI3
	Chr6:77073751-77084489	0	
	Chr9:44667843-44795721	0	
AST/ ALT	Chr1:173288426-174429270	4	TNN,TNR,KIAA0040,SCARNA3
	Chr2:121450996-123193472	5	TSN,CLASP1,MKI67IP,TFCP2L1,RNU4ATAC
	Chr3:5510570-5514102	0	
	Chr3:84782471-84784814	0	
	Chr4:172611447-172614496	0	
	Chr4:48788531-48849514	0	
	Chr7:22702971-22707984	0	
	Chr11:4926841-4930593	0	
	Chr13:106183224-106695599	0	
	Chr14:23001245-24313152	48	REC8,GZMH,NRL,CHMP4A,PSME2,DHRS2,DHRS4L1,RIPK3,AP1G2,SDR39U1,PSME1,LTB4R2,RABGGTA,IRF9,FITM1,C14orf21,NFATC4,GZMB,THTPA,LTB4R,NGDN,PCK2,TINF2,TM9SF1,FAM158A,DHRS4,TGM1,DHRS1,ADCY4,IPO4,GMPR2,TSSK4,JPH4,CMA1,MDP-1,RNF31,DHRS4L2,WDR23,LRRC16B,CPNE6,CIDEB,KIAA1305,CBLN3,KIAA0323,NEDD8,CTSG,C14orf167,C14orf165
ALT	Chr1:72541492-72583724	0	
	Chr3:175301024-176121434	0	
	Chr3:194317766-194360584	0	
	Chr3:58683409-58684433	0	
	Chr4:122501906-122504585	0	
	Chr5:57361772-57369278	0	
	Chr5:75155872-75479220	0	

	Chr6:119014075-119139790	0	
	Chr7:142929078-143198980	4	FAM115C,LOC441294,CTAGE6,LOC154761
	Chr9:104852719-104862217	0	
	Chr9:1105998-1315179	0	
	Chr11:79553671-79645152	0	
	Chr12:107007904-107314707	2	WSCD2,CMKLR1
	Chr12:111893559-112564198	13	OAS2,SDS,SDSL,PLBD2,RASAL1,TPCN1,DTX1,DDX54,C12orf52,IQCD,SLC24A6,LHX5,LOC387885
	Chr12:52741396-53668999	21	LOC100240735,ZNF385A,LACRT,ITGA5,SMUG1,HNRNPA1,DCD,PPP1R1A,MUCL1,COPZ1,CBX5,GPR84,NCKAP1L,KIAA0748,LOC100240734,PDE1B,NFE2,GTSF1,GLYCAM1,HNRPA1L-2,LOC400043
	Chr16:46036433-47063550	5	SIAH1,ABCC11,LONP2,PHKB,ABCC12
	Chr16:47069610-48586106	4	ZNF423,N4BP1,CBLN1,C16orf78
	Chr16:72954547-73014090	1	CLEC18B
	Chr18:52339709-52856818	2	WDR7,TXNL1
	Chr18:63110881-63118233	0	
	Chr20:43763418-45210381	29	ACOT8,TP53RK,CD40,SNX21,PLTP,UBE2C,WFDC13,WFDC3,TNNC2,ZSWIM3,ZSWIM1,NEURL2,NCOA5,SLC35C2,ZNF334,SLC12A5,SLC2A10,LOC100240726,PCIF1,CDH22,C20orf123,ELMO2,SLC13A3,SPINT4,DNTTIP1,MMP9,ZNF335,CTSA,C20orf165
	Chr22:24019061-24248709	2	IGLL3,LRP5L

(B)

Trait	Copy Number regions	No. Genes	Genes
AST	Chr1:199155367-200961788	25	RNPEP,PTPN7,IGFN1,ELF3,PPP1R12B,UBE2T,SHISA4,SYT2,CACNA1S,TNNT2,TIMM17A,KIF21B,LAD1,NAV1,IPO9,GPR37L1,ARL8A,LGR6,LMOD1,PKP1,CSRP1,PHLDA3,TNNI1,TMEM9,RPS10P7
	Chr1:219016303-223441941	21	HHIPL2,SUSD4,DUSP10,TP53BP2,CNIH4,TLR5,DNAH14,C1orf65,TAF1A,MIA3,HLX,AIDA,NVL,WDR26,CNIH3,FAM177B,DISP1,FBXO28,DEGS1,CAPN2, CAPN8
	Chr10:72716105-73401893	4	CDH23,SLC29A3,C10orf54,PSAP
	Chr11:25870713-27225055	5	MUC15,SLC5A12,FIBIN,BBOX1,ANO3

	Chr15:39996964-40459070	6	PLA2G4E,VPS39,TMEM87A,PLA2G4D,PLA2G4F,GANC
	Chr16:14897352-14967222	1	NPIP
	Chr20:48228324-48562412	1	CEBPB
	Chr21:29429536-31660765	33	CLDN17,KRTAP26-1,KRTAP27-1,KRTAP23-1,KRTAP13-2,KRTAP13-4,KRTAP15-1,KRTAP19-2,KRTAP19-3,KRTAP19-4,KRTAP19-5,KRTAP19-6,KRTAP19-7,KRTAP6-3,KRTAP6-2,KRTAP22-1,KRTAP6-1,KRTAP20-1,KRTAP20-2,KRTAP21-2,KRTAP21-1,KRTAP8-1,KRTAP11-1,GRIK1,BACH1,KRTAP13-1,KRTAP13-3,KRTAP19-8,KRTAP20-3,KRTAP19-1,KRTAP24-1,CLDN8,KRTAP25-1
	Chr5:158988746-159912305	9	CCNJL,TTC1,SLU7,ADRA1B,FABP6,C5orf54,C1QTNF2,PTTG1,PWWP2A
	Chr6:67859000-70188705	1	BAI3
AST/ ALT	Chr1:173288426-174429270	3	TNN,TNR, SCARNA3
	Chr2:121450996-123193472	5	TSN,CLASP1,MKI67IP,TFCP2L1,RNU4ATAC
	Chr14:23001245-24313152	43	REC8,GZMH,NRL,CHMP4A,PSME2,DHRS2,RIPK3,AP1G2,SDR39U1,PSME1,LTB4R2,RABGGTA,IRF9,FITM1,C14orf21,NFATC4,GZMB,THTPA,LTB4R,NGDN,PCK2,TINF2,TM9SF1,FAM158A,DHRS4,TGM1,DHRS1,ADCY4,IPO4,GMPR2,TSSK4,JPH4,CMA1,RNF31,DHRS4L2,LRRC16B,CPNE6,CIDEB,CBLN3,NEDD8,CTSG,C14orf167,C14orf165
ALT	Chr7:142929078-143198980	1	FAM115C
	Chr12:107007904-107314707	2	WSCD2,CMKLR1
	Chr12:111893559-112564198	12	OAS2,SDS,SDSL,PLBD2,RASAL1,TPCN1,DTX1,DDX54,C12orf52,IQCD,SLC24A6,LHX5
	Chr12:52741396-53668999	16	ZNF385A,LACRT,ITGA5,SMUG1,HNRNPA1,DCD,PPP1R1A,MUCL1,COPZ1,CBX5,GPR84,NCKAP1L,PDE1B,NFE2,GTSF1,GLYCAM1
	Chr16:46036433-47063550	5	SIAH1,ABCC11,LONP2,PHKB,ABCC12
	Chr16:47069610-48586106	4	ZNF423,N4BP1,CBLN1,C16orf78
	Chr16:72954547-73014090	1	CLEC18B
	Chr18:52339709-52856818	2	WDR7,TXNL1
	Chr20:43763418-45210381	27	ACOT8,TP53RK,CD40,SNX21,PLTP,UBE2C,WFDC13,WFDC3,TNNC2,ZSWIM3,ZSWIM1,NEURL2,NCOA5,SLC35C2,ZNF334,SLC12A5,SL

		C2A10,PCIF1,CDH22,C20orf123,ELMO2,SLC13A3,SPINT4,DNTTIP1,MMP9,ZNF335,CTSA
Chr22:24019061-24248709	1	LRP5L

Supplementary Table 3.1. Summaries of the four diseases and one pathway associated with hepatic biomarkers AST or ALT.

	Name	ID	Definition
Diseases	hepatocellular carcinoma	MESH:D006528	A primary malignant neoplasm of epithelial liver cells.
	liver neoplasm	MESH:D008113	Tumors or cancer of the liver.
	liver cell adenoma	MESH:D018248	A benign epithelial tumor of the liver.
	drug-induced liver injury	MESH:D056486	A spectrum of clinical liver diseases ranging from biochemical abnormalities to acute liver failure, caused by drug metabolites.
Pathway	hepatitis C pathway	KEGG:05160	A major cause of chronic liver disease.

Supplementary Table 3.2. Summaries of non-redundant 22, 25, and 332 genes identified in the CEU, JPT, and YRI individuals.

Chromosome	CEU	JPT	YRI
Chr1		CREG1,REG4,LOC100129534,CKS1B,PRAMEF4	RBP7,EIF3I,LAMTOR2,ANKRD65,IFI6,VWA1,ISG20L2,KTI12,RPS27,FAM46B,TACSTD2,MIXL1,MIR181B1,PYCR2,LOR,IER5,IVL,CCDC28B,C1orf63,S100A2,AMIGO1,PITHD1,SNAPIN,FAM58BP,ZNF436,OR6K3,NUDT17,OR6N1,APOBEC4,LOC100506801
Chr2	PCBP1		MZT2A,MRPL53,CYP4F30P,HOXD8,ABHD1,PROM2,GPR148,UCN,ZFP36L2,TLX2,LIMS3,LOC100130451,SNORD94,FOXI3,LOC647012,NMUR1,RETSAT,ARL4C,LOC401010,GPR75,RDH14,PCGF1,HOXD12,GDF7,FER1L5,RESP18,PREB,LIMS3L,DQX1,BOLA3-AS1,CCDC74B
Chr3	PIGZ		MYNN,RTP2,FLJ42393,OXSM,SERP1,CYB561D2,SSR3,TLR9,ABHD14B,DNAJB8-AS1,C3orf71,GHSR,RPL35A,OR5H15,PAQR9,LOC401074
Chr4		LOC100287327	NAA11,PABPC4L,SFRP2,CXCL10,IL2,BBS12,MIR3138,DKFZP434I0714,LOC644248,CXCL6
Chr5	LOC100133050	PCDHB17	NPY6R,HIGD2A,APBB3,HINT1,LOC100132356,GPR151,MIR4803
Chr6	BAK1, TSPYL4		GGNBP1,NOL7,CAHM,RRP36,HIST1H3H,LOC100289495,CLPSL2
Chr7	HSPB1, HYALP1	GSTK1,GATSL2,TRIP6	GAL3ST4,ZNF394,C7orf34,EPHB6,GTF2IRD2,ATP6V1F,DLX5,ZNF467,FABP5P3,WBSCR28,HOTTIP,SNORA15,PRSS3P2,MPLKIP,SOSTDC1,ARHGEF35,STAG3L3

Chr8	SCXB, SCXA	SNHG6	LOC100133267,NUDT18,DKK4,SPAG11B,GPT,SPAG11A,MFSD3,DEFB4B,PROSC,FABP9,REXO1L2P,C8orf73,PMP2,DEFB107B,DEFB130,DEFB107A,PCAT1,C8orf69
Chr9	AQP7P3	LCN15,EXD3	TOMM5,LOC100128593,ASB6,LCN6,LINC00092,DPM2,IFNA16,HSPA5,CREB3,FAM122A,ANKRD20A3,LOC100129722,ANKRD20A2,LOC286297,C9orf173
Chr10		PGAM1,MRPS16	CHCHD1,SYCE1,PLAU,UTF1,TFAM,MIR603,FAM21C,MARK2P9
Chr11	OR5M1, OR4C6	ARL1	TIMM10,CTSW,TMEM133,LOC120824,C11orf1,LAMTOR1,CLP1,NUDT22,GYLTL1B,B3GAT3,LOC221122,SCGB1D4,C11orf24,APOC3,JRKL,APOA4,KCNA4,OR10A3,NUDT8,TRIM64C,OR9G4,OR8H1,OR52N2,OR52B6,OR51F2,OR5L2,OR10A2,POLD4
Chr12	LOC100506451,NANOGNB,LOC100505978,LOC100131733		MMP19,OR10AD1,C3AR1,CLEC1B,SLC9A7P1,C12orf39,AVPR1A,SP7,MYF5,SELPLG,DCD,C12orf68
Chr13			LINC00460
Chr14		SIVA1	MIR154,INF2,SNORD114-6,BCL2L2,OR10G3,LINC00523,RD3L
Chr15	NDNL2		LOC283663,SCARNA14,OR4N3P,RHOV,LOC100289656,SPATA8,LINC00593,C15orf59,ISLR,LOC253044,MEX3B
Chr16	NPW,BCL7C		ZNF688,VPS35,NAGPA,CMTM2,MT4,IRX6,RRN3,ATP6V0C,PSMB10,NME4,LOC653786,ASPHD1,PSMD7,ZNF785,CCDC101,LOC100128788,HCFC1R1,PRSS8,C16orf59,FOXC2,NRN1L,U

			BE2MP1,NTN3,NTAN1,ZNF689,EXOC3L1,SPN,PKD1P1,MARVELD3,MIR328,ELMO3,DDX11L10
Chr17		HIGD1B,KCTD11	KRT14,TMEM93,MRPS23,WNK4,MRPL27,CSF3,TSEN54,HAP1,CYB5D1,KRT33A,TBC1D3P2,SAT2,SPDYE4,PIPOX,C17orf102,GRB7,KRTAP4-4,RNASEK,KRTAP9-8,CSH1,KRTAP4-11,TUBG1,ORMDL3,MIR4726,LIMD2
Chr18			LOC644669,SLC25A52
Chr19	SIRT6	CIRBP,ICAM4	CCDC8,GCDH,TRAPPC2P1,CEBPG,SWSAP1,LRFN3,KIR2DL1,LOC100288123,DMRTC2,DNASE2,VN1R2,CALR,FPR1,RPL13AP5,SIGLEC16,LIN37,ZNF580,CLEC11A,RPL13A,MIR519B,LOC100134317,PPP1R15A,CEBPA
Chr20			FRG1B,LOC100131496,SPAG4,SCAND1,DEFB116,C20orf202,SUMO1P1
Chr21	OLIG2		LINC00163,KRTAP10-8,KRTAP12-3,TFF1,KRTAP10-7,LINC00162
Chr22	P2RX6P	C22orf29,DNAJB7	CHCHD10,ARVCF,CBX6,LGALS1,GALR3,C1QTNF6,FAM109B

Supplementary Table 3.3. Summaries of the pathways (A), drugs (B), and diseases (C) associated with ethnic disparities.

(A)

Ethnic	**Total numbers**	**Pathway ID**	**Pathway name**
CEU-JPT-YRI	3	REACT:111102	Signal Transduction
		KEGG:04740	Olfactory transduction
		KEGG:01100	Metabolic pathways
CEU-YRI	6	KEGG:04141	Protein processing in endoplasmic reticulum
		REACT:71	Gene Expression
		KEGG:00563	Glycosylphosphatidylinositol-anchor biosynthesis
		REACT:21257	Metabolism of RNA
		REACT:1675	mRNA Processing
		REACT:578	Apoptosis
CEU-JPT	0		
JPT-YRI	5	KEGG:05200	Pathways in cancer
		REACT:111217	Metabolism
		REACT:115566	Cell Cycle
		KEGG:04146	Peroxisome
		REACT:6900	Immune System
CEU	5	KEGG:00531	Glycosaminoglycan degradation
		KEGG:04010	MAPK signaling pathway
		KEGG:04370	VEGF signaling pathway
		KEGG:05146	Amoebiasis
		KEGG:03040	Spliceosome
JPT	7	KEGG:00980	Metabolism of xenobiotics by cytochrome P450
		KEGG:05222	Small cell lung cancer
		KEGG:00982	Drug metabolism - cytochrome P450
		REACT:115655	Metabolism

		KEGG:04621	NOD-like receptor signaling pathway
		KEGG:00480	Glutathione metabolism
		KEGG:00010	Glycolysis / Gluconeogenesis
YRI	100	KEGG:03013	RNA transport
		REACT:11123	Membrane Trafficking
		KEGG:04145	Phagosome
		KEGG:03060	Protein export
		KEGG:00250	Alanine, aspartate and glutamate metabolism
		KEGG:00230	Purine metabolism
		KEGG:04144	Endocytosis
		REACT:604	Hemostasis
		KEGG:04650	Natural killer cell mediated cytotoxicity
		REACT:383	DNA Replication
		REACT:78	Post-Elongation Processing of the Transcript
		KEGG:04020	Calcium signaling pathway
		KEGG:04940	Type I diabetes mellitus
		KEGG:05016	Huntington's disease
		KEGG:00520	Amino sugar and nucleotide sugar metabolism
		KEGG:03018	RNA degradation
		KEGG:05144	Malaria
		KEGG:00330	Arginine and proline metabolism
		KEGG:03440	Homologous recombination
		REACT:1788	Transcription
		KEGG:00532	Glycosaminoglycan biosynthesis - chondroitin sulfate
		KEGG:04140	Regulation of autophagy
		KEGG:04310	Wnt signaling pathway
		REACT:75800	Meiotic Synapsis (mouse)
		KEGG:05332	Graft-versus-host disease
		REACT:17015	Metabolism of proteins
		REACT:116125	Disease

REACT:111155	Cell-Cell communication
KEGG:00061	Fatty acid biosynthesis
KEGG:05150	Staphylococcus aureus infection
KEGG:04962	Vasopressin-regulated water reabsorption
KEGG:05320	Autoimmune thyroid disease
KEGG:05110	Vibrio cholerae infection
KEGG:03320	PPAR signaling pathway
KEGG:05322	Systemic lupus erythematosus
KEGG:05120	Epithelial cell signaling in Helicobacter pylori infection
KEGG:03420	Nucleotide excision repair
KEGG:04623	Cytosolic DNA-sensing pathway
KEGG:04620	Toll-like receptor signaling pathway
REACT:13505	Proteasome mediated degradation of PAK-2p34
KEGG:04916	Melanogenesis
KEGG:04350	TGF-beta signaling pathway
KEGG:04977	Vitamin digestion and absorption
KEGG:04610	Complement and coagulation cascades
KEGG:03010	Ribosome
KEGG:04360	Axon guidance
KEGG:04672	Intestinal immune network for IgA production
KEGG:04612	Antigen processing and presentation
KEGG:04914	Progesterone-mediated oocyte maturation
KEGG:04966	Collecting duct acid secretion
KEGG:04062	Chemokine signaling pathway
REACT:6850	Cdc20:Phospho-APC/C mediated degradation of Cyclin A
KEGG:04660	T cell receptor signaling pathway
KEGG:00190	Oxidative phosphorylation

KEGG:00380	Tryptophan metabolism
KEGG:00534	Glycosaminoglycan biosynthesis - heparan sulfate
KEGG:05330	Allograft rejection
KEGG:04975	Fat digestion and absorption
KEGG:00310	Lysine degradation
KEGG:00510	N-Glycan biosynthesis
KEGG:00561	Glycerolipid metabolism
KEGG:04622	RIG-I-like receptor signaling pathway
KEGG:04114	Oocyte meiosis
KEGG:05215	Prostate cancer
KEGG:05020	Prion diseases
KEGG:03030	DNA replication
KEGG:00760	Nicotinate and nicotinamide metabolism
KEGG:04080	Neuroactive ligand-receptor interaction
KEGG:05152	Tuberculosis
KEGG:00260	Glycine, serine and threonine metabolism
REACT:216	DNA Repair
REACT:1762	3' -UTR-mediated translational regulation
KEGG:04270	Vascular smooth muscle contraction
KEGG:00240	Pyrimidine metabolism
KEGG:03015	mRNA surveillance pathway
REACT:27166	Transcriptional Regulation of Adipocyte Differentiation in 3T3-L1 Pre-adipocytes
KEGG:05162	Measles
KEGG:05142	Chagas disease (American trypanosomiasis)
REACT:111183	Meiosis
KEGG:03430	Mismatch repair
KEGG:04640	Hematopoietic cell lineage
REACT:115492	Developmental Biology
KEGG:00564	Glycerophospholipid metabolism

REACT:111045	Developmental Biology
REACT:27235	Meiotic Recombination (mouse)
REACT:89750	Hemostasis
KEGG:05160	Hepatitis C
KEGG:04630	Jak-STAT signaling pathway
KEGG:03410	Base excision repair
KEGG:03050	Proteasome
KEGG:00071	Fatty acid metabolism
KEGG:00830	Retinol metabolism
KEGG:04060	Cytokine-cytokine receptor interaction
REACT:15518	Transmembrane transport of small molecules
KEGG:05323	Rheumatoid arthritis
KEGG:05221	Acute myeloid leukemia
KEGG:04514	Cell adhesion molecules (CAMs)
REACT:13685	Neuronal System
KEGG:05143	African trypanosomiasis
KEGG:04142	Lysosome

(B)

Ethnic	**Total numbers**	**Drug ID**	**Drug name**	**Indication**
CEU-JPT-YRI	42	DB00250	Dapsone	For the treatment and management of leprosy and dermatitis herpetiformis.
		DB00943	Zalcitabine	For the treatment of Human immunovirus infections.
		DB00648	Mitotane	For treatment of inoperable adrenocortical tumours.

DB01356	Lithium	Lithium is used as a mood stabilizer, and is used for treatment of depression and mania.
DB01169	Arsenic trioxide	For induction of remission and consolidation.
DB00369	Cidofovir	For the treatment of CMV.
DB01060	Amoxicillin	For the treatment of infections of the ear, nose, and throat, the genitourinary tract, the skin and skin structure.
DB00544	Fluorouracil	For the treatment of superficial basal cell carcinomas.
DB01101	Capecitabine	For the treatment of patients with metastatic breast cancer.
DB00126	Vitamin C	Used to treat vitamin C deficiency, scurvy, delayed wound and bone healing, urine acidificatio.
DB00563	Methotrexate	For the treatment of gestational choriocarcinoma.
DB00459	Acitretin	For the treatment of severe psoriasis in adults.
DB00091	Cyclosporine	For treatment of transplant rejection, rheumatoid arthritis.

DB00290	Bleomycin	For palliative treatment in lymphomas.
DB01206	Lomustine	For the treatment of primary and metastatic brain tumors.
DB01262	Decitabine	For treatment of patients with myelodysplastic syndromes French-American-British.
DB01234	Dexamethasone	For the treatment of endocrine disorders, rheumatic, dermatologic diseases, allergic statesc.
DB00262	Carmustine	For the treatment of brain tumors, multiple myeloma, Hodgkin's disease and Non-Hodgkin's lymphomas.
DB04690	Camptothecin	Investigated for the treatment of cancer.
DB00928	Azacitidine	For treatment of patients with the following French-American-British myelodysplastic syndrome subtypes.
DB01008	Busulfan	For use in combination with cyclophosphamide.
DB00997	Doxorubicin	For the treatment of Koposi's sarcome

		connected to AIDS.
DB00977	Ethinyl Estradiol	For treatment of moderate to severe vasomotor symptoms.
DB00381	Amlodipine	For the treatment of hypertension and chronic stable angina.
DB00787	Aciclovir	For the treatment and management of herpes zoster, genital herpes, and chickenpox
DB01143	Amifostine	For reduction in the cumulative renal toxicity in patients with ovarian cancer.
DB00678	Losartan	May be used as a first line agent to treat hypertension.
DB00681	Amphotericin B	Used to treat potentially life threatening fungal infections.
DB00322	Floxuridine	For palliative management of gastrointestinal adenocarcinoma metastatic to the liver.
DB00900	Didanosine	For use the treatment of HIV-1 infection in adults.
DB00640	Adenosine	Used as an initial treatment for the termination.
DB00515	Cisplatin	For the treatment of metastatic testicular tumors.
DB00970	Dactinomycin	For the treatment of Wilms' tumor,

		childhood rhabdomyosarcoma.
DB00851	Dacarbazine	For the treatment of metastatic malignant melanoma.
DB00959	Methylprednisolone	Adjunctive therapy for short-term administration.
DB00482	Celecoxib	For relief and management of osteoarthritis, rheumatoid arthritis.
DB06151	Acetylcysteine	Acetylcysteine is used as a mucolytic and in the management of paracetamol overdose.
DB00297	Bupivacaine	For the production of local or regional anesthesia or analgesia for surgery.
DB00163	Vitamin E	Vitamin E is protective against cardiovascular disease.
DB00317	Gefitinib	For the continued treatment of patients with locally advanced platinum-based or docetaxel chemotherapies.
DB00987	Cytarabine	For the treatment of acute non-lymphocytic leukemia, acute lymphocytic leukemia.

		DB00855	Aminolevulinic acid	For the treatment of moderately thick actinic keratoses of the face or scalp.
CEU-YRI	9	DB00499	Flutamide	For the management of locally confined Stage B2-C and Stage D2.
		DB01248	Docetaxel	For the treatment of patients with locally advanced or metastatic breast cancer after failure of prior chemotherapy.
		DB00668	Epinephrine	Used to treat anaphylaxis and sepsis.
		DB00248	Cabergoline	For the treatment of hyperprolactinemic disorders.
		DB00305	Mitomycin	For treatment of malignant neoplasm of lip, oral cavity.
		DB00242	Cladribine	For the treatment of active hairy cell leukemia.
		DB00958	Carboplatin	For the initial treatment of advanced ovarian carcinoma.
		DB00254	Doxycycline	Doxycycline is indicated for use in respiratory tract infections.
		DB01167	Itraconazole	For the treatment of the fungal infections pulmonary.
CEU-JPT	0			

JPT-YRI	5	DB00295	Morphine	For the relief and treatment of severe pain.
		DB00158	Folic Acid	For treatment of folic acid deficiency, megaloblastic anemia.
		DB00196	Fluconazole	For the treatment of fungal infections.
		DB00783	Estradiol	For the treatment of urogenital symptoms.
		DB00898	Ethanol	For therapeutic neurolysis of nerves or ganglia.
CEU	2	DB01177	Idarubicin	For the treatment of acute myeloid leukemia.
		DB00996	Gabapentin	For the management of postherpetic neuralgia.
JPT	0			
YRI	14	DB00281	Lidocaine	For production of local or regional anesthesia.
		DB00603	Medroxyprogesterone	Used as a contraceptive and to treat secondary amenorrhea.
		DB01592	Iron	Used in preventing and treating iron-deficiency anemia.
		DB01042	Melphalan	For the palliative treatment of multiple myeloma.
		DB00523	Alitretinoin	For topical treatment of cutaneous lesions.

	DB00422	Methylphenidate	For use as a treatment for a stabilizing with a behavioral chidren.
	DB01119	Diazoxide	Used parentally to treat hypertensive emergencies.
	DB01225	Enoxaparin	For the prophylaxis of deep vein thrombosis.
	DB00724	Imiquimod	For the topical treatment of clinically typical.
	DB00224	Indinavir	Indinavir is an antiretroviral drug of HIV infection.
	DB00333	Methadone	For the treatment of dry cough, drug withdrawal syndrome.
	DB01181	Ifosfamide	Used as a component of chemotherapeutic regimens.
	DB00448	Lansoprazole	For the treatment of acid-reflux disorders.
	DB00813	Fentanyl	For the treatment of cancer patients with severe pain.

(C)

Ethnic	**Total numbers**	**Disease ID**	**Disease name**
CEU-JPT-YRI	123	MESH:D009382	Neoplasms, Unknown Primary
		MESH:D003557	Phyllodes Tumor
		MESH:D000386	AIDS-Related Complex

MESH:D055756	Meningeal Carcinomatosis
MESH:D020202	Cerebral Hemorrhage, Traumatic
MESH:D000754	Anemia, Refractory, with Excess of Blasts
MESH:D055623	Keratosis, Actinic
MESH:D045262	Reticulocytosis
MESH:D015620	Histiocytic Disorders, Malignant
MESH:D002389	Catatonia
MESH:D009188	Myelitis, Transverse
MESH:D005134	Eye Neoplasms
MESH:C535533	Intrahepatic cholangiocarcinoma
MESH:D009894	Opportunistic Infections
MESH:D051346	Mobility Limitation
MESH:D006192	Haemophilus Infections
MESH:D007968	Leukoencephalopathy, Progressive Multifocal
MESH:D002921	Cicatrix
MESH:D019968	Sexual and Gender Disorders
MESH:D013899	Thoracic Neoplasms
MESH:D013086	Spermatic Cord Torsion
MESH:D011349	Proctitis
MESH:D014987	Xerostomia
MESH:D013832	Thiamine Deficiency
MESH:D054138	Sinus Arrest, Cardiac
MESH:D011529	Protozoan Infections, Animal
MESH:D054537	Atrioventricular Block
MESH:D006102	Granuloma, Laryngeal
MESH:D005356	Fibromyalgia
MESH:D014134	Tracheal Neoplasms
MESH:C535648	Familial primary gastric lymphoma
MESH:D004679	Encephalomyelitis
MESH:D009182	Mycosis Fungoides
MESH:D004695	Endocardial Fibroelastosis

MESH:D010997	Pleural Neoplasms
MESH:D015840	Oculomotor Nerve Diseases
MESH:D016919	Meningitis, Cryptococcal
MESH:D007232	Infant, Newborn, Diseases
MESH:D012872	Skin Diseases, Vesiculobullous
MESH:D005185	Fallopian Tube Neoplasms
MESH:D011252	Pregnancy Complications, Neoplastic
MESH:D011832	Radiation Injuries
MESH:D020434	Abducens Nerve Diseases
MESH:D004483	Ectropion
MESH:D013924	Thrombophlebitis
MESH:D018785	Tricuspid Atresia
MESH:D015866	Uveitis, Posterior
MESH:D057896	Striae Distensae
MESH:D006646	Histiocytosis, Langerhans-Cell
MESH:D000757	Anencephaly
MESH:D010255	Paranasal Sinus Neoplasms
MESH:D006562	Herpes Zoster
MESH:D007019	Hypoproteinemia
MESH:D003139	Common Cold
MESH:D054438	Leukemia, Myeloid, Chronic, Atypical, BCR-ABL Negative
MESH:D020232	Kluver-Bucy Syndrome
MESH:D016411	Lymphoma, T-Cell, Peripheral
MESH:D020828	Pseudobulbar Palsy
MESH:C535668	Adrenocorticotropic hormone deficiency
MESH:D007638	Keratoconjunctivitis Sicca
MESH:D013684	Telangiectasis
MESH:D000381	Agraphia
MESH:D018268	Adrenocortical Carcinoma

MESH:D014328	Trophoblastic Neoplasms
MESH:D045745	Scleroderma, Limited
MESH:C538525	Mitochondrial encephalopathy
MESH:D001762	Blepharitis
MESH:D004172	Diplopia
MESH:C536495	VACTERL association
MESH:D003218	Condylomata Acuminata
MESH:D002283	Carcinoma, Bronchogenic
MESH:D009006	Monosomy
MESH:D002291	Carcinoma, Papillary
MESH:D019559	Capillary Leak Syndrome
MESH:D010236	Paraganglioma, Extra-Adrenal
MESH:C538370	Retroperitoneal liposarcoma
MESH:D013952	Thymus Hyperplasia
MESH:D007829	Laryngostenosis
MESH:D004379	Duodenal Neoplasms
MESH:D004407	Dysgerminoma
MESH:D044504	Enterocolitis, Neutropenic
MESH:D018236	Carcinoma, Embryonal
MESH:D002494	Central Nervous System Infections
MESH:D025242	Spondylarthropathies
OMIM:146850	IMMUNE SUPPRESSION
MESH:D060831	Hand-Foot Syndrome
MESH:D010307	Parotid Neoplasms
MESH:D011128	Polyradiculopathy
MESH:D020237	Alexia, Pure
MESH:D055154	Dysphonia
MESH:C537844	Nonseminomatous germ cell tumor
MESH:D005533	Foot Dermatoses
MESH:D002357	Cartilage Diseases
MESH:D004933	Esophageal Atresia
MESH:D012811	Sigmoid Neoplasms
MESH:D020240	Apraxia, Ideomotor
MESH:D018325	Hemangioblastoma

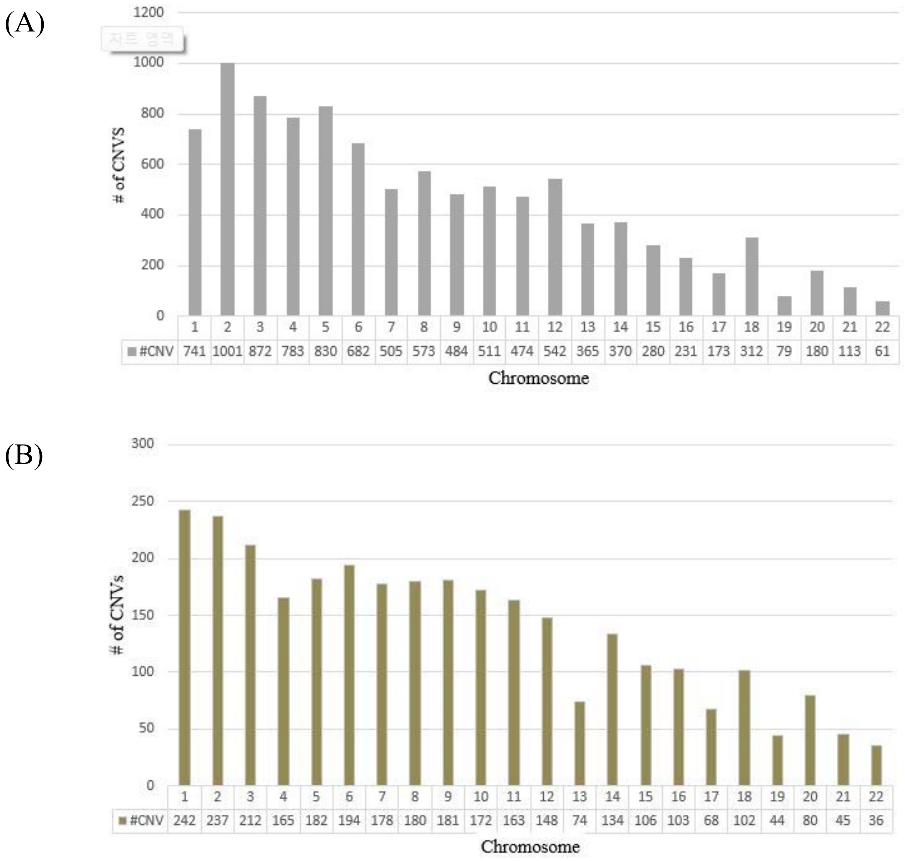
		MESH:D001984	Bronchial Neoplasms
		MESH:D013122	Spinal Diseases
		MESH:D016543	Central Nervous System Neoplasms
		MESH:D020149	Manganese Poisoning
		MESH:D008480	Mediastinitis
		MESH:D010257	Paraneoplastic Syndromes
		OMIM:613290	HEARING LOSS, CISPLATIN-INDUCED, SUSCEPTIBILITY TO
		MESH:D014523	Urethral Neoplasms
		MESH:D004443	Echinococcosis
		MESH:D016918	Arthritis, Reactive
		MESH:D004701	Endocrine Gland Neoplasms
		MESH:D016781	Toxoplasmosis, Cerebral
		MESH:D020069	Shoulder Pain
		MESH:D016400	Lymphoma, Large-Cell, Immunoblastic
		MESH:D010192	Pancreatic Pseudocyst
		MESH:D009442	Neurilemmoma
		MESH:D012817	Signs and Symptoms, Digestive
		MESH:D007007	Hypohidrosis
		MESH:D015490	HTLV-I Infections
		MESH:D005155	Facial Nerve Diseases
		MESH:D054038	Posterior Leukoencephalopathy Syndrome
		MESH:D015477	Leukemia, Myelomonocytic, Chronic
		MESH:C538011	Eales disease
		MESH:D008664	Metal Metabolism, Inborn Errors
		MESH:D007939	Leukemia L1210
		MESH:D010304	Paronychia
CEU-YRI	39	MESH:D007953	Leukemia, Radiation-Induced
		MESH:D015448	Leukemia, B-Cell
		MESH:D014719	Vesicovaginal Fistula
		MESH:D014627	Vaginitis

MESH:D007499	Iris Diseases
MESH:D002528	Cerebellar Neoplasms
MESH:D018227	Sarcoma, Clear Cell
MESH:D001025	Aortitis
MESH:D007943	Leukemia, Hairy Cell
MESH:D004940	Esophageal Stenosis
MESH:D018410	Pneumonia, Bacterial
MESH:D015422	Scleral Diseases
MESH:D019462	Syncope, Vasovagal
MESH:D010167	Pallor
MESH:D001747	Urinary Bladder Fistula
MESH:D012912	Sneezing
MESH:D013492	Suppuration
MESH:D051302	Paroxysmal Hemicrania
MESH:D001661	Biliary Tract Neoplasms
MESH:D002828	Choristoma
MESH:D048550	Hepatic Insufficiency
MESH:D020518	Focal Nodular Hyperplasia
MESH:D017714	Community-Acquired Infections
MESH:D010390	Pemphigoid, Benign Mucous Membrane
MESH:D010034	Otitis Media with Effusion
MESH:D007500	Iritis
MESH:D002575	Uterine Cervicitis
MESH:D009209	Myofascial Pain Syndromes
MESH:D010033	Otitis Media
MESH:D034321	Hyperamylasemia
MESH:C536783	T-Lymphocytopenia
MESH:D017577	Cutaneous Fistula
MESH:C538268	Auditory neuropathy
MESH:D057112	Corneal Perforation
MESH:D015792	Retinal Dysplasia
MESH:D015441	Leprosy, Tuberculoid
MESH:D014262	Tricuspid Valve Insufficiency

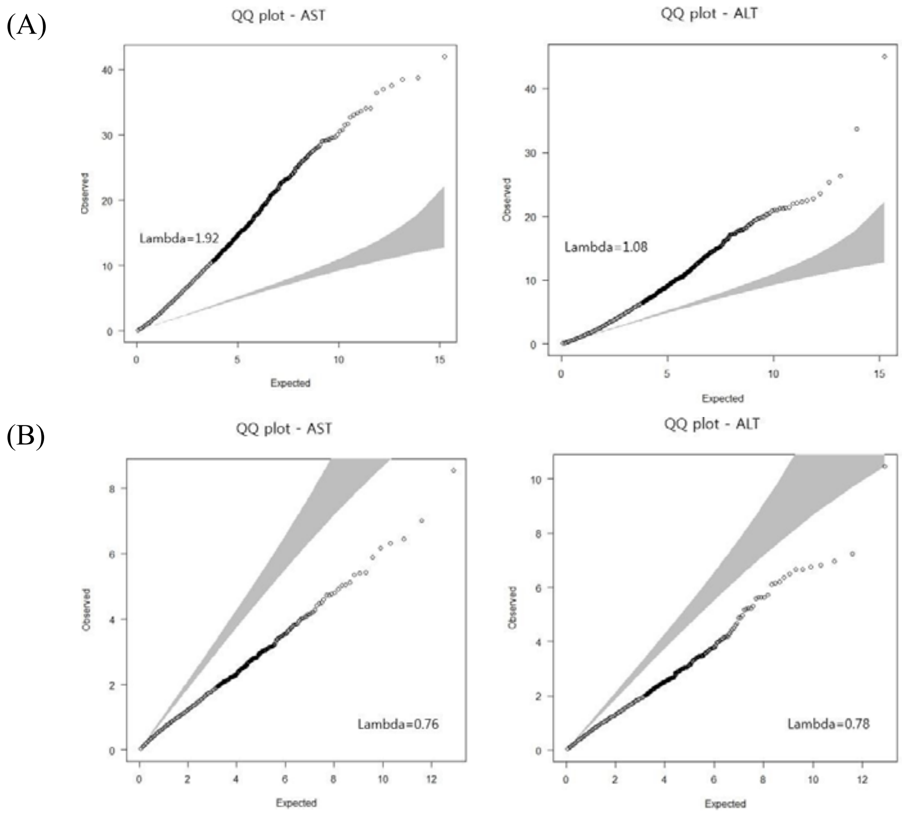
		MESH:D020516	Brachial Plexus Neuropathies
		MESH:D014550	Urinary Incontinence, Stress
CEU-JPT	0		
JPT-YRI	24	MESH:D014009	Onychomycosis
		MESH:D018302	Neoplasms, Neuroepithelial
		MESH:D013978	Tibial Fractures
		MESH:D001206	Ascorbic Acid Deficiency
		MESH:D006558	Herpes Genitalis
		MESH:D016388	Tooth Loss
		MESH:D018677	Tooth Injuries
		MESH:D020277	Polyradiculoneuropathy, Chronic Inflammatory Demyelinating
		MESH:D001657	Biliary Dyskinesia
		MESH:D014008	Tinea Pedis
		MESH:C535464	Conotruncal cardiac defects
		MESH:D020268	Alcohol-Induced Disorders, Nervous System
		MESH:D014860	Warts
		MESH:D006560	Herpes Labialis
		MESH:D001028	Aortopulmonary Septal Defect
		MESH:D027601	Polyomavirus Infections
		MESH:D005242	Fecal Incontinence
		MESH:D012614	Scurvy
		MESH:D020918	Complex Regional Pain Syndromes
		MESH:D006819	Hyaline Membrane Disease
		MESH:D013182	Sprue, Tropical
		MESH:C531767	Edema of the optic disc
		MESH:D048949	Labor Pain
		MESH:D000267	Tissue Adhesions
CEU	3	MESH:D010591	Phantom Limb
		MESH:D020432	Trochlear Nerve Diseases
		MESH:D014847	Vulvitis

JPT	0		
YRI	46	MESH:D011528	Protozoan Infections
		MESH:C535700	Malignant mesenchymal tumor
		MESH:D001759	Blastomycosis
		MESH:D003291	Conversion Disorder
		MESH:D008172	Lung Diseases, Fungal
		MESH:D004184	Dirofilariasis
		MESH:D004454	Echolalia
		MESH:D046350	Porphyria, Variegate
		MESH:C538542	Sexual precocity
		MESH:C537095	Brachydactyly with hypertension
		MESH:D018746	Systemic Inflammatory Response Syndrome
		MESH:D006944	Hyperglycemic Hyperosmolar Nonketotic Coma
		MESH:D023981	Sarcoma, Myeloid
		MESH:D003440	Croup
		MESH:D000377	Agnosia
		MESH:D020206	Subarachnoid Hemorrhage, Traumatic
		MESH:C531616	Primary amyloidosis
		MESH:D045602	Steatorrhea
		MESH:D016460	Granuloma Annulare
		MESH:D056650	Vulvodynia
		MESH:D053120	Respiratory Aspiration
		MESH:D016574	Seasonal Affective Disorder
		MESH:D020220	Facial Nerve Injuries
		MESH:D015673	Fatigue Syndrome, Chronic
		MESH:D006491	Hemothorax
		MESH:D008946	Mitral Valve Stenosis
		MESH:C536855	Fanconi like syndrome
		MESH:D006660	Histoplasmosis
		MESH:D003874	Dermatitis Herpetiformis
		MESH:C536610	Familial cerebral cavernous malformation
		MESH:D004842	Epispadias

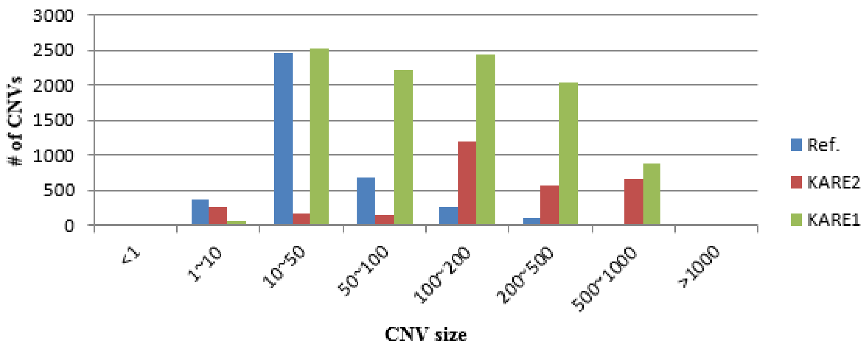
	MESH:C537372	Multi-centric Castleman's Disease
	MESH:D008260	Macroglossia
	MESH:D020433	Trigeminal Nerve Diseases
	MESH:D002279	Carcinoma 256, Walker
	MESH:D014010	Tinea Versicolor
	MESH:D017789	Granuloma, Pyogenic
	MESH:D006551	Hernia, Hiatal
	MESH:D001988	Bronchiolitis
	MESH:D000274	Adiposis Dolorosa
	MESH:D016263	AIDS-Associated Nephropathy
	MESH:D018437	Brown-Sequard Syndrome
	MESH:C538169	Acitretin embryopathy
	MESH:D017499	Porokeratosis
	MESH:D034161	Pelvic Infection
	MESH:D021081	Chronobiology Disorders



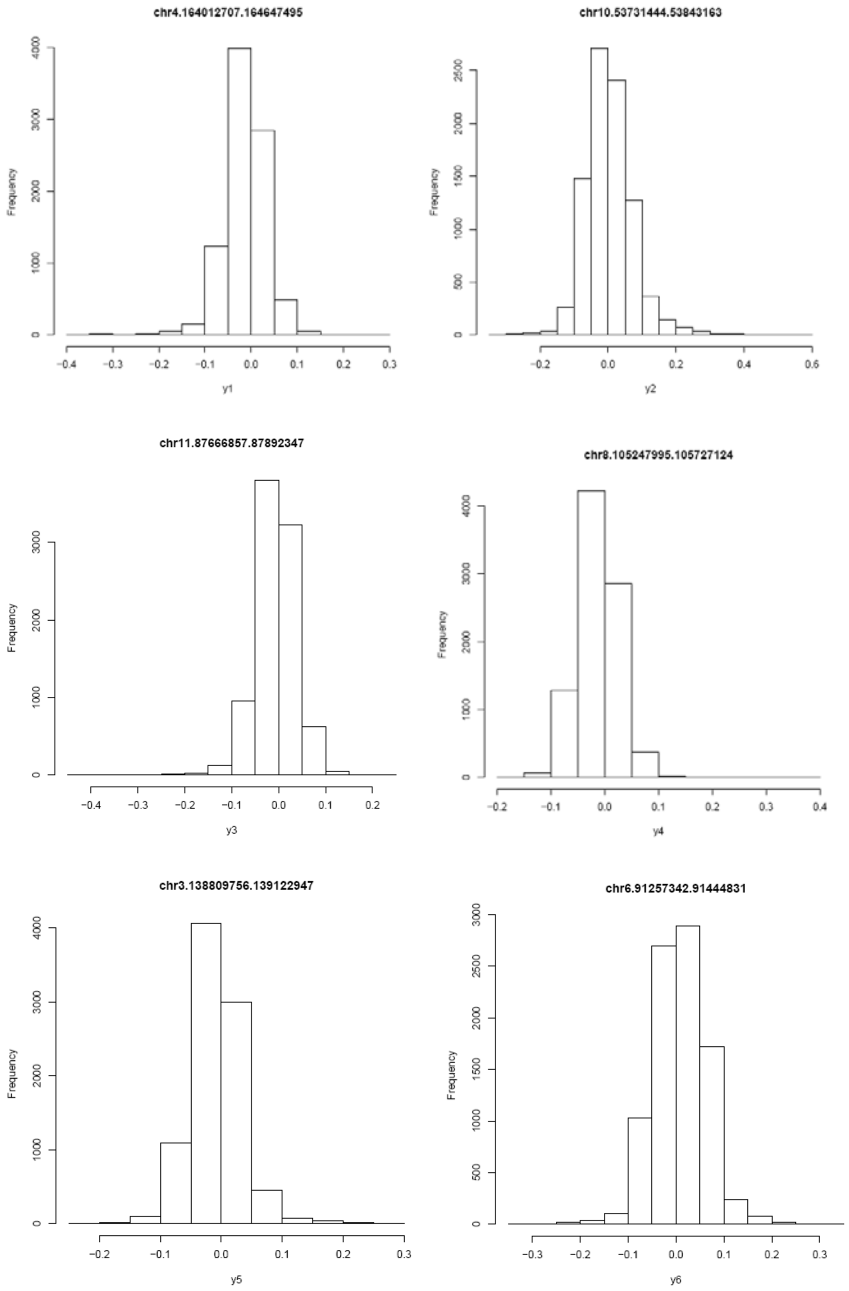
Supplementary Figure 2.1. Bar graph of the total number of the significant CNVs on each chromosome in KARE1 (A) and KARE2 (B).

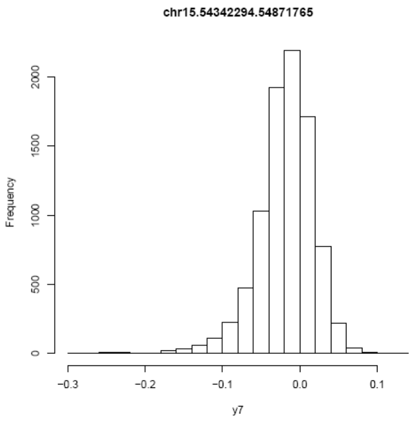


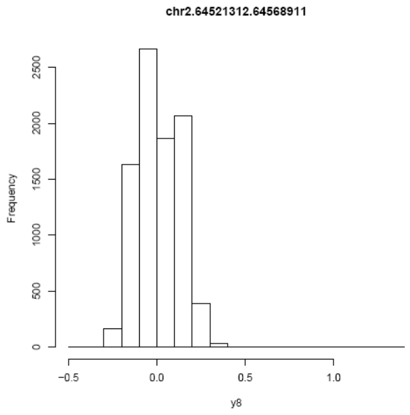
Supplementary Figure 2.2. Quantile-quantile (QQ) plot for AST or ALT results for GWAS in KARE1 (A) and KARE2 (B). In the QQ plot, the horizontal and the vertical axis indicates expected and observed *p*-values, respectively. The lambda values (median[obs]/median[exp]) is shown.

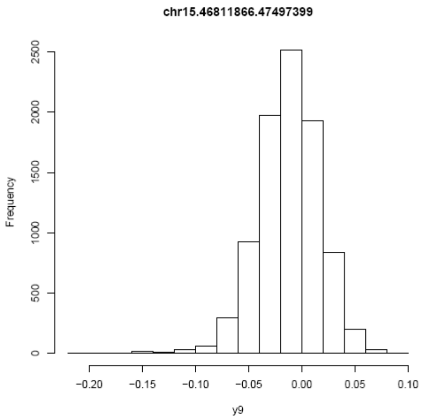


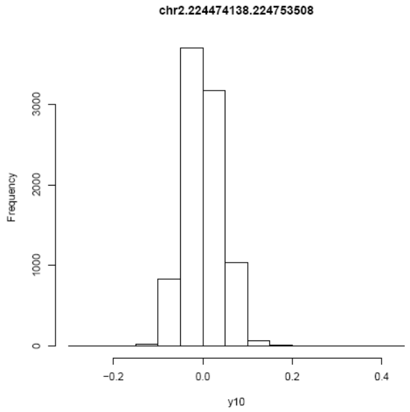
Supplementary Figure 2.3. The distribution of the number of CNVs in this study compare to previously found CNVs in same Korean populations.

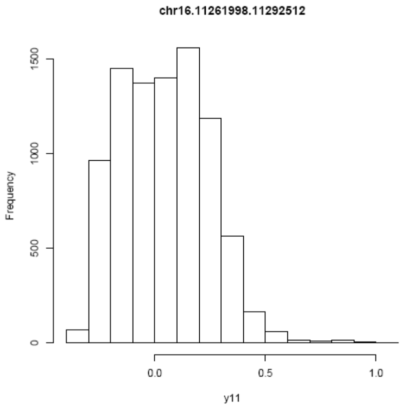

chr4.164012707.164647495
Frequency
y1
chr10.53731444.53843163
Frequency
y2
chr11.87666857.87892347
Frequency
y3
chr8.105247995.105727124
Frequency
y4
chr3.138809756.139122947
Frequency
y5
chr6.91257342.91444831
Frequency
y6

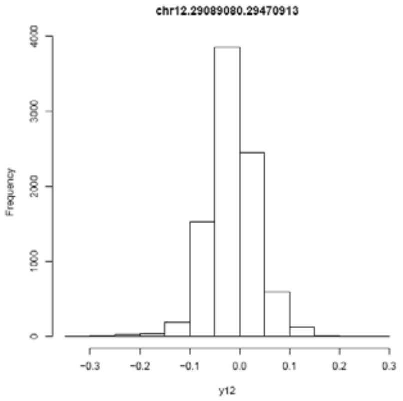

chr15.54342294.54871765
Frequency
0
500
1000
1500
2000
−0.3
−0.2
−0.1
0.0
0.1
y7

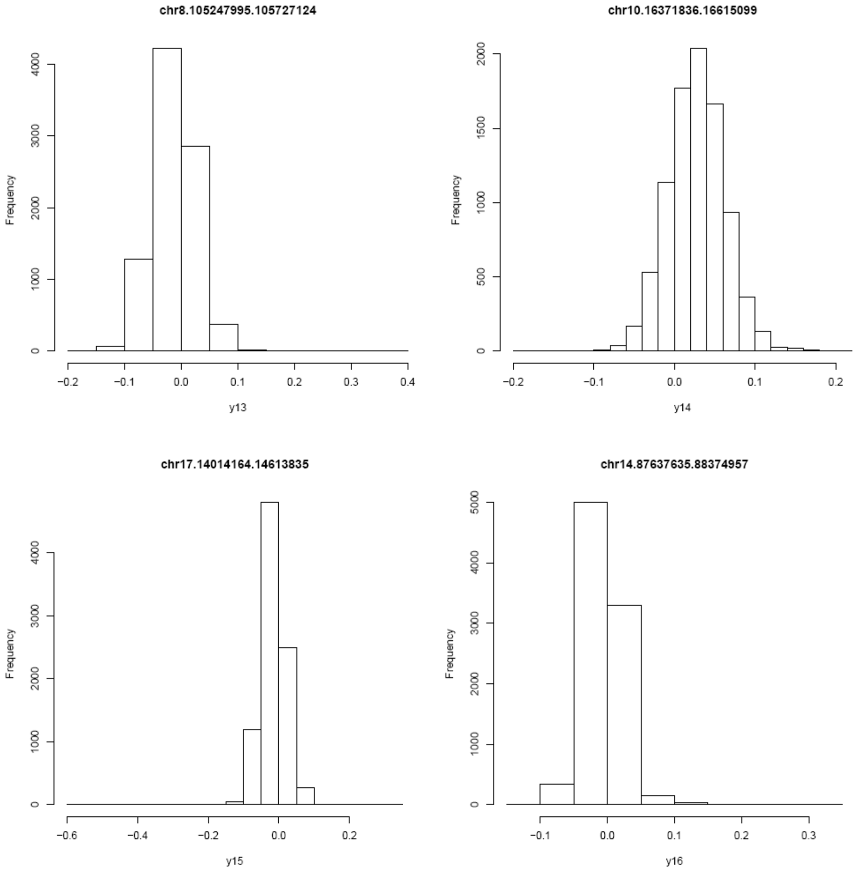

chr2.64521312.64568911
Frequency
0
500
1000
1500
2000
2500
−0.5
0.0
0.5
1.0
y8


chr15.46811866.47497399
Frequency
0
500
1000
1500
2000
2500
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
y9

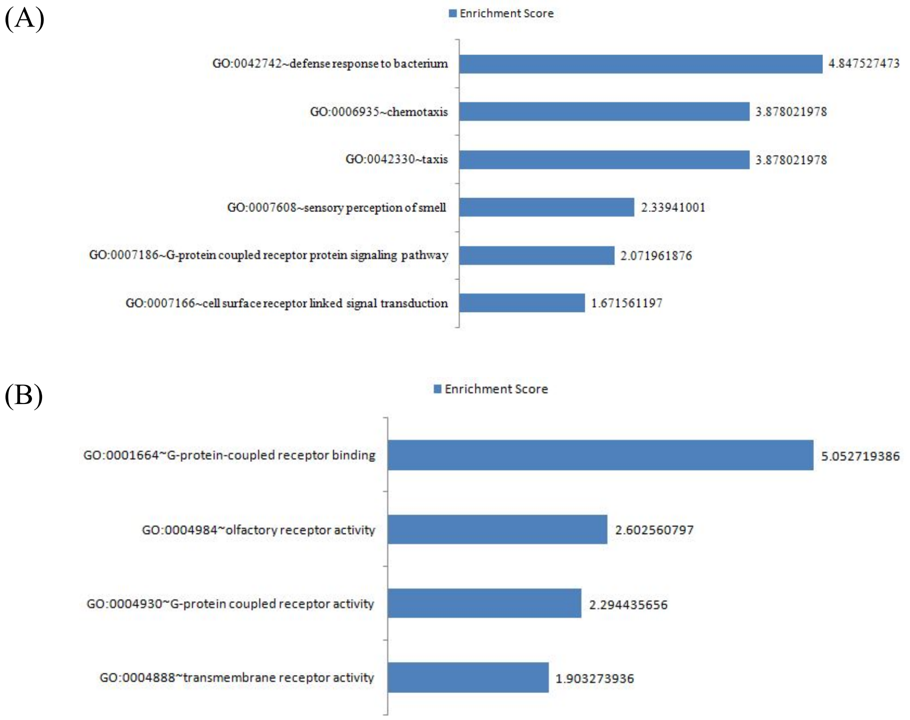

chr2.224474138.224753508
Frequency
0
1000
2000
3000
−0.2
0.0
0.2
0.4
y10


chr16.11261998.11292512
Frequency
0
500
1000
1500
0.0
0.5
1.0
y11


chr12.29089080.29470913
Frequency
0
1000
2000
3000
4000
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
y12



Supplementary Figure 2.4. The log2 ratio distributions of the 16 significant CNVs.



Supplementary Figure 3.1. Gene functional classifications for AST or ALT based on the DAVID tool. All categories are with the significant enrichment groups, with ranges of 1.67-4.85 (A: Biological Process) and 1.9-5.05 (B: Molecular Function).

요약(국문초록)

구조 변이 기반 인간 게놈 특성 규명을 위한 생물정보학 연구

김효영

농생명공학부 동물생명공학전공

서울대학교 대학원 농업생명과학대학

지난 몇 년 동안 질병 관련 유전체 구조적 변이 (단일염기 다형성과 유전자 복제 수 변이) 연구에 대한 노력이 계속되고 있다. 단일염기 다형성은 참조유전체와 비교하여 DNA 염기서열에서 하나의 염기서열의 차이를 가지고 유전자 복제 수 변이는 1,000 개 이상의 구조적 변이이다. 전장유전체연관분석은 유전체 구조적 변이와 질병에 관한 후보유전자를 찾는데 많이 연구되고 있다. 데이터 마이닝은 복잡하고 많은 양의 정보를 통찰하는데 중요하다. 이러한 생물학적 네트워크는 연구자가 정보를 통하여 복잡한 문제에 대한 의미론적 해답을 찾는데 도움을 준다. 따라서, 이 논문의 목표는 한국인에서 간 질병과 관련된 유전적 변이를 찾고, 간 기능이나 인종 차이에 영향을 미치는 생물학적 네트워크를

구축하여 이에 대한 의미론적 해답을 찾고 유전체 구조적 변이에 대한 시각화 툴을 구축하는데 있다.

제 1 장에서는 유전자 복제 수 변이, 전장유전체연관분석과 생물학적 네트워크에 관하여 기술하였다. 1) 유전자 복제 수 변이에 대한 개요와 원천 및 찾는 방법을 기술하였고 연구동향과 질병에서의 역할을 정리하였다. 2) 전장유전체연관분석에 대한 개요와 배경을 정리하였고 방법 및 결과를 요약하였다. 3) 생물학적 네트워크에 관한 개요 및 연구동향을 정리하였다.

제 2 장에서는 한국인에 관한 간 형질과 유전자 복제 수 변이의 메타연관분석을 수행하였다. KARE1 파트에서는 1) 한국인 8,842 명에 대해 총 10,162 개의 유전자 복제 수 변이를 찾았고, 2) 간 형질에 대한 유전자 복제 수 변이의 영향을 보기 위하여 단일 선형 회귀 분석을 수행하였다. 그 결과, AST 와 ALT 에 대해서 각각 100 개와 16 개가 유의하게 나왔다. 3) 그 유의한 유전자 복제 수 변이의 지역에 39 개의 유전자가 위치해 있었고 4) 그 유전자에 대해 기능적 분류 분석 결과, 간 관련 후보유전자로서 인정이 되었다. KARE2 파트에서는 KARE1 파트의 반복 유전체연관분석을 수행하였다. 1) 한국인 407 명에 대해 총 3,046 개의 유전자 복제 수 변이를 찾았고, 2) 단일 선형 회귀 분석을 이용하여 유전자 복제 수 변이와 간 형질과의 연관분석을 수행하였다. 그 결과, AST 와 ALT 에 대해서 각각 32 개 (140 개의 유전자)와 42 개 (172 개의 유전자)가 유의하게 나왔다. 3) 반복분석결과, 한국인의 유전자 복제 수 변이와 간 관련하여 총 9 개의 유전자가 유의하게 나왔다.

제 3 장에서는 간 기능과 인종 차이를 나타내는 유전자 복제 수 관련 생물학적 네트워크를 구축하였다. 노드는 유전자, 질병, 대사, 화학물질, 약, 임상정보, 변이 등으로 구성되어있고, 연결은 유전자-질병, 유전자-변이, 유전자-화학물질, 대사-질병, 대사-화학물질, 화학물질-약, 질병-임상정보, 임상정보-약 등으로 구성되어있다. 생물학적 네트워크 분석을 통해 한국인 간 기능 유전자 복제 수 변이 관련 총 4 개의 질병과 1 개의 대사회로 및 7 개의 약을 밝혀내었고, 인종 차이 유전자 복제 수 변이 관련 총 3 개의 질병과 1 개의 약 및 5 개의 대사회로를 밝혀내었다.

제 4 장에서는 유전자 복제 수 변이와 단일염기다형성의 시각화를 위한 툴을 구축하였다. 총 6 개의 메뉴로 1) 유전자 복제 수 변이나 단일염기다형성의 위치에 풍부한 요소 검사와 2) 염색체상의 변이 위치 분포 3) log2 ratio 분포 4) binning 단위 당 변위 분포 5) homozygosity 분포 6) cytomapping 시각화로 구성되어있다. 이 툴은 값으로 나타나는 변이로부터 생물학적 의미를 쉽게 이해하는데 도움을 주고, 또한 어떤 설치나 다운로드 없이 쉽게 이용 가능하다.

전장유전체 연관분석을 통해 한국인의 유전자 복제 수 변이와 간 형질 관련 유력한 후보유전자를 찾을 수 있었고, 간 질병과 인종차이 유전자 복제 수 변이관련 의미론적 생물학 네트워크를 구축할 수 있었다. 또한 다양한 유전자 복제 수 변이 연구를 함으로써 축적되어온 변이 시각화를 위한 총집합적 툴을 개발하였다. 이러한 네트워크와 시각화 툴은 질병이나 인종 관련

유전자 복제 수 변이의 의미론적 생물학 의미 발견이 가능하고 시각화 툴은 값으로 나타나는 유전자 복제 수 변이로부터 생물학적 해석에 도움이 된다.

주요어: 생물학적 네트워크, 시각화, 유전자 복제 수 변이, 전장유전체연관분석, 한국인.

학번: 2007-30877