EXPLORING COPY NUMBER VARIATIONS IN A THAI POPULATION

CHAIWAT NAKTANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (BIOCHEMISTRY)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2015

Thesis
entitled

EXPLORING COPY NUMBER VARIATIONS IN A THAI POPULATION

.....
Mr. Chaiwat Naktang
Candidate

……
Lect. Varodom Charoensawan,
Ph.D. (Molecular Biology)
Major advisor

……
Lect. Natini Jinawath,
M.D., Ph.D. (Molecular Pathology)
Co-advisor

……...
Prof. Sumalee Tungpradapkul,
Ph.D. (Molecular Biology)
Co-advisor

……
Lect. Bhoom Suktitipat,
M.D., Ph.D. (Genetic Epidemiology)
Co-advisor

……
Prof. Banchong Mahaisavariya,
M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

…….……...
Asst. Prof. Kittisak Yokthongwattana,
Ph.D. (Agricultural & Environmental Chemistry)
Program Director
Master of Science Program in Biochemistry
Faculty of Science, Mahidol University

Thesis
entitled

EXPLORING COPY NUMBER VARIATIONS IN A THAI POPULATION

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Biochemistry)
on
February 24, 2015

.....
Mr. Chaiwat Naktang
Candidate

……..	……
Lect. Varodom Charoensawan, Ph.D. (Molecular Biology) Member	Prof. Mathurose Ponglikitmongkol, Ph.D. (Molecular Biology) Chair
……..	……...
Lect. Bhoom Suktitipat, M.D., Ph.D. (Genetic Epidemiology) Member	Lect. Natini Jinawath, M.D., Ph.D. (Molecular Pathology) Member
……..	……...
Prof. Sumalee Tungpradapkul, Ph.D. (Molecular Biology) Member	Mr. Surakameth Mahasirimongkol, M.D., Ph.D. (International Health) Member
……..	……
Prof. Banchong Mahaisavariya, M.D., Dip Thai Board of Orthopedics Dean Faculty of Graduate Studies Mahidol University	Prof. Skorn Mongkolsuk, Ph.D. (Biological Science) Dean Faculty of Science, Mahidol University

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Varodom Charoensawan, for his support, insights and suggestions to my thesis. Moreover, he always kindly listens and to gives advice. In addition, I would like to give a special thanks to my co-advisor, Dr. Natini Jinawath, for all the invaluable supervision and background genetic knowledge that helps complete my thesis. I also thank Dr. Bhoom Suktitipat for his time in various helpful discussions. Last, but not least, I would like to thank Mr.Wuttichai Mhuantong, Miss Thitima Tularak and Miss Paramita Artiwet for their kinds help on the CNV calling and suggestions about programming languages.

Chaiwat Naktang

EXPLORING COPY NUMBER VARIATIONS IN A THAI POPULATION

CHAIWAT NAKTANG 5536368 SCBC/M

M.Sc.(BIOCHEMISTRY)

THESIS ADVISORY COMMITTEE: VARODOM CHAROENSAWAN, Ph.D.
NATINI JINAWATH, M.D., Ph.D. BHOOM SUKTITIPAT, M.D., Ph.D.
SUMALEE TUNGPRADAPKUL, Ph.D.

ABSTRACT

Copy Number Variation (CNV) is one of the major structural variations in a human genome. CNV has been associated with several human diseases such as neurodevelopmental diseases, including Autism Spectrum Disorder (ASD), and neuropsychiatric diseases. Recently, reference CNVs in normal subjects from certain populations, such as African-American, Caucasian and East Asian, are available from a number of CNV databases. These CNV databases can facilitate clinical interpretation of CNVs, which can be categorized into three main groups: pathogenic (disease-related), unknown clinical significant, or benign. So far there is no normal CNV database available for the Thais, and existing CNV databases of different ethnic groups are by no mean an ideal reference for the CNV interpretation for the Thai population, due to divergent genetic backgrounds. In this study, we combine the genome-wide Single Nucleotide Polymorphism (SNP) genotyping data from previous studies, consisting of 3,017 Thai subjects with no known genetic disorders. We perform CNV discovery from these datasets using PennCNV and CNV Workshop software to ensure the highest possible confident of CNV calls, and using the combining CNV sets to create the largeest CNV reference for the Thais to date. Moreover, we perform population analysis by using the program Plink to compare the Thai population with eleven HAPMAP3 populations. Hierarchical clustering analysis (HCA) using frequency of candidate genes is used to assess similarity between the Thai population and other HAPMAP3 populations. The results show that CNVs found in the Thai population cluster with other Asian populations. Having population-specific CNV database will improve the accuracy for the interpretation of clinical significant CNVs in the Thais, and serve as one of the most informative population-specific CNV reference databases for population geneticists.

KEY WORDS: COPY NUMBER VARIATION (CNV) / SINGLE NUCLEOTIDE POLYMORPHISM (SNP) / THAI POPULATION / DATABASE

80 pages

การสำรวจหาการแปรผันของจำนวนชุดดีเอ็นเอในประชากรคนไทย

EXPLORING COPY NUMBER VARIATIONS IN A THAI POPULATION

ชัยวัฒน์ นาคทั่ง 5536368 SCBC/M

วท.ม. (ชีวเคมี)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : วโรดม เจริญสวรรค์, Ph.D., ณฐินี จินาวัฒน์, M.D., Ph.D
ภูมิ สุขธิติพัฒน์, M.D., Ph.D., สุมาลี ตั้งประดับกุล, Ph.D

บทคัดย่อ

การแปรผันของจำนวนชุดดีเอ็นเอ (Copy Number Variation) เป็นหนึ่งในการแปรผันทางโครงสร้าง (Structural Variation) ที่สำคัญในข้อมูลทางพันธุกรรมทั้งหมดของมนุษย์ (human genome) ในปัจจุบันการแปรผันของจำนวนชุดดีเอ็นเอได้เกี่ยวข้องกับโรคที่เกิดในมนุษย์หลายๆโรคด้วยกันเช่น โรคทางระบบประสาทที่เกี่ยวข้องกับพัฒนาการ เช่น โรคออทิสติกและโรคทางจิตเวช ในปัจจุบันได้มีการสร้างฐานข้อมูลของการแปรผันของจำนวนชุดดีเอ็นเอในคนปกติจากหลายๆกลุ่มประชากร เช่น คอเคเชี่ยน แอฟริกัน-อเมริกา เอเชียตะวันออก ซึ่งข้อมูลเหล่านี้จะสามารถช่วยในการแปรผลทางคลีนิกของการแปรผันของจำนวนชุดดีเอ็นเอที่พบในคนไทย ซึ่งสามารถแบ่งออกมาได้สามกลุ่มดังนี้ 1.กลุ่มที่ก่อให้เกิดโรค, 2.กลุ่มที่ยังไม่ทราบความสำคัญ, 3.กลุ่มที่ไม่ก่อให้เกิดโรค แต่เราพบว่าในฐานข้อมูลเหล่านี้ไม่มีข้อมูลของการแปรผันของจำนวนชุดดีเอ็นเอในคนไทย จึงทำให้ไม่สามารถนำไปใช้ในการแปรผลของการแปรผันของจำนวนชุดดีเอ็นเอในคนไทยได้ เนื่องจากความหลากหลายทางพันธุกรรมระหว่างกลุ่มประชากรนั้นแตกต่างกัน โดยในงานวิจัยชิ้นนี้คณะผู้ทำวิจัยได้รวบรวมข้อมูลจาก SNP Genotyping ซึ่งประกอบไปด้วยคนไทยจำนวน 3,017 คนและไม่มีการรายงานโรคทางพันธุกรรมในกลุ่มคนไทยเหล่านี้ โดยคณะผู้ทำวิจัยได้การสำรวจหาการแปรผันของจำนวนชุดดีเอ็นเอจากข้อมูลเหล่านี้โดยใช้โปรแกรม PennCNV และ CNV Workshop ซึ่งทั้งสองโปรแกรมนี้เป็นที่ใช้สำหรับหาการแปรผันของจำนวนชุดดีเอ็นเอจาก SNP Genotyping Array และเพื่อความแม่นยำที่สูงขึ้นคณะผู้ทำการวิจัยได้ใช้ข้อมูลจากทั้งสองโปรแกรมเพื่อใช้ในการสร้างฐานข้อมูลของการแปรผันของจำนวนชุดดีเอ็นเอในกลุ่มประชากรคนไทย นอกจากนั้นคณะผู้ทำการวิจัยได้ทำการเปรียบเทียบการแปรผันของจำนวนชุดดีเอ็นเอในคนไทยและในกลุ่มประชากร HAPMAP3 โดยใช้โปรแกรม plink มาทำ Hierarchical Clustering Analysis (HCA) โดยใช้ความถี่ของการแปรผันของจำนวนชุดดีเอ็นเอที่ผ่านการคัดเลือกมาทำการจัดกลุ่มประชากรคนไทยและกลุ่มประชากร HAPMAP3 ซึ่งจากผลการทดลองพบว่าการแปรผันของจำนวนชุดดีเอ็นเอสามารถจัดกลุ่มร่วมกับกลุ่มของประชากรในกลุ่มเอเชียตะวันออก และจากข้อมูลนี้สามารถที่จะใช้เป็นแหล่งอ้างอิงในการแปรผลของจำนวนชุดดีเอ็นเอที่ยังไม่ทราบความสำคัญทางคลีนิกในคนไทยได้ในอนาคตต่อไป

80 หน้า

CONTENTS

CONTENTS (cont.)

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURE

LIST OF FIGURE (cont.)

LIST OF ABBREVIATIONS

aCGH	Comparative Genomic Hybridization Array
bp	Base pair
CNV	Copy Number Variation
CNVR	Copy Number Variation Region
DNA	Deoxyribonucleic acid
FISH	Fluorescence In Situ Hybridization
GWAS	Genome-Wide Association Studies
HCA	Hierarchical Clustering Analysis
Kbp	Kilobase pairs
LOH	Copy neutral loss of heterozygosity
Mbp	Megabase pairs
NGS	Next Generation Sequencing
SNP	Single Nucleotide Polymorphism

CHAPTER I
INTRODUCTION

Copy Number Variation (CNV) is known for the major sources of genetic variation in a human genome, however, it has been largely neglected until the past decade, as compared to other generatic variations such as Single Nucleotide Polymorphisms (SNPs). Previous studies have revealed that CNV is more common in a human genome than previously known, as at least 12 CNVs have been estimated in any individual on average (Iafrate et al. 2004; Sebat et al. 2004). CNVs account for 4 Mbp (1 in every 800 bp) of genetic difference between individuals, which is higher than SNPs in terms of fraction in the human genome, as SNPs normally account for approximately 2.5 Mbp (1 in every 1,200 bp). Therefore, the total genetic variation between any two individuals is significantly greater than previously thought (at least 0.2% of the genome). More than 0.12% of this is structural variations, and 0.08% is nucleotide variations (Sebat 2007). At present, many studies found CNVs in both normal and diseased individuals. Many diseases are found associated with CNVs, such as important genetic diseases including autoimmune, neuropsychiatric and cardiovascular diseases (Fanciulli, Petretto, & Aitman, 2010). CNVs can convey a phenotype by the following mechanisms: (a) gene dosage, (b) gene interruption, (c) gene fusion, (d) position effects and unmasking of recessive alleles, or functional polymorphism (Feuk, Carson, and Scherer 2006).

Generally, the techniques used to detect CNVs can be divided into three main categories: (1) PCR-based detection methods, (2) microarray-based detection methods, and (3) sequencing-based detection methods. These three methods differ in the precision, throughput, and resolution (Li and Olivier 2013). While microarray- and sequencing-based techniques are commonly used to detect CNVs at a genome-wide scale, PCR-based technique is commonly used as a method of choice to detect target well-characterized CNVs. Next generation sequencing (NGS) technology provides sensitive and accurate tools for detecting CNVs. However, NGS requires a high depth

of coverage for sufficient accuracy in calling the CNVs, so it is presently not yet a commonly used technique for detecting CNVs in a large number of samples due to high cost. On the other hand, SNP genotyping arrays are currently more attractive compared to the other two due to its several advantages. This technique provides both intensity and genotype of each SNP so it can be used for both SNP genotyping and CNV detection in the same experiment. It does not requires as much sample when compare to aCGH array. Moreover, SNP array allows parallel analysis of a large samples, saving both cost and time (Hughes et al. 2011).

At present, there are several databases that provide both genotype and phenotype information, which can be used for CNV interpretation. The DGV database (MacDonald et al. 2014) provides a valuable catalogue for interpreting control data, whereas the DECIPHER database (Firth et al. 2009)provides a useful catalogue of pathogenic variants and associated phenotype for interpreting pathogenic data. However, the data in these databases are mainly European, African and certain East Asian ethnic groups, but do not include Thai individuals. So far there is no comprehensive Thai samples that case be used as controls, and thus no reference for interpretation of CNVs in the Thai population. Moreover, recent studies in Korean, Chinese and European show population-specific CNVs in each population, so the interpretation of CNVs from these combined population should be done with caution. (Chen et al., 2011; Lou et al., 2011; Yim et al., 2010). For this reason, this project has been conceived to develop a reference Thai CNV database, which we hope to help both clinicians and scientists to interpret CNV which found in the Thai population.

We first compile the genetic variation information from multiple SNP genotyping array studies. The combined dataset contains 3,017 Thai individuals from multiple published genome-wide association studies (GWAS), of which majority of the subjects suffered from infectious diseases and no known genetic disease. Together with my collaborators, which I shall specify in more details in the main text, we propose to characterize CNVs in a Thai population, with an aim to establish a reference database of CNVs of normal Thai individuals. I then focus on the comparison of the occurrence patterns of CNVs among the Thais, as well as compare with the CNVs of different ethnic groups publicly available. In doing so we have identified both shared and unique CNVs in the Thai population, in order to establish a

population-specific references, which can help diagnosis of genetic diseases of the Thai population in the future.

CHAPTER II
OBJECTIVES

1) To characterize and determine the frequency of Copy Number Variations (CNVs) in the Thai population.

2) To make CNV genotypic information available for public access through a database (as part of a team, which I shall clearly describe). This database will contain the CNV frequencies, types of CNVs, and genomic locations.

3) To explore the CNV occurrence patterns in the Thai population, and perform a comparative analysis against the CNVs found in other populations.

CHAPTER III
LITERATURE REVIEWS

3.1 Human Genome and Genetic Variation

The human genome is the main storage of genetic information in human. The genetic information is encoded in the form of Deoxyribonucleic Acid (DNA), which is the blueprint for creating all the human cells. All the DNA sequences are stored within 23 pairs of chromosomes (22 autosomal chromosomes and a pair of ssex chromosome) within the nuclei in all the cells. A haploid human genome (containing only one copy of each of the chromosomes) comprises approximately 3 billions bp of DNA. The discovery of the double-helix DNA structure by James Watson and Francis Crick in 1953 (Watson & Crick, 1953) has lead to the crucial first step in molecular biology of gene and genome (NHGRI, 2012).

To obtain the complete sequence of the entire human genome was a big challenge in the genomic area after the structure of DNA was uncovered. The Human Genome Project was an international collaboration that was established to undertake this task (Major Changes in Our DNA Lead to Major Changes in Our Thinking, 2012). The project was designed to help scientist improve the understanding of the sequences that make up human DNA and their functions. After the Human Genome Project was completed in 2003, it provided researchers with the first reference human genome, as well as the location of genes within the genome (Venter et al., 2001).

Having the first complete human genome, there were still many other questions yet to be answered: Are there any difference in human genome sequences among individuals? Do these differences in genomic sequence lead to human disease? Can genetic variations make us more susceptible to certain diseases? (Major Changes in Our DNA Lead to Major Changes in Our Thinking, 2012). The answers to these questions require much deeper knowledge of genetic variation, beyond a single reference genome.

The early studies of human genomic variation that started soon after the completion of the Human Genome Project confirmed that the genome sequence of any two individuals is up to 99.9% identical (NHGRI 2012b;Venter et al. 2001) and that the 0.1% are genetic variations that can cause phenotypic differences between each individual in the population, such as physical chracteristics (for example, hair and eye colors), disease susceptibility and drug responses.

The genetic variations between individual genomes of different populations can take many forms, namely single nucleotide polymorphisms (SNPs); tandem repeats; insertions and deletions (indels); Copy Number Variations (CNVs), which change the number of times a copy of DNA segment occurs: chromosome rearrangements such as inversions and translocations and copy neutral loss of heterozygosity (LOH) (Ku, Loy, Salim, Pawitan, & Chia, 2010). One of the main purposes of studying the genetic variation between humans is to identify the polymorphism associated with various phenotypes and diseases.

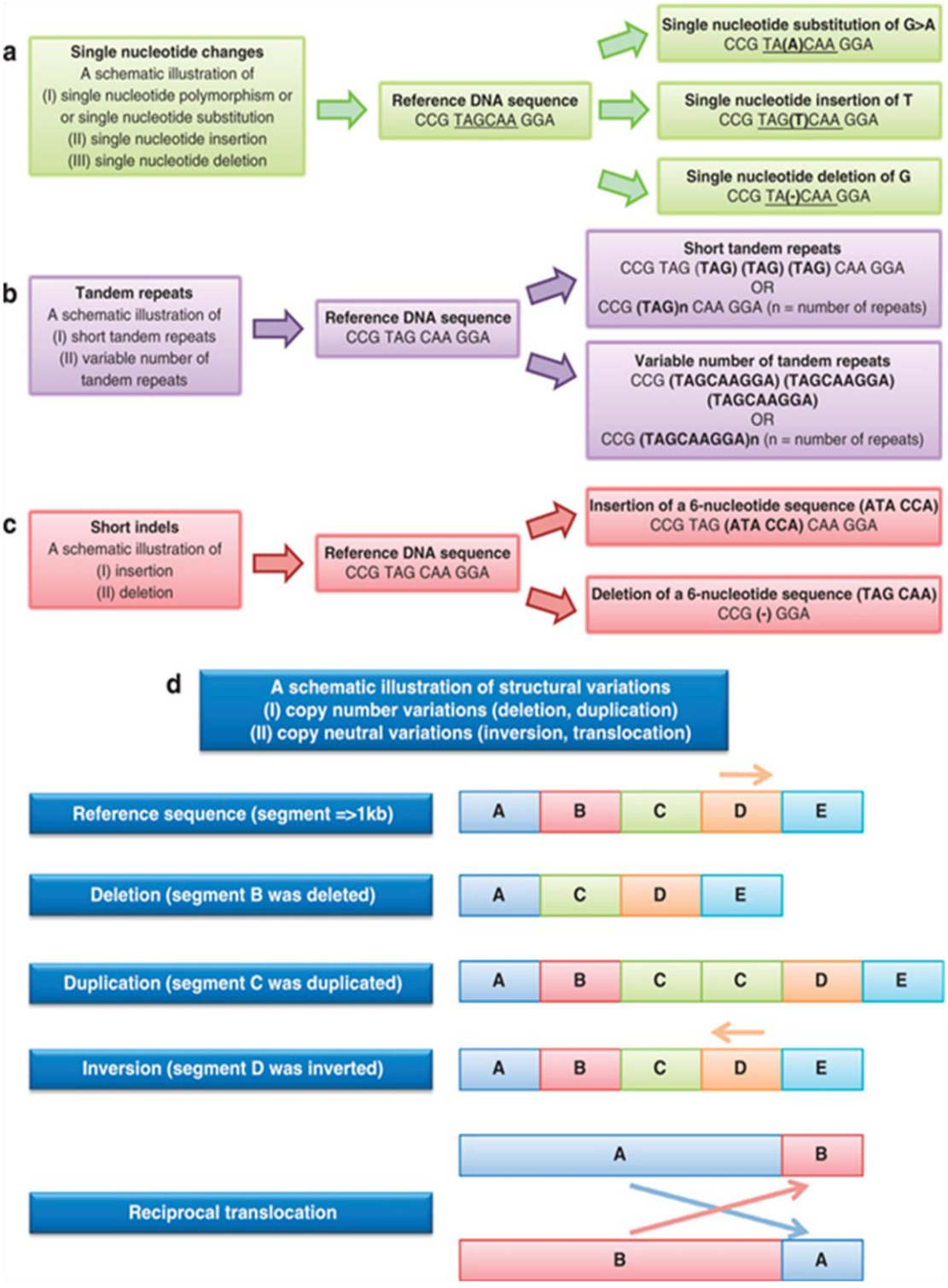


Fig3.1 A schematic illustration of (a) single nucleotide changes; (b) tandem repeats; (c) short indels; (d) structural variations from Ku *et al* (Ku et al. 2010).

Single Nucleotide Polymorphisms or SNPs are one of the most common genetic variations in human (Ku et al. 2010). If there are multiple possible different nucleotides at a particular position with the least common nucleotide occuring more than one percent in the population, it will be classified as a SNP. In the human genome, we can find on average one SNP in every one kilobase pair (Kbp) of DNA sequence. This means approximately three million SNPs are present in each individual's genome (Day, 2010). However, we do not fully understand the functions of all of the SNPs and some may or may not carry any important genetic information.

The International HapMap Project has been established with the aim to identify the common patterns of genetic variations in the human genome and to make this information freely available to the public. It helps the scientists to extract the informative SNPs from several million to roughly 500,000 tag SNPs to cover the entire genome. In fact, the SNPs on the same chromosome are frequently inherited in a block known as haplotypes. This means a few SNPs might be enough to represent all SNPs within that block. The specific SNPs that identify the haplotypes are called tag SNPs (NHGRI, 2012). This knowledge is very helpful to expedite the development of high-throughput genotyping SNP arrays, which is an important technique for genotyping millions of SNPs at the same time.

SNPs occur in a much larger number in the genome as compared to other types of genetic variations such as microsatellites, which used to be widely focused on the past, and can be rapidly assayed in a single experiment. For this reason, SNPs are widely used as a genetic marker in Genome Wide Association Study (GWAS). In a GWAS, comparisons between different case and control can help to determine which SNPs are associated with specific diseases. These SNPs can be used to help assess whether a patient who may be at risk for certain genetic diseases, for example, autoimmune diseases, including systemic lupus erythematosus, type 1 diabetes, rheumatoid arthritis, and other metabolic diseases (e.g. type 2 diabetes, obesity) (Visscher, Brown, McCarthy, & Yang, 2012). In addition to SNPs, in recent years, there are more and more studies focusing on another type of genetic variation, known as CNVs (Conrad, Andrews, Carter, Hurles, & Pritchard, 2006; Girirajan, Campbell, & Eichler, 2011; Osiriphun et al., 2009; Redon et al., 2006; Valsesia et al., 2012; Kai Wang et al., 2007a). These studies include the CNVs discovered in both normal

(Iafrate et al., 2004; Perry et al., 2008; Redon et al., 2006) and diseased individuals (Firth et al., 2009; Zhang, Gu, Hurles, & Lupski, 2009) individuals.

3.2 Characteristics of Copy Number Variations

3.2.1 What is CNV and its Categories?

Copy Number Variation (CNV) is defined as a gain or a loss of a segment of DNA with a size more than 1 Kbp. CNVs are present in different copy numbers as compared to the reference genomes. Because the term CNV does not directly indicate a clinical significance, deep understanding of CNVs and their clinical relevance will allow us to categories CNVs based on their functional or clinical significance. Generally speaking, CNVs can be categorized into three groups: pathogenic CNVs, benign CNVs, and CNVs of unknown clinical significance (CNVUS), as recommended of The American College of Medical Genetics guidelines for array based technique (Kearney, Thorland, Brown, Quintero-Rivera, & South, 2011). There is a general relationship between the size of CNVs and their clinical significance. The larger and contiguous CNVs have a higher chance to cover important genomic regions, which are subsequently more likely to modify expression levels of genes associated with disease (Lee, Iafrate, & Brothman, 2007). However, there are also cases where large CNVs are benign and small CNVs are clinically relevant so consideration of genomic content within and around the CNV intervals is crucial (Kearney et al. 2011).

Pathogenic CNVs are the easiest to identify because the deleted or duplicated CNV regions normally contain or overlap with a gene or a genomic region associated with a clinical disorder reported in the OMIM database (Lee et al., 2007). In contrast, benign CNVs are mostly determined when the deleted or duplicated regions contain no annotated gene or the CNVs overlapped with well-known characterized regions, such as salivary amylase gene. The CNVs commonly found in a population (more than one percent of the population) are also defines as benign CNVs. The last type of CNV is of unknown clinical significance (CNVUS). This type

represent the CNVs neither defined as pathogenic or benign, or do not contain any well-supported evidence. Some unknown CNVs may appear like pathogenic CNVs, such as there are well defined breakpoints and phenotypes. Other CNVUS may look like benign CNVs, for example, the CNVs reported only in a few cases in the databases but not common in population (Kearney et al. 2011). Fig.3.2 shows the major criteria that can be used to distinguish between pathogenic and benign CNVs (Lee et al., 2007).

	Major criteria	Characteristic of pathogenic CNVs	Characteristic of benign CNVs
1	a. CNV is inherited from a healthy parent	No	Yes
	b. CNV is inherited from an affected parent	Yes	No
2	a. CNV is similar to a CNV in a healthy relative	No	Yes
	b. CNV is similar to a CNV in an affected relative	Yes	No
3	a. CNV overlaps in genomic imbalance in a CNV database for healthy individuals (for example, Database of Genomic Variants)	No	Yes
	b. CNV overlaps in genomic imbalance in a CNV database for afftected individuals (for example, DECIPHER)	Yes	No
4	CNV contains morbid OMIM genes	Yes	No
5	a. CNV is a gene rich	Yes	No
	b. CNV is gene poor	No	Yes
	Minor criteria	Characteristic of pathogenic CNVs	Characteristic of benign CNVs
1	a. CNV is a deletion	Yes	No
	b. CNV is a homozygous deletion	Yes	No
2	a. CNV is duplication	No	Yes
	b. CNV is an amplication (gain of more than one copy)	Yes	No
3	CNV is >3 Mb in size	Yes	No
4	CNV is devoid of known regulatory elements	No	Yes

Fig 3.2 The influence factor for risk assessment of a CNV, adapted from Lee *et al*. ((Lee et al., 2007).

3.2.2. How many CNVs are there in a human genome?

In 2004, two studies have reported that CNVs are widespread in human genomes and represent a large proportion of all the genetic variation (Iafrate et al., 2004; Sebat et al., 2004). The first generation CNV map of human genome was created by Redon *et.al* in 2006. They used 270 individuals from four ancestral populations including those from Europe, Asia, and Africa. Two platforms were used in this study namely single-nucleotide polymorphism (SNP) genotyping arrays, and BAC clone-based comparative genomic hybridization. The results showed that there were on average 1,447 copy number variable regions (CNVRs) covering 360 Mbp (12% of the genome) in these populations. Many new CNVs have been since discovered from subsequent studies (Conrad et al. 2006, Redon et.al in 2006,Valsesia et al. 2012). One of the more recent studies has shown that roughly 15% of the human genome was affected by CNVs and an average of 12 CNVs exist in an individual compared with the reference genome (Li & Olivier, 2013). With the development of techniques such as SNP array, which can detect CNVs at a higher resolution, the amount of CNVs analyzed per genome has increased.

Database of Genomic Variant (DGV) is a public database that catalogues structural variations (SVs) including CNVs found in the genomes of normal individuals from global populations. The contents and data in DGV come from many studies and various platforms. Earlier DGV data were collected from low-resolution microarrays, which can result in high false negative and false positive rates. Currently, higher-resolution microarrays and next generation sequencing technologies have been used to detect CNVs in individual genomes, which significantly improves the accuracy of the DGV database. The number of published SV data that have been added to DGV has significantly increased as summarized in Fig 3.3. The current version of DGV consists of 55 published studies, comprising > 2.5 million entries identified in > 22,300 genomes (MacDonald, Ziman, Yuen, Feuk, & Scherer, 2014). DGV contains various types of genetic variations and most of these data are CNVs as shown in Table 3.1. The latest version of DGV identified 44% of variants from microarray studies, and the rest of the variants using sequencing studies (53%), and other targeted approaches including FISH/PCR and Optical Mapping (3%) (MacDonald et al., 2014).

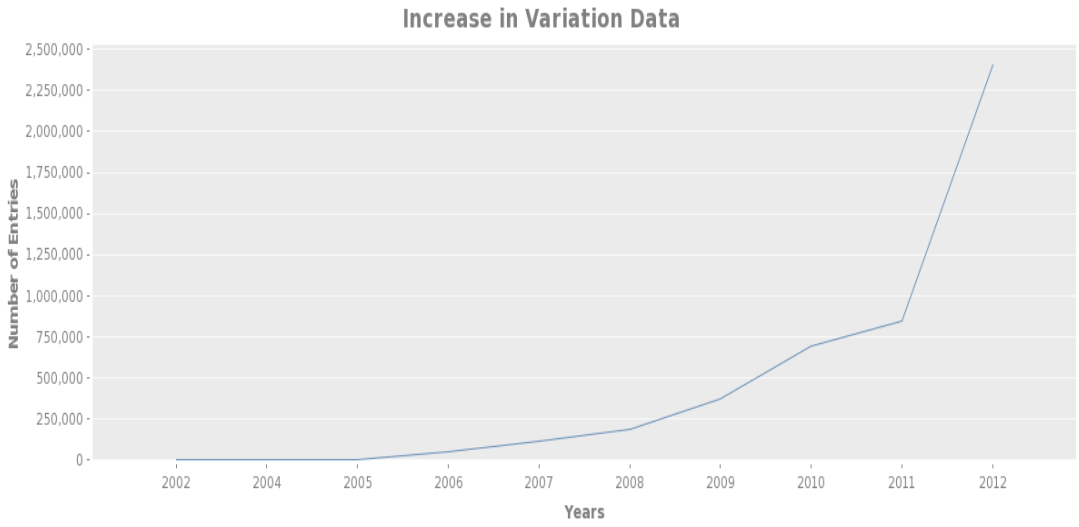


Fig 3.3 Graph showing the increase in published structural variation data that have been added to DGV database since its start in 2004; the numbers reflect the years of publications. The data were obtained from http://dgv.tcag.ca/dgv/app/statistics, last accessed in May 2014.

Table 3.1 The total number of studies and samples reported in the DGV database (the July 2013 update, mapped to the GRCh37 assembly). The data were obtained from http://dgv.tcag.ca/dgv/app/statistics, last accessed in May 2014.

Database content	Number of entries
Studies	55
Unique samples	14 316
Variant regions	202 431
Deletion	77 268
Duplication	668
Loss	64 185
Gain	24 891
Gain + loss	3850
Insertion	24 140
Inversion	1149
Complex	4090
Unknown	2189
Variant calls	2 393 718
CNV	2 391 408
Inversion	2310
Filtered variants	3 900 253

3.2.3 Size and chromosomal location of CNVs

The size distribution of CNVs in DGV is shown in Fig 3.4. The size of CNVs ranged from 50 bp to more than 1 Mbp. Most CNVs size ranged from 1-10 Kbp and the size distribution tended to follow a normal distribution. The difference in size of CNVs may also result from different detection methods. It is known that the detectable size of CNV is limited by a technique used to identify CNVs. In early days, Giemsa staining of chromosomes (karyotype) was the first method used to detect CNVs that are larger than 5 Mbp using regular light microscopy Fluorescence in situ hybridization (FISH) provides more precise detection than Giemsa staining but it still detects CNVs in a few hundred kbp range. Additionally, BAC array is a type of array-based method where the resolution depends on the insertion size of the BAC clone, which is approximately 200 Kbp (Vandeweyer & Kooy, 2013). Detecting smaller CNVs is difficult using these methods mentioned above. The resolution of the SNP arrays varies across the genome and the detection limit can go down as low as 10–40 Kbp (Carter, 2007). The density of probes also influences the CNV detection more probes able to detect more CNVs.

The locations of CNVs on the chromosomes tend to be equally distributed (Fig 3.5). CNVs can be found in all chromosomes. However, CNV enriched regions are in subtelomeric and pericentromeric regions of the chromosomes due to their high repetitive content

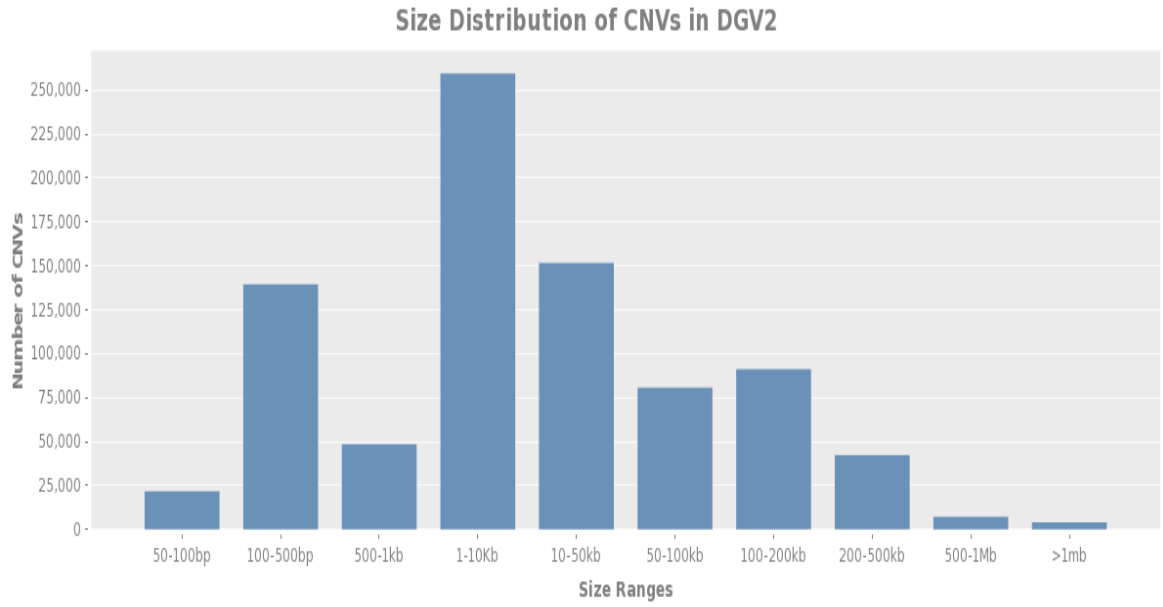


Fig 3.4 Size distribution of CNV in the DGV database (http://dgv.tcag.ca/dgv/app/statistics, last accessed in May 2014).

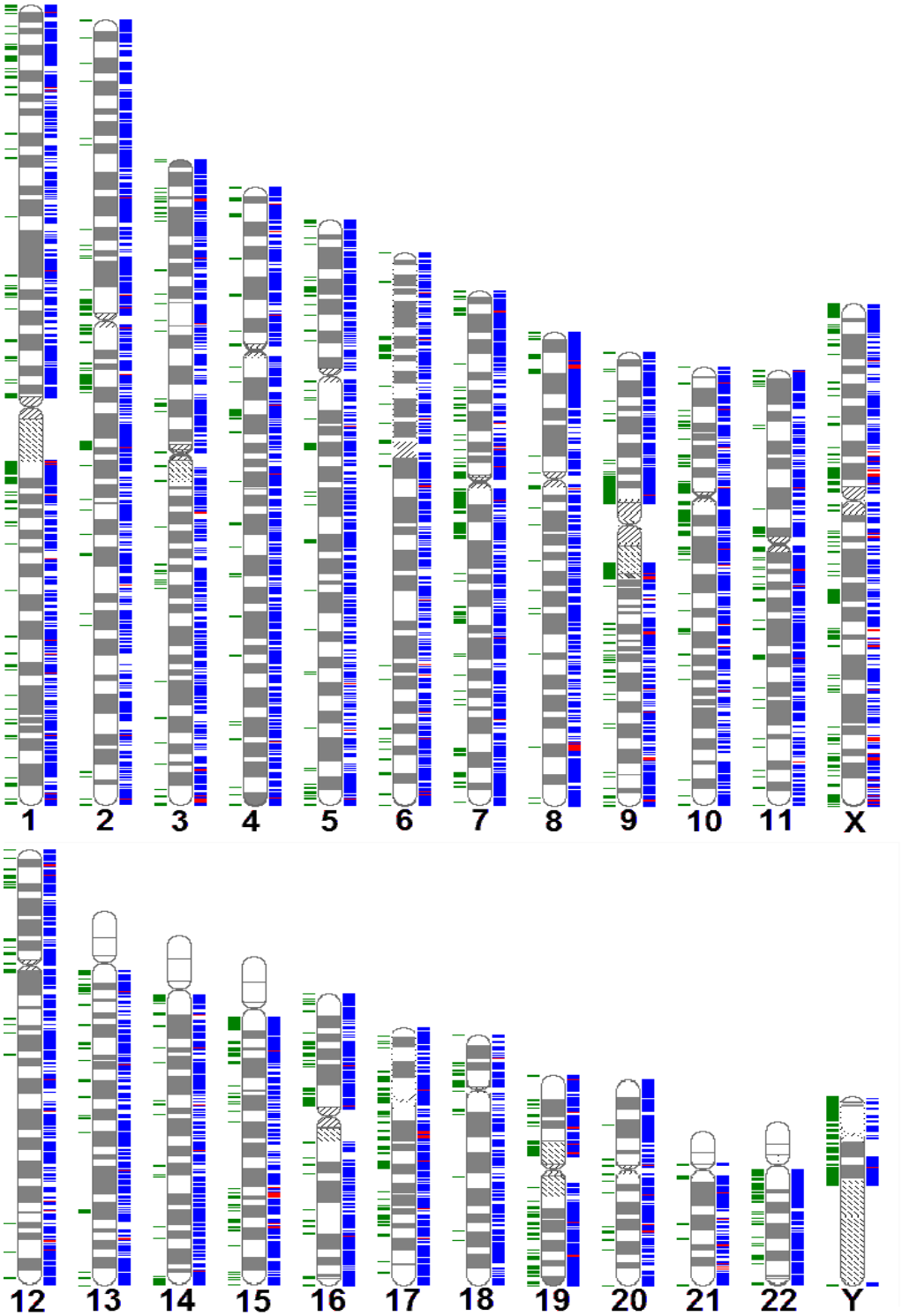


Fig 3.5 A Genome-wide view of CNVs in the human genome . Reported CNVs are represented as blue bars, inversion breakpoints were represented as red bars, segmental duplications were represented as green bars. The data were obtained from http://dgv.tcag.ca/dgv/app/statistics, last accessed in May 2014.

3.3 Method to detect CNVs

Copy Number Variation can be detected by a number of methods. In the past, Giemsa staining was one of the first methods used to detect copy number variant in chromosomes. However, the resolution of this technique is relatively low and it can detect only CNV larger than 5 Mbp using a light microscope. Although Giemsa staining has a low resolution, this technique was once the primary method used to detect chromosome rearrangements. Fluorescence in situ hybridization (FISH) is another method used to detect copy number variants under a microscope. FISH provide higher resolution and more precision than the Giemsa staining. This technique is based on the hybridization of fluorescently labeled probes to specific complementary sequences in the genome. At present, there are three main techniques that are more widely used to detect CNVs: (1) PCR based detection method, (2) microarray based detection method and (3) sequencing detection method. These three methods differ in the precision, throughput, and resolution. Due to their popularity, here I will focus on genome-wide CNV discovery platforms, that is, DNA microarrays (CGH and SNP) and sequencing detection methods (Vandeweyer & Kooy, 2013).

3.3.1 CGH array

The principle of CGH arrays (aCGH) is based on hybridization between probes and samples DNA. Both samples and reference probes are labeled with fluorescent dyes of different colors (red or green), then the samples are hybridized with probes attached to the arrays (Fig 3.6). The signal intensity ratios between the samples and references along the length of each chromosome is used to assess the copy number within each region. In the CGH array, probes can be genomic fragments cloned in a variety of vectors such as bacterial artificial chromosomes (BACS), cDNA, or long synthetic oligonucleotides. The genomic location of each probe is designed specifically to a region of interest. After the hybridization step, the signal ratio between the samples and references can be used to indicate copy number loss or copy number gain in the target regions. Although the resolution of CGH array is not as high as the one obtained from recent SNP arrays, the signals obtained from a few CGH probes tend to be more reliable than those obtained from few adjacent SNPs, whereas

the allele specific copy number cannot be inferred directly from CGH array as in the SNP array (Valsesia et.al 2013) .

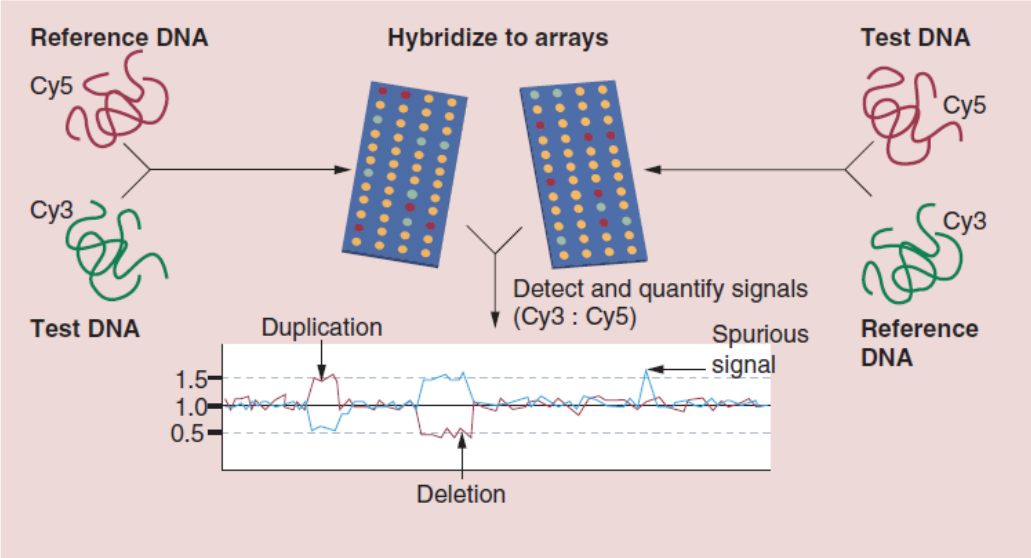


Fig 3.6 Principle of comparative hybridization array, adapted from Geert Vandeweyer and GV, Kooy RF (Vandeweyer & Kooy, 2013).

3.3.2 SNP array

The second type of microarray that can also be used to assess CNVs in individuals is the single nucleotide polymorphism genotyping array (SNP array). It provides a number of advantages over the CGH array. Firstly, although SNP arrays are not primarily designed for CNV analysis, they can be used to detect both SNP genotyping and CNV at the same time using appropriate algorithms (described below). Secondly, the SNP array is a cost effective and allows researchers to scale-up the number of samples tested on a small budget. Moreover, it uses a smaller amount of sample per experiment when compared to other techniques such as CGH array (Winchester, Yau, & Ragoussis, 2009).

Another difference between SNP array and aCGH is the number of samples that are hybridized to the array. Array CGH method uses different labels (red/green) on reference and test sample hybridize to probe on the same array. On the contrary, only one sample is hybridized to the SNP array and then the data is compared *in silico* to a reference dataset (Vandeweyer & Kooy, 2013). At present,

There are two major companies manufacturing SNP arrays. They use different approaches to determine the SNP genotype. Illumina (Illumina Inc., CA, USA) (Fig 3.7) uses a different nucleotide labeled color in single base extension reaction to determine a genotype of each SNP probe. Then, these samples will be hybridize to probes on the array, which contains complementary DNA sequences. The differently labeled nucleotides will bind to the sample specifically at SNP position that allows the researchers to determine the genotype of each SNP.

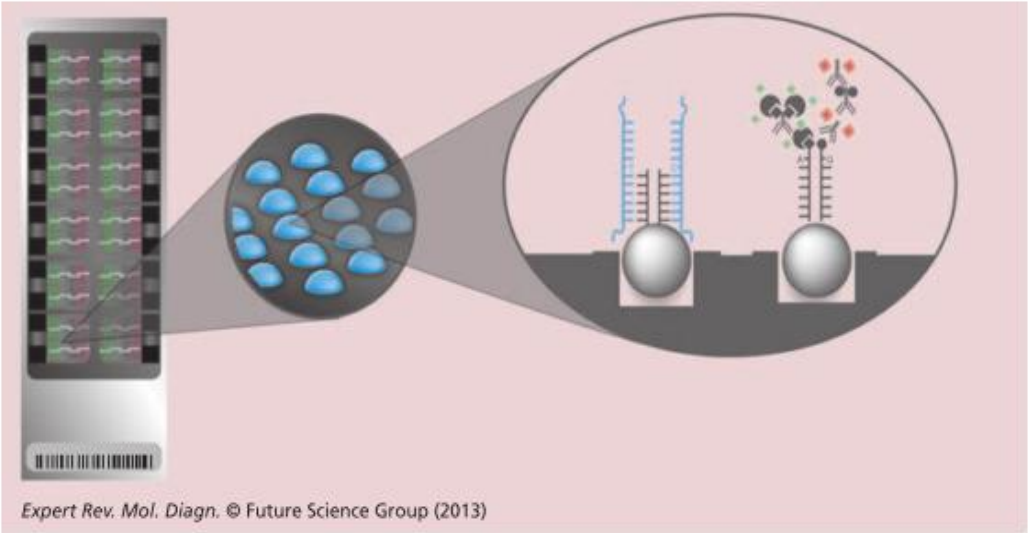


Fig 3.7 Principle of the Illumina SNP array. Each of the chip contains several thousands to millions different beads. Each bead contains several thousand replicates of one type of probe targeting to adjacent SNP location. A labeled nucleotide is inserted at the SNP position, adapted from Geert Vandeweyer and GV, Kooy RF (Vandeweyer & Kooy, 2013).

The Affymetrix method (Affymetrix Inc., CA, USA) uses a different approach to determine SNP genotype. It uses different probes to target each alleles in single-color hybridization. After DNA samples were fragmented and labeled with fluorochrom. it is hybridized to the array where it binds specifically to a perfect match probe (Nowak, Hofmann, & Koeffler, 2009).

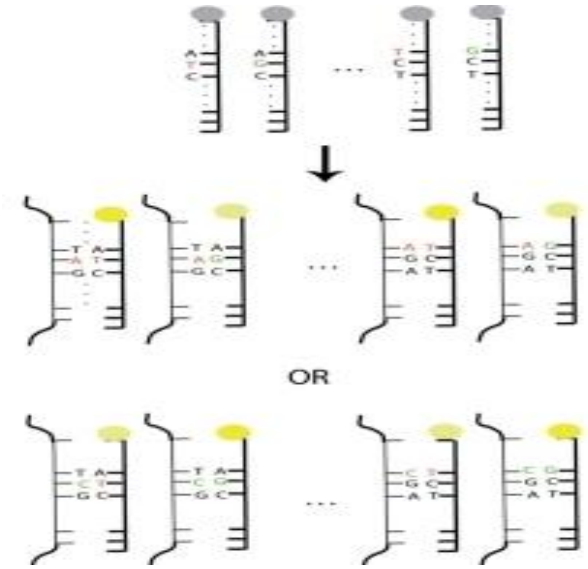


Fig 3.8 Principle of the Affymetrix SNP array. Each SNP probe has 25 bp for both alleles, and the location of the SNP locus varies from probe to probe. The DNA binds to both probes regardless of the allele carried, but binds more efficiently when it is complementary to all the 25 bases (bright yellow) rather than mismatching the SNP site (dimmer yellow) adapted from Thomas LaFramboise et al (LaFramboise, 2009).

Both Illumina and Affymetrix SNP determination methods provide two different types of information. The first type of information is the intensity of all SNPs. The SNP intensity is then normalized to two, representing two copies of each allele on the chromosome. The homozygous deletion results in a zero and heterozygous deletion results in value of one. Duplications or amplifications result in values of three or higher (Nowak et al., 2009). The second type of information is the intensity ratio between A and B alleles. This information provides a measure for each SNP genotype (AA, AB and BB genotype). When combined with copy number data, this information allows the detection of loss of heterozygosity (LOH). Because there are more than a million probes in an array, it provides a big advantage for the investigation of SNP association coupled to CNV analysis in GWAS studies and it can

be used to infer the copy number of any given genomic region containing the SNP markers (Vandeweyer & Kooy, 2013).

3.3.3 Next Generation Sequencing

Since Next Generation Sequencing (NGS) has been applied to genome sequencing, it has also become sensitive and relatively accurate approaches for accessing CNVs. Currently, many platforms are available such as Illumina HiSeq/MiSeq, Life Technologies Ion Torrent/Ion Proton, Life Technologies SOLiD, and various platforms in development. Although different platforms use different in the biochemistry for sequencing, but the workflows are conceptually similar. NGS starts with library construction by random fragmentation of DNA and ligate each fragment to common adapter sequence. The second step, the clusters of amplicon are generated by Polymerase Chain Reaction (PCR), derived from either single location on a planar substrate or to the surface of the bead. The final step is sequencing and imaging which consists of alternative enzyme bichemistry (Fusté, 2012). Most of the platforms use sequencing by synthesis, which can be either polymerase (Mitra, Shendure, Olejnik, & Church, 2003) or ligase (Mardis, 2008). Data information is received from the imaging of each cycle.

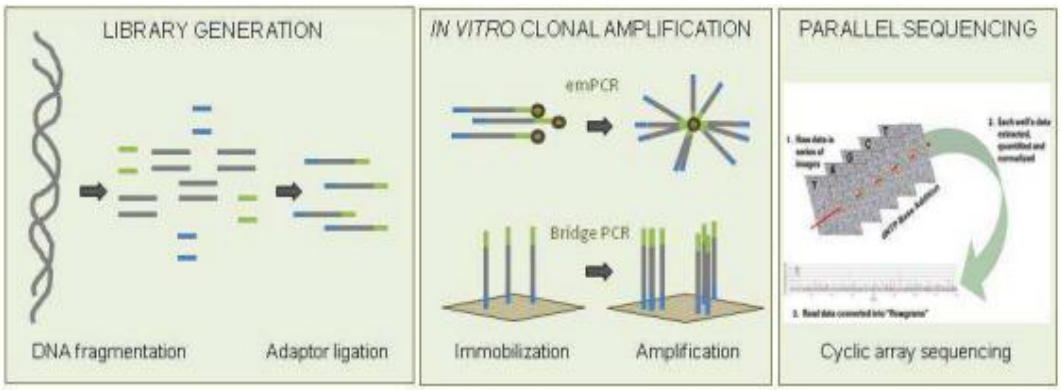


Fig 3.9 Work flow of Next Generation of DNA Sequencing (Fusté, 2012).

Algorithms for detection of structural variation from NGS data fall into three main categories: discordant paired-end reads, split reads, or depth of coverage

(Abel & Duncavage, 2013). Whole genome sequencing allow the scientist to detect the structural variation at single nucleotide resolution. However, NGS has some drawbacks when used to detect CNV. Firstly, NGS provides short length sequence so many read can not uniquely map to the genome. A high depth of coverage is needed for more precision. Cost is also one important factor to select detection method. If the requirement is to screen a large sample size, NGS may not be a cost effective technique. Second, alignment in repeat-rich regions is particular problematic reads aligning to multiple locations. Third, if the insertion is larger than insert size the short reads sequenced from this region will not align (Xi et al., 2010). Therefore more advanced and sophisticated platforms and algorithms for detecting SVs from sequence data are required.

3.4 Algorithms for CNV detection using SNP array

Because of the large number of probes and their great coverage of the genome, many algorithms have been created to detect CNVs from the SNP array data. High-density SNP array data provide both alleic ratio and high quality intensity information for each locus to assess copy number variation. Each SNP marker in the array is represented as an A or B allele, and a and b are designed as normalized intensity data for A and B alleles, respectively (Li & Olivier, 2013). For CNV estimation, a and b are transformed into R and Θ. R is a measure of total signal intensity from two alleles and R = a + b. Θ is a measure of the relative allelic intensity ratio and equals arctan(a/b)/(Θ/2). Both R and Θ values are then transformed to two parameters, which are the most important for CNV detection from SNP arrays. The first one is logR ratio (LLR) which is defined as log2(Robserved/Rexpected), where Rexpected is measured form the reference sample. The B allele frequency (BAF) is the normalized measure of the relative signal intensity ratio of B and A alleles (Li & Olivier, 2013). It depends on the Θ values for three canonical genotype clusters (AA, AB, BB) generated from the reference samples (such as the Hapmap samples) (K. Wang & Bucan, 2010). The majority of the detection algorithms generally have two categories: Hidden Markov model (HMM) and Circular Binary Segmentation (CBS). The Hidden Markov Model (HMM) is a statistical process that creates a model

dependent structure between the copy numbers at nearby SNPs. The HMM assumes that only copy number state of the most preceding SNP can affect the true copy number state at each SNP.

PennCNV is one of the HMM-based algorithms that we decided to use in this study. In the HMM state-transition matrix, it was created from both state-specific and distance-dependent transition probabilities. By this method, it can create a more realistic model for state transition between different copy number states (Li & Olivier, 2013). PennCNV uses six-state definition instead of gain, loss and normal when compared with other algorithms. Moreover, the PennCNV also uses several sources of useful data, population allele frequency for each SNP, the distance between adjacent SNPs and family information, to improve the accuracy in calling CNV. Smaller size CNVs (with a median size of 12 Kbp) can be efficiently identified by PennCNV, smaller than previous studies (Kai Wang et al., 2007a).

Circular binary segmentation algorithm (CBS) is a modified version of binary segmentation, which was firstly developed for the aCGH analysis. Detection in a small change segment that occurs between large segments is the limitation of the binary segmentation algorithm. It detects only a single change point at a time (Olshen, Venkatraman, Lucito, & Wigler, 2004). CBS is an adjustment of the change-point strategy, allowing for tertiary splits by connecting the two chromosomal ends. CBS starts by dividing the chromosome into neighboring regions of equal copy number by modeling discrete copy number gains and losses, then assesses the significance of the proposed splits using a permutation reference distribution (Karimpour-Fard, Dumas, Phang, Sikela, & Hunter, 2010).

CNVworkshop is an example of a CBS-based algorithm. The genotyping data provides useful parameters including allelic ratios used to detect copy number of the segment. However, each genotyping platform also has its unique properties so CNVworkshop applies three major steps to detect CNVs from every genotype platform including segmentation, calculation of genotype-specific statistics, and CNV determination. Both Log R ratio (LRR) and B-allele frequency (BAF) are used to determine CNV (Gai et al., 2010). The above or below zero of LRR value indicate possible duplication and deletion at the segmentation step. Additional statistics have been used to make a quality CNV determination: standard deviation of LRRs by

sample and also by chromosome, and mean LRR for each chromosome and each segment (Gai et al., 2010). Similarly to BAF, three separate statistics have been calculated for each segment: percentage of SNPs that have BAF between 0.4-0.6, b2.sd for normal diploid segment as well as homozygous AA alleles (0), homozygous BB alleles (1), and heterozygous AB alleles (0.5) and b3.sd for monoallelic duplication: AAA (0), BBB (1), ABB (0.67), and AAB (0.33) (Gai et al., 2010)). In addition to this, CNVworkshop provides a threshold for several different genotyping platform: Illumina 550 K Illumina 610-Quad, 660-Quad, and Affymetrix 6.0 arrays to make more accuracy CNV calling. Annotation using CNVworkshop is automatic for gene content, known disease loci, and gene-based literature references. This information can be easily obtaind, sorted, filtered and visualized in a web-based presentation.

3.5 Common mechanisms of copy number variation formation

Changes in copy numbers of DNA fragments alter the chromosome structure. In general, there are two main mechanisms that can change the structure of a chromosome: Homologous Recombination (HR) and non-homologous recombination. HR requires a sequence identity around 50 bp in E.coli and up to 300 bp in mammals and human. In contrast to HR mechanism, non-homologous recombination requires only small micro homology or no homology at all (Hastings, Lupski, Rosenberg, & Ira, 2009).

HR is a more accurate repair mechanism that uses another identical sequence to repair the damaged DNA sequence. Non-allelic Homologous Recombination (NAHR) is an example of the HR mechanism that is caused by Low Copy Repeat (LCR). Normally, LCR is spread out throught the entire human genome and used for repairing damage sequences in the same positions on the chromosome. LCR contains more than 97% identity in the DNA sequence and a distance of approximately 10 Mbp from another LCR (Hastings et al., 2009). Because of the identity in their DNA sequences of LCRs, they can lead to misalignment of a chromosome or chromatid at non-allelic positions on the same chromosome. The result of this mechanism is unequal crossing over and thus leads to deletion or

duplication of a segment flanked by a direct orientation of the LCR (Stankiewicz & Lupski, 2010). If an NAHR falls in between inverted LCRs, it will cause inversion of the genomic segment as shown in Fig 3.10.

Non-Homologous end joining (NHEJ) is an example of the non-homologous recombination mechanism. Normally, NHEJ is a mechanism responsible for repairing double stranded DNA breaks by joining the two ends together. However, in some cases, an NHEJ may lead to a small deletion or insertion, this frequently occurs in mitochondria DNA. NHEJ is different from the NAHR mechanism because it does not require LCR to mediate the recombination mechanism (Hastings et al. 2009, Stankiewicz and Lupski 2010). The major differences between NAHR and NAHJ mechanisms are the location and size. When a deletion or duplication event occurs at a particular region, sizes and positions vary in each individual. Because NHEJ does not require LCR, it does not occur at the same region in every person, which is different from the NAHR mechanism.

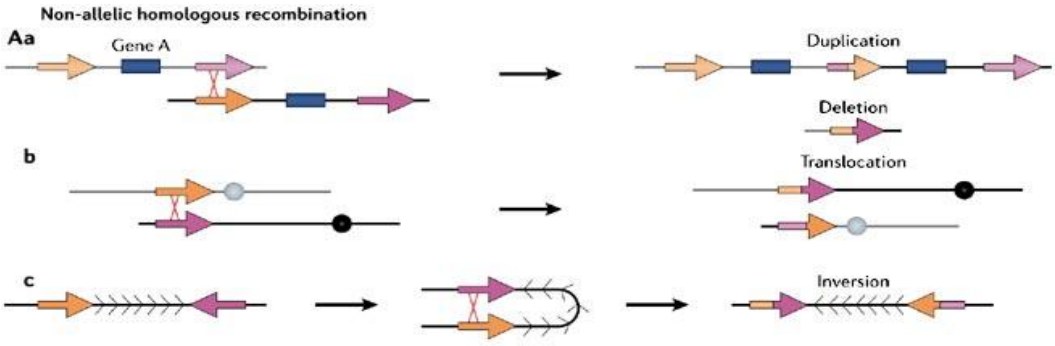


Fig 3.10 Duplications and deletions can occur as a result of NAHR between adjacent duplicated sequences (Aa). Translocations could result from an exchange between segmental duplication on non-homologous chromosomes (Ab). Inversions can occur as a consequence of recombination between inverted intrachromosomal duplications (Ac) adapted from Jeffrey A. Bailey & Evan E. Eichler (Bailey & Eichler, 2006).

3.6 Studies of CNV in human phenotypes and diseases

CNVs that play a role in human disease were first recognized in the 1980s, in the case of alpha-thalassaemia (Fanciulli et al., 2010). The CNV at the alpha globin locus has been shown to be a causative genetic trait of this disease. As present, we know that CNVs are widespread in the human genomes and many have been reported to be associated with a number of genetic diseases, including autoimmune and infectious disease susceptibility, as well as neuropsychiatric and cardiovascular diseases (Fanciulli, Petretto, and Aitman 2010, Almal and Padh 2012).

3.6.1 CNV in infectious and autoimmune diseases

CCL3L1, a cheomokine gene involved in the immunoregulatory and inflammatory process, is located on the long arm of chromosome 17(17q12). This gene encodes a HIV-1-suppressive chemokine, which is a major co-receptor for CCR5. In Caucasian populations an increase in the copy number of the CCL3L1 gene can reduce the susceptibility to HIV (Almal & Padh, 2012). Moreover, it is also interesting that high copy numbers of the CCL3L1 gene has been associated with rheumatoid arthritis, as demonstrated in two separate Caucasian populations (New Zealand and UK populations) (Almal & Padh, 2012). In addition, a similar observation has been observed in the beta-defensin gene. High and low copy numbers of this gene have influences on association to disease in different ethnic groups. The chemokine proteins, whose functions are linked to the innate and adaptive immune responses, were encoded by these genes. In three independent cohorts, deletion of the beta-defensin 2 (HBD-2) gene, the leading cause of colonic Crohn's disease, and high HBD-2 copy number was associated with psoriasis in the Dutch and German populations (Fanciulli et al., 2010).

3.6.2 CNV in neurological disease

CNVs also have been reported to be associated with neurological and psychiatric diseases, including autism, schizophrenia, epilepsy and Parkinson syndromes (Fanciulli, Petretto, and Aitman 2010, Almal and Padh 2012). Autism is a high prevalence neurological disorder. It can be found in between 15-60 of 10,000 children (Fanciulli, Petretto, and Aitman 2010). The symptom characteristics include limitation or absent verbal communication, a lack of reciprocal social interaction or responsiveness, and restricted, stereotypical, and ritualized patterns of interests and behavior (Zhang et al., 2009). There are several loci associated with autism susceptibility such as duplication of 15q11-13 (AUTS4; MIM 608636) and deletion of 16p11.2 (AUTS14; MIM611913) (Almal and Padh 2012). The recurrent micro-deletion or micro-duplication on the 16p11.2 accounts for approximately 1% of all the cases. There is also a de novo CNV that was deleted at the locus that overlaps genes such as DIA1 SHANK3 and neurexin1, with known function related to neurological activity (Fanciulli, Petretto, and Aitman 2010, Almal and Padh 2012).

Schizophrenia is also another high prevalence neurological disorder associated with CNVs. Schizophrenia is a debilitating illness with both neurological and other psychiatric features (Zhang et al., 2009). Three deletion loci on position 1q21.1, 15q11.2, and 15q13.3 were found to be associated with schizophrenia and psychosis. These three loci have also been confirmed by the International Schizophrenia Consortium in case and control studies (Almal and Padh 2012). The genes interrupted by CNVs include MYT1L, CTNND2 and ASTN2.

3.6.3 CNV in cardiovascular disease

The potential role of CNVs in cardiovascular disease came from the report of an overlap region between a CNV and loci which association to cardiovascular traits. For example, the deletion in the LDLR gene has been found in affect patients with familial hypercholesterolemia (FH) disorder. In addition, the LPA gene on chromosome 6 encodes an atherogenic apolipoprote which is the primary determinant of the plasma lipoprotein and a risk factor for atherosclerosis (Almal and Padh 2012).

At present, we now also have a database that collects and catalogs the CNV data linked to multiple diseases: DECIPHER (Database of Chromosomal Imbalances using Ensembl Resources). This database will help the identification and interpretation of pathogenic genetic variations in patients with rare diseases and genetic disorders (Firth et al., 2009). Up until now, approximately 70 symptoms from over 21,000 patients have been categorized in DECIPHER. These diseases included Charcot-Marie-Tooth syndrome (CMT), Angelman syndrome, Cri du Chat Syndrome and many other syndromes in the database.

3.7 Influence of CNV on phenotype

As described in the introduciton, CNVs can be regarded as benign, or have a subtle influence on phenotype (for example, they can modify drug response), or causes diseases (for example duplication or deletion in associated disease genes). Generally, CNVs can convey a phenotype by the four possible mechanisms: (a) gene dosage, (b) gene interruption, (c) gene fusion, (d) positional effects and unmasking of recessive alleles, or functional polymorphisms (Zhang et al., 2009).

Changes in a copy number of a gene by duplication or deletion can cause disease. PMP22 is an example of the effect of CNV on a dosage sensitive gene. The gene is located within the 1.4-Mbp CMT1A region of the chromosome 17p12 (Zhang et al., 2009) Charcot-Marie-Tooth disease type 1A (CMT1A) is one of the most common inherited peripheral neuropathies in human, characterized by a decrease in nerve conduction velocities (NCVs). The duplication of PMP22 results in decreased NCV, and leads to the CMT1A disease (Walsh et al., 2008), and PMP22 deletions, leading to Hereditary Neuropathy, with liability to Pressure Palsies (HNPP).

CNVs can also disrupt gene functions by deletion or duplication, and cause functional loss or modification. One example is a study in schizophrenia, where the researchers collected data from approximately 150 patients and 268 ancestrial records. This genome-wide study revealed a number of genes in pathways relating to brain development, which were disrupted by the CNVs. Interestingly, the disrupted genes have been found in cases only, but not in controls. For example, ERBB4 (receptor tyrosine-protein kinase erbB-4), encoding a type I transmembrane tyrosine kinase receptor for neuregulins, is disrupted by a 399-Kbp deletion in a patient (LIFTON et al., 1992).

The fusion of different genes or their regulatory sequence can be caused by genomic rearrangements. It can also generate a gain of functional mutation. This mechanism can be observed among cancer associated patients with specific somatic chromosomal translocations and can also be found in other diseases (Zhang et al., 2009). An autosomal dominant disorder, Glucocorticoid-remediable aldosteronism or GRA is characterized by hypertension with variable hyperaldosteronism. NAHR on chromosome 8q results in fusing the 5' regulatory region of 11 beta-hydroxylase to the coding sequences of aldosterone synthases. This particular mutation can account for all the physiological abnormalities of GRA in animals as well as human (Velagaleti et al., 2005). In addition to this, CNVs can have an effect on expression or regulation of a nearby gene outside the CNV region. Velagaleti et al. reported that two translocations that have breakpoints mapped to approximately 900 Kbp upstream and 1.3 Mbp downstream of SOX9 could cause compomelic dysplasia disease (Velagaleti et al., 2005). Hemizygous deletion at one locus may uncover expression of a recessive allele or functional polymorphism. For example, in a Sotos syndrome patient, the activity of plasma coagulation factor 12 (FXII) is dependent on the remaining FXII allele (Kurotaki et al., 2005).

CHAPTER IV
MATERIALS AND METHODS

4.1 Materials

4.1.1 Data

SNP genotyping data derived of the Thai populations (Chantarangsu et al. 2011; Jongjaroenprasert et al. 2012; Mahasirimongkol et al. 2012) CNV data of the HAPMAP3 population (Altshuler et al., 2010)

4.1.2 Programs

CNV Workshop (Gai et al., 2010) PennCNV (Kai Wang et al. 2007) PLINK (Purcell et al., 2007)

4.2 Methods

4.2.1 Ethical approval and consent

Ethical approval to use the patients in the Thai populations was obtained from the research Ethical Committee of Ramathibodi Hospital, Mahidol University, the Local research Ethical Committee of Department of Medical Science, Ministry of Public Health, Local research Ethical Committee of Center for genomic medicine and RIKEN.

The patients agreed to the consent forms for all parts of the study with the research medical doctors or research nurses or research coordinators present to answer the questions. In this study, the patients were not required to participate for all experiments and they could retire from the study any time without any consequence

.

4.2.2 Study Samples

To undertake a large-scale whole genome study, 3,427 Thai individuals who were enrolled in the population-based cohort were genotyped with Illumina microarray of different platforms. The samples, selected in this study, were obtained from the previous genome-wide association studies (GWAS) in the Thai individuals (Table 5.1). No known genetic diseases were reported in any of these subjects. Previously analyzed and published CNVs from 11 different ethnicities from the HAPMAP3 project were downloaded from "http://HapMap.ncbi.nlm.nih.gov/downloads /cnv_data/hm3_cnv_submission.txt" and the number of samples in each population are summarized in Table 5.5.

4.2.3 CNV genotyping platform

In this study, we used three different types of genotyping platforms to detect CNVs; Illumina HumanHap550 Genotyping BeadChip, Illumina HumanHap 610 Quad BeadChip, and Illumina OmiExpress BeadChip. Each genotype platform contains specific numbers of probes used to detect both SNP genotype and CNV. Illumina HumanHap 550 BeadChip provides more than 550,000 tagSNP markers for detection CNVs whereas Illumina HumanHap610 Quad BeadChip, comprising a space between each probe equal to 4.7 Kbp, and contains 620,901 tagSNP markers. Consequently, Human OmiExpress showed the highest probe number of 730,525 tagSNP markers.

4.2.4 Sample Quality Control (QC)

The first sample quality control was conducted based on the SNP genotypes. We excluded samples that contained the SD of log-R ratio > 0.3, and the SNP call rate of < 98%, or with self-reported/genotype-derived sex inconsistency. In the HAPMAP3 population, we also applied the same filtering criteria as the Thai population in order to exclude low quality CNV data. CNVs with less than 30 Kbp per SNP density, less than 5 SNPs (>5 Kbp) for deletion CNVs, and less than 10 SNPs (>10 Kbp) for duplication CNVs and CNVs with more than 50% overlap with centromeric and telomeric regions were ignored.

4.2.5 CNV Calling

In this study, two different calling algorithms were used to annotate CNVs. The first one is CNVworkshop, which is based on the Circular Binary Segmentation (CBS) algorithm, explained in a previous section, and the second one is PennCNV, which is based on Hidden Markov Model (HMM) trained from know CNV data. The signal intensity from SNP genotyping arrays were used to call CNVs in our combined Thai population by these two algorithms. Using PennCNV, a number of parameters were applied to ensure the most accurate model. Firstly, signal intensity data of both A and B alleles were normalized and transformed into Log R Ratio (LRR) and B Allele Frequency (BAF) by using the GenomeStudio software (Illumina, San Diego, CA, USA), then the four possible states of CNVs: homozygous deletion, heterozygous deletion, normal copy number, and copy duplication were predicted using these four parameters (LRR, BAF, the chromosome position of SNP probe, and population frequency of B allele file (pfb) for the Thai population).

The second algorithm (CNV Workshop) was also used to identify the CNVs. Conceptually, the segments of chromosome that display a change in signal intensity, being different from their neighbors, were identified. Both mean LRR and the distribution of BAF were used to predict the copy number of each segment. The default parameter settings were used to call CNVs in both programs. The default paremeters of PennCNV are as follows : minsnp =3, medianadjust, sdadjust, and bafadjust = 1 in order to reduce false positve rate for low-quality samples empirically. For CNV Workshop, the default parameters were shown in Appendix A..

4.2.6 CNV discovery in the Thai population.

CNVs with more than 60% overlapping in length from CNVworkshop and PennCNV were selected for further downstream analyses. To reduce false positive rates, we excluded the CNVs with less than 30 Kbp per SNP density, less than 5 SNPs (>5 Kbp) for deletion CNVs, and less than 10 SNPs (>10 Kbp) for duplication CNVs. The CNVs on the Y chromosome and CNVs with more than 50% overlapping with centromeric and telomeric regions were also excluded. We also excluded individuals with more than 100 CNVs, as these were unusually high likely from an error of genotyping array.

4.2.7 CNV distribution among population

The CNV comparison between the Thai and Hapmap populations were done using the PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/cnv.shtml) software to perform an association study. Frequency of CNVs in the pairwise comparison between the Thai and each of the 11 HAPMAP3 populations were computed. Map location probes were created from unique start and end CNV locations in each pairwise comparison. The empirical statistical significance level was calculated using 5,000 permutations. The map locations with statistically significantly different frequencies were selected with p-value < 0.01. The overlapping genes with these map locations were chosen and the frequencies were calculated. The hierarchical clustering analysis was performed using frequencies of CNV overlapping with these genes in each population. These CNV frequencies were transformed to have a mean of 0 and a variance of 1. The heatmap package in R version 3.1 was used to create hierarchical clustering using the Euclidean distance with the Ward clustering method

4.2.8 Copy number variable region

Copy Number Variable Region (CNVR) is a widely used term that we employed in this study to represente a discrete region of overlapping CNVs. The CNVR was created after combining data from Thai and HAPMAP3 populations. The GenomicRanges package in R was used to collapse overlapping regions to discreet regions. The gene list, based on hg 18, from the PLINK software resource page (http://pngu.mgh.harvard.edu/~purcell/plink/dist/glist-hg18) was used to identified an overlapping CNVR. The frequency of CNVR in each population was calculated from the number of people that have a CNV in CNVR region, divided by the total number of people in each population. The degree of match between the Thais and HAPMAP3 was calculated by creating the CNVRS for the Thais and HAPMAP3 separately and then comparing the common CNVRs between two populations. Only CNVR containing more than one individual was included in this analysis.

4.2.9 Principle Component Analysis of Thai CNV

Principle Component Analysis (PCA) is a statistical method used for reducing a dimension of a large dataset to fewer principal components (PCs) that can

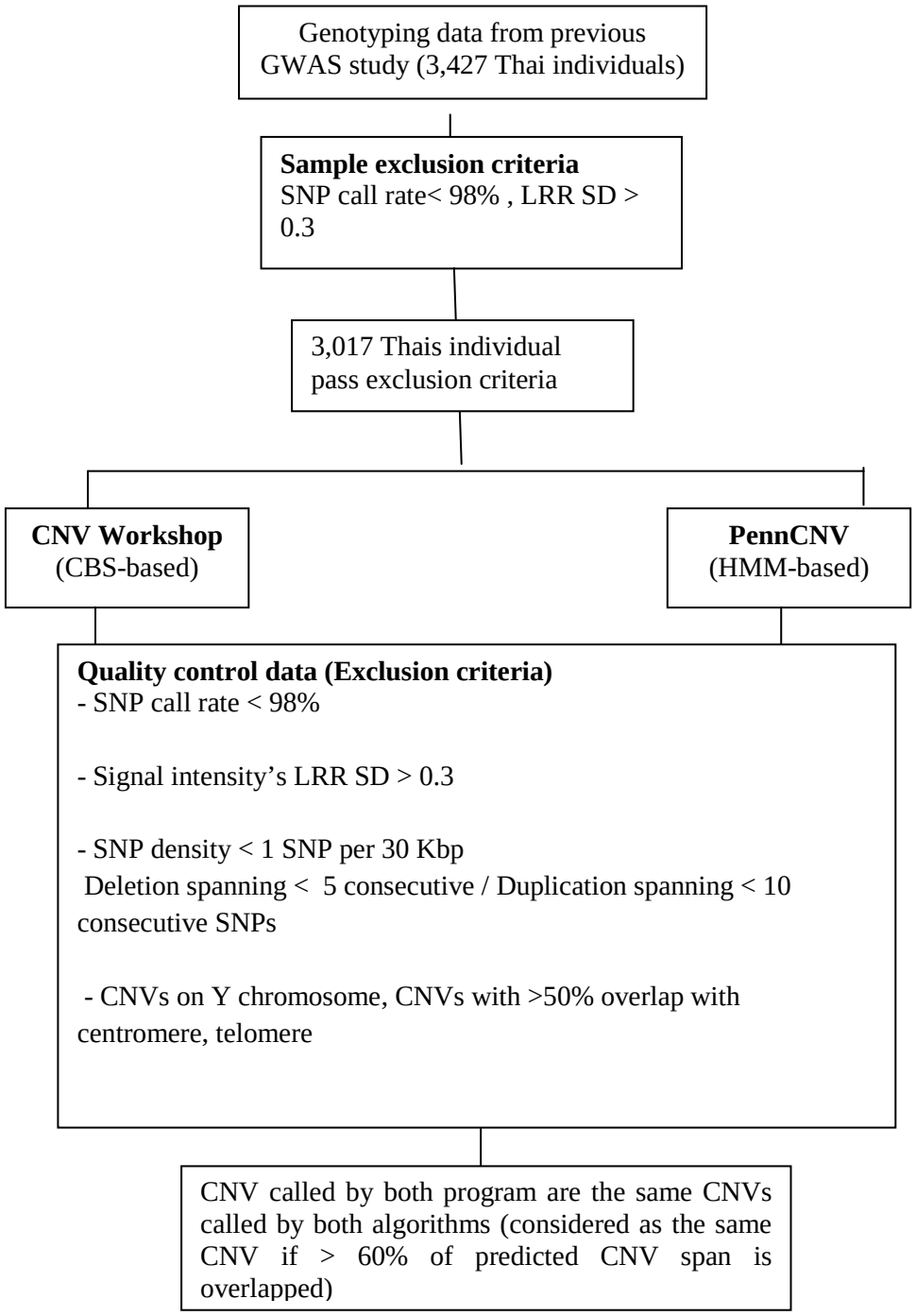
explain the main patterns of the data. In this study, we used PCA to explore the relationship between copy state and phenotypic group. The state of CNVs in 1,014 CNVRs were assigned as follows: zero and one for deletion, two for normal copy number, three and four for duplication and five for multiple state. In another PCA plot, we assigned zero if they do not contain a CNV and one as containing CNV in each CNVR. The table of CNV state in each CNVR was created in order to compute eigenvalue and eigen vector by applying "prcomp" in the R statistical package (R Development Core Team, 2013).

4.2.10 Cluster Analysis of Thai CNV

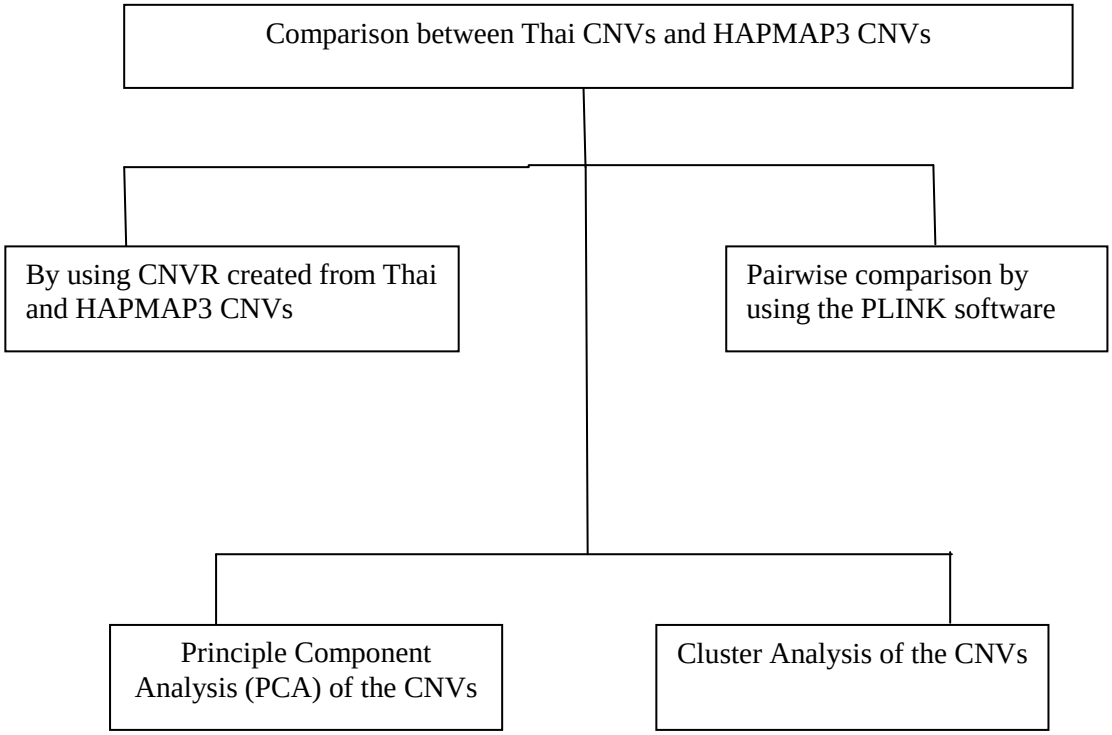
We sampled ten random individuals of each disease group to perform clustering analysis. The group of disease, types of SNP genotyping arrays, and the CNV calling software used were the three factors that we used to explore the distribution of the Thai CNV data through these parameters. Cluster analysis was then performed and the result was used to construct a dendrogram based on the similarity matrix data using the Ward's method from the R statistical package.

Experimental strategies

Part I CNV discovery in Thai population.



Part II CNV comparison between Thai and HAPMAP3 population



CHAPTER V
RESULTS

5.1 Characteristic of Thai CNV

5.1.1 Basic Characteristic of Thai CNV

Here, I summarize the combined Thai samples from multiple GWAS studies in Table 5.1. The total number of the Thai samples is 3,427. To filter low quality data, the following criteria have been used to exclude low quality samples: samples with the SD of log-R ratio > 0.3, SNP call rate of < 98%, or with self-reported/genotype-derived sex inconsistency (as described in Suktitipat, Naktang, et al. 2014). After this filtering process, the number of all samples decreased from 3,427 to 3,017 individuals. The percentages of samples excluded are also shown in Table 5.1.

Table 5.1 GWAS studies containing genomic data from 3,017 Thai individuals in the final CNV dataset, after excluding low quality data.

Reference	Type of SNP array	Number of subjects	Total	Excluded (%)
Jongjaroenprasert et al, 2012	Illumina Human610-quad	289	330	12.424
Mahasirimongkol et al, 2012	Illumina Human610-quad	463	484	4.339
Wattanapokayakit et al (unpublished data)	Illumina HumanOmniexpress-12	517	685	24.536
Chantarangsu et al, 2011	Illumina HumanHap550-Duo v3	56	165	66.061
Chantarangsu et al, 2011	Illumina Human610-quad	167	210	20.476
Mahasirimongkol et al, 2012	Illumina Human610-quad	856	868	1.382
Nuinoon et al, 2010	Illumina Human610-quad	669	685	2.336
	total	3017	3427	11.964

PennCNV and CNVworkshop, HMM- and CBS-based algorithms respecitively, have been used to call the CNVs. As part of the team (including Mr. Wuttichai Mhuantong, Miss Thitima Tularak and Miss Paramita Artiwet, as appearing in Suktitipat et al, PLoS ONE, 2014), we have used both programs to call CNVs from this combined Thai dataset. The called CNVs with more than 60% intersecting regions between the two programs are counted as the same CNVs. The two programs show approximately 70% common CNVs (29,436 CNVs), as shown in Figure 5.1.

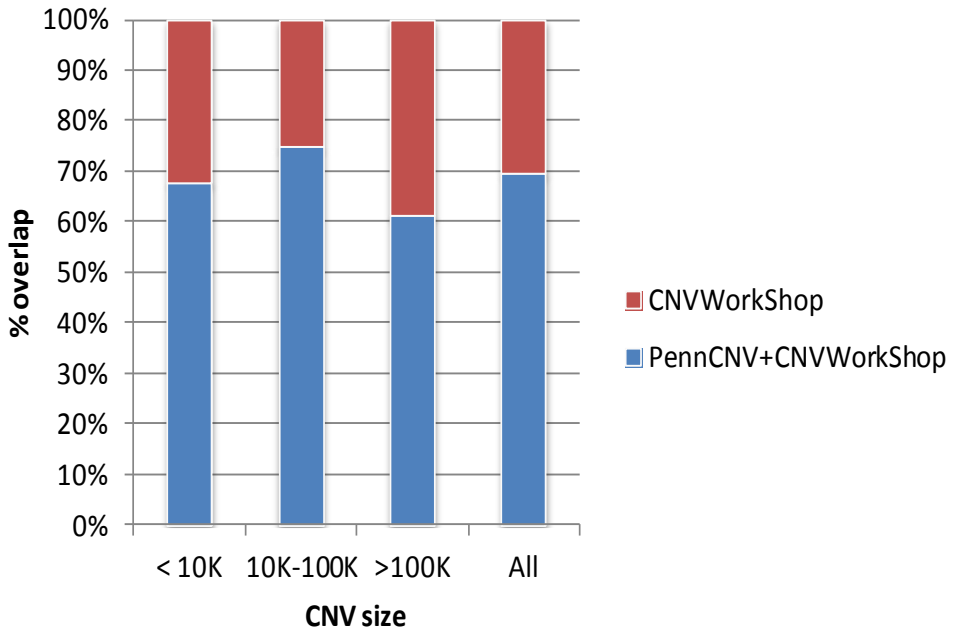


Fig 5.1 Percentages of common CNVs identified by CNVworkshop and PennCNV.

To obtain reliable and consistent CNV results, we further excluded regions with a low number of markers (<1 SNPs per 30 Kbp), small deletions (< 5 consecutive SNPs) or small duplications (<10 consecutive SNPs). The CNVs overlapping with centromeric and telomeric regions were also excluded due to a high false positive CNV prediction rate (Kai Wang et al., 2007c). After this second filtering process, there were 23,458 CNVs left for further comparison and statistical analyses.

In the next step, we confirmed that the CNVs obtained from both programs can indeed be used to represent a normal Thai population. To answer this question, we used the clinical-related CNV information from the DECIPHER database. Among the

23,458 CNVs identified in the Thai population, 1,505 CNVs (6.42%) overlap with the 70 known genetic syndromes from the DECIPHER database (https://decipher.sanger.ac.uk/syndromes) (data not shown). Characteristics of the Thai CNVs are shown in Table 5.2. The median size of Thai CNVs was 59.80 Kbp (ranged from 5kb to 427.5 Kbp). We found approximately 7.77 CNVs per individuals (ranging from 1 to 23). The number of deletion CNV was higher than duplication CNV (79 % and 21 % respectively). The median size of CNVs was 40,811 bp for deletion and 122,757 bp for duplication CNVs.

To represent unique CNV regions, we used a CNVR defined as the union region of any amount of overlapping CNVs found in more than one individual. Deletion CNVRs were found more frequently than duplication. The CNVR possessed a median size of approximately 95.06 Kbp (ranging from 5.18 Kbp to 4275.08 Kbp). We found 7.35 CNVR, accounting for 8.72% of genome coverage, per individual genome.

Table 5.2 Thai CNVs and their CNVR characteristics.

	CNV	CNVR
Total count	23,458	1014
CN-gain count	4,879	165
CNV-loss count	18,579	538
Multiplex		311
Average number per genome	7.77	7.35
Median size (range) (Kbp)	59.80 (5.0-4275.08)	95.06 (5.18-4275.08)
Median size of CN-gains	122.76 (100.45-4275.08)	137.34 (14.67-1491.4)
Median size of CN-losses	40.81 (5.0-3893.87)	37.5 (5.18-2144.0)
Genome coverage		261.77Mb (8.72%)

5.1.2 The percentage of Thai CNVs overlapping with known genes

Because changes in the copy number of genes can cause diseases, the number of gene-overlapping CNVs was identified. We used a gene list, downloaded from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/downloads.html) in the hg 18 version. In total, the number of overlapping CNVs with a gene is slightly

different from the number of CNV: no gene overlap accounts for 47.75% (11202/23458) and 52.25% (12256/23458) overlap a gene, respectively. Figure 5.2 illustrates this basic characteristic of the annotated Thai CNVs. Note that the CNV calling and preliminary analyses were performed as part of a team as explained above, whereas the rest of statistical analyses described below were performed by myself.

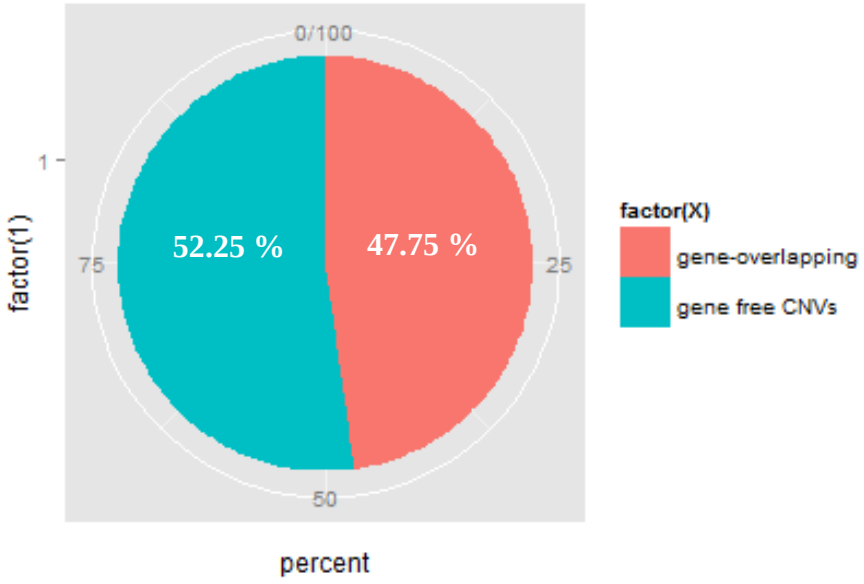


Fig 5.2 The percentage of Thai CNV, which overlap or without overlap with known genes.

5.1.3 Principle Component Analysis of Thai CNV

Principle Component Analysis (PCA) was used to explore whether there is a correlation between each copy number state and disease group (where the samples were taken from Table 5.3). The proportion of the variance in each major component tells us how much of the variance in the data set was accounted for by the different principal components. The first component, PC1, accounted for 14.12% of the variance while the second component accounted for 8.66%. The last component, PC16, accounts for approximately 1%. As a result, PC1 and PC2 were selected to plot

the PCA graph, as they captured the highest proportion of variance in the Thai CNV data

The states of CNVs in each CNVR were used to calculate PCA and the results of PCA analysis were shown in Figure 5.3. Different colors represent phenotypes (infectious disease) groups where this particular data set was taken from. The result contained six upper and six lower groups, while each group included a mixture of phenotypes. Such a pattern, which can neither be explained by the algorithms or parameters nor ignored, used to call the CNVs or other technical procedures performed during the analyses.

As a result, we were not able to confidently comment on this peculiar PCA pattern. It is possible, however, that the fact that the first two PCAs explain only 22% of all the variances might not be sufficient.

In addition to the CNV state, we performed another PCA using presence (1) and absence (0) patterns of each CNV in each sample. The result is also shown in Figure 5.3 on the right. On the bottom of Figure 5.3 we show another PCA, where color labeles each SNP genotyping platform to observe whether the type of SNP array can affect the CNVcalling results..

Unsurprisingly, the PCA plot included only two groups instead of three groups so that means the type of SNP genotyping array did not influence CNV calling. From all of the results, although we were not able to explain the two clusters explicitly, it was clear that samples from each disease group are distributed throughout both clusters. Importantly, this suggests that different disease groups included in our CNV calling did not affect how the calling algorithms performed. Alternatively, one could interpret that there are no underlying genetic differences in terms of CNVs, between patients with these infectious diseases.

Table 5.3 The proportion of variance with each component were captured in Thai CNV data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.187	0.930	0.735	0.577	0.521	0.494	0.480	0.461
Proportion of Variance	0.141	0.087	0.054	0.033	0.027	0.024	0.023	0.021
Cumulative Proportion	0.141	0.228	0.282	0.315	0.343	0.367	0.390	0.411
	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
Standard deviation	0.451	0.435	0.424	0.415	0.390	0.383	0.354	0.342
Proportion of Variance	0.020	0.019	0.018	0.017	0.015	0.015	0.013	0.012
Cumulative Proportion	0.432	0.451	0.469	0.486	0.501	0.516	0.528	0.540

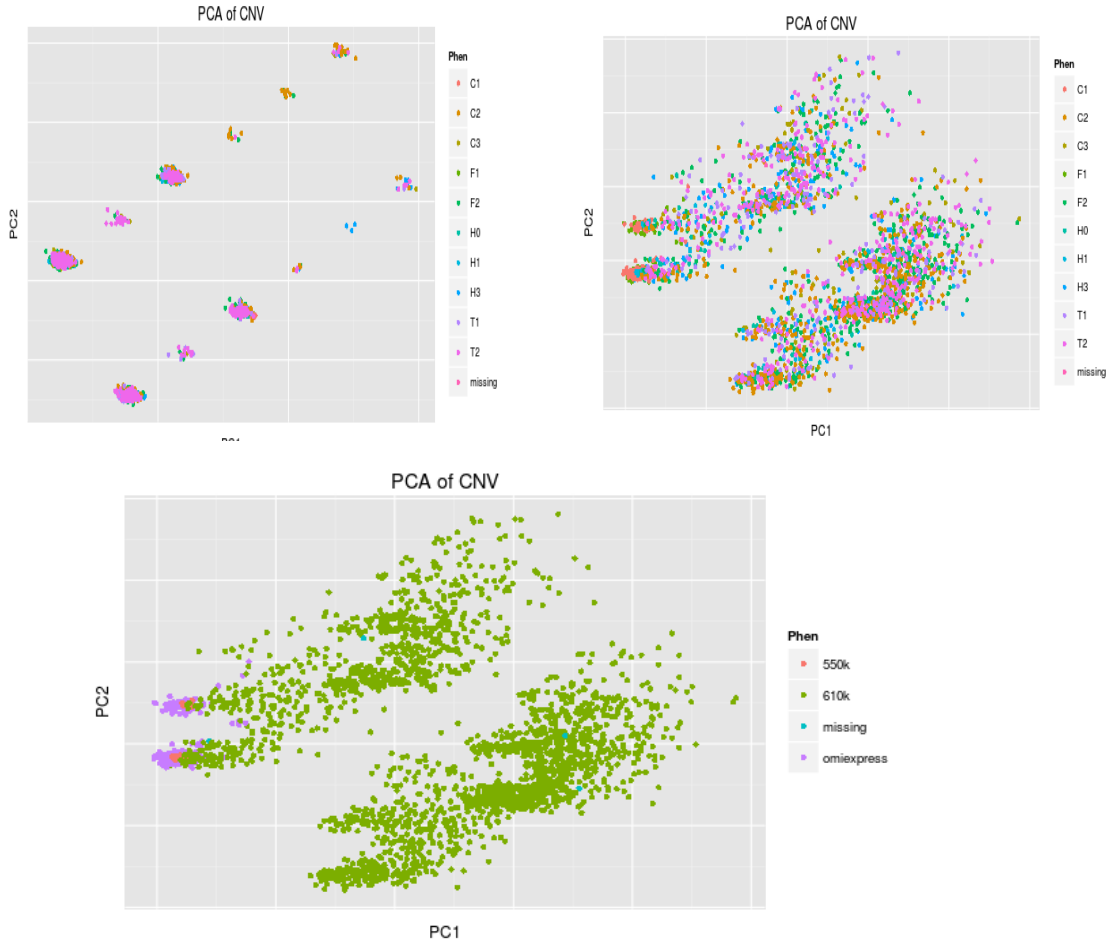


Fig 5.3 PCA plots of the Thai CNVs by copy states (upper left), presence and absence CNVs (upper right) and type of genotyping array (bottom) were used to explore a correlation between each factors and diseases.

5.1.4 Cluster Analysis of Thai CNVs

A dendrogram is another correlation analysis method we employed to group the samples according to how similar they are. In this study, we generated a dendrogram to explore whether the Thai CNV calls can be self-categorized according to the different disease groups, types of SNP genotyping arrays, and CNV calling algorithms. First, ten samples from each disease group were randomly selected and a cluster analysis using R statistics package was performed based on the states of CNVs (0,1,3,4) across the 1,014 CNVRs (Fig 5.4 upper panel). The similarities were then constructed using the Ward's method. According to the structure of this dendrogram, the Thai CNVs could not be grouped based on the distinct diseases affecting the subjects, nor the different types of SNP array used (Fig 5.4, middle panel), or the specific algorithms using to identify them (Fig 5.4, lower panel). Therefore, this confirmed that none of these parameters showed any effect on the distribution of our Thai CNV data. This result is also in line with our previous PCA described above (Fig 5.3).

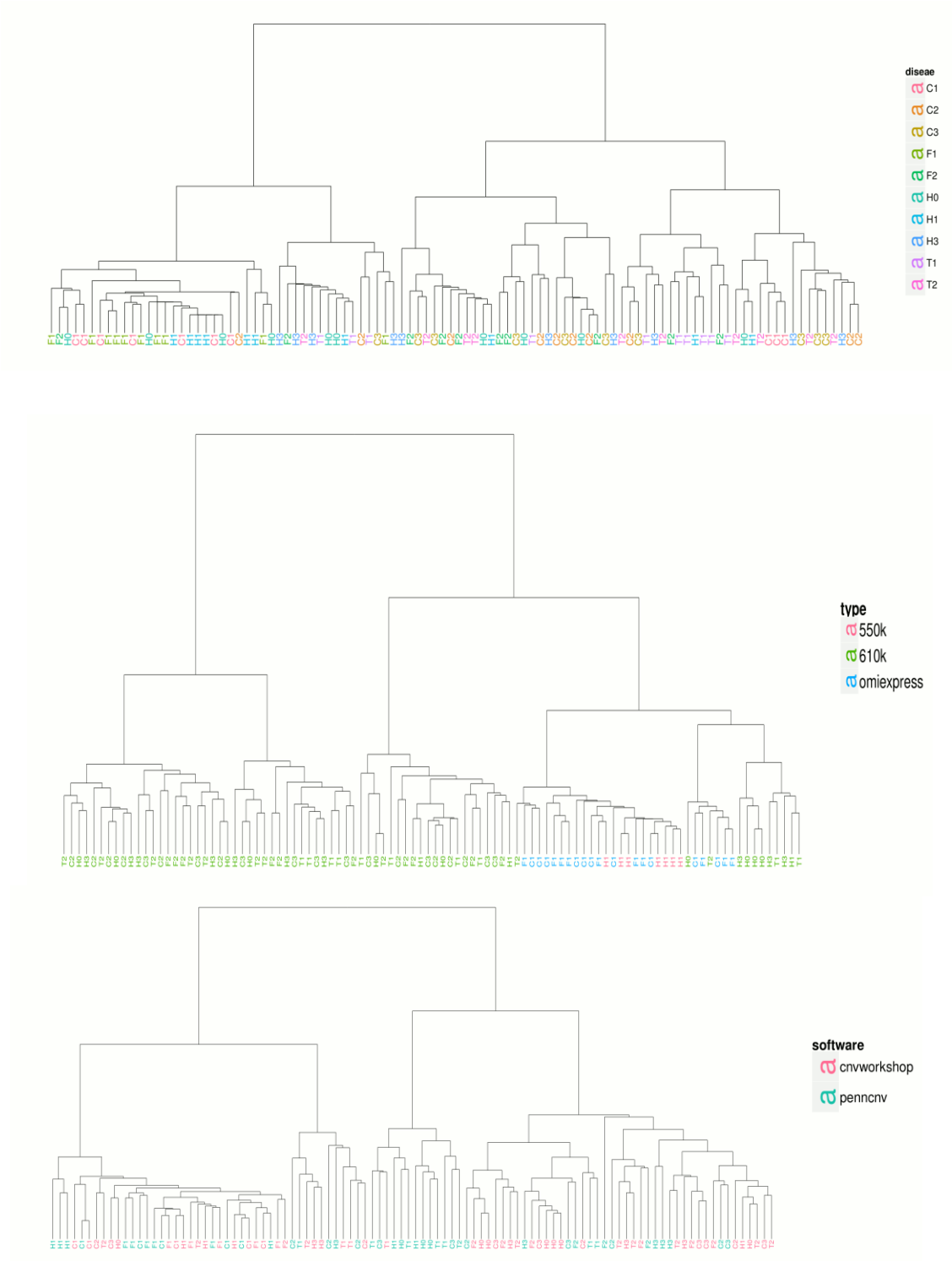


Fig 5.4 Dendrogram based on Ward's method depicting the relationship of CNV's state among group of disease (upper), type of array (middle) and CNV calling software (bottom).

5.2 Characteristics of Chromosome X in the Thai CNVs

5.2.1 Basic Characteristics of Chromosome X CNV

Table 5.4 Thai CNVs and their CNVR characteristics in chromosome X.

	CNV	CNVR
Total count	7,079	32
CN-gain count	2,382	0
CNV-loss count	4,697	21
Multiplex		11
Average number per genome	4.79	4.78
Median size (range) (Kbp)	97.91(0.7-5757.1)	616.80(32.10-25161.24)
Median size of CN-gains	146.63(4.91-5757.1)	0
Median size of CN-losses	73.6(0.7-4017.52)	87.8(32.10-537.87)

The CNV characteristics of chromosome X in the Thais were identified in the same way as autosomal chromosomes. However, only PennCNV contains functions to call CNVs on the X chromosome. After the same filtering steps, the number of deletion CNV was higher than the number of gain CNV 4,697 and 2,382 CNV, respectively. We found 4.79 CNVs in chromosome X per individual and the median size of CNV gain was larger than CNV loss. After CNVR conversion, there were more deletion CNVR than duplication CNVR. When compare to autosomal Thai CNVs, the number of CNVs per individual was higher than chromosome X same as in genome coverage.

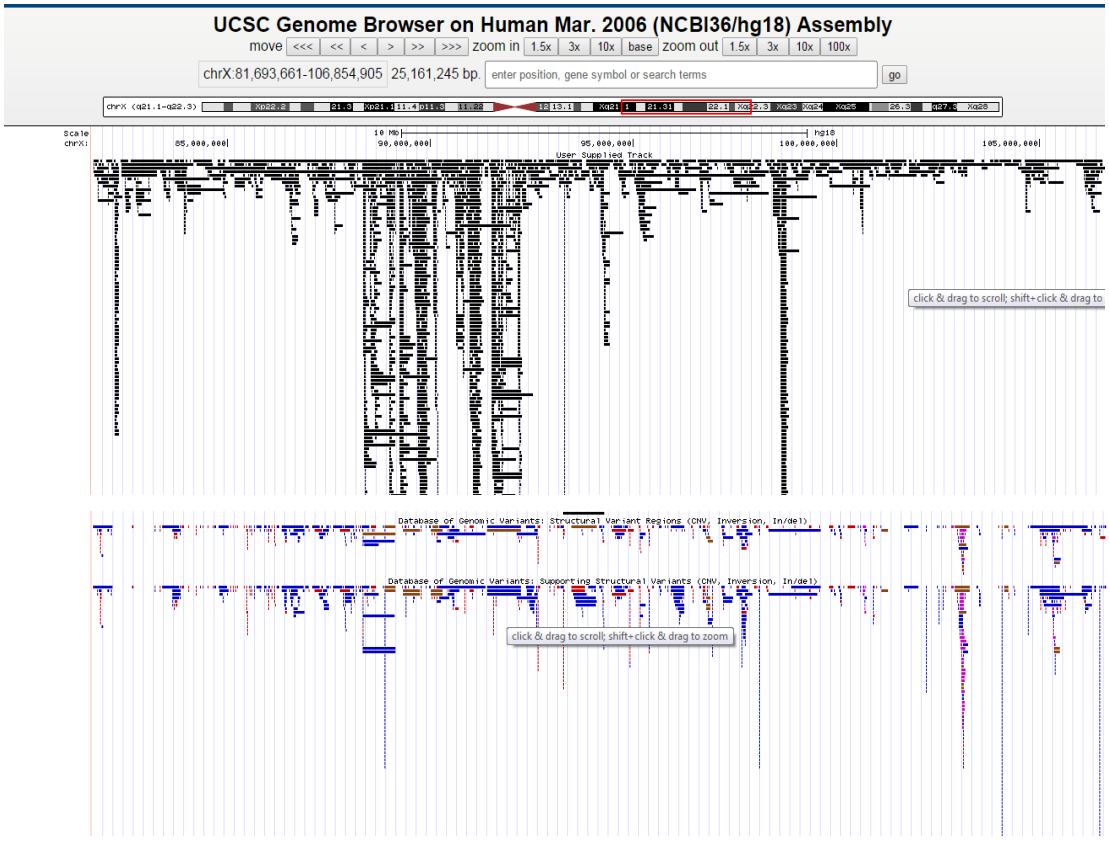


Fig 5.5 The most common CNV regions in chromosome X of the Thai population compared to the DGV database.

5.2.2 Common CNV in chromosome X

The most common CNV region in chromosome x (hg18 location chrX:81693661-106854905) has been found 38% of this Thai population. This particular region was used to compare to another database. The DGV (Database Of Genomic Variant) database was selected to compare with our data. This database included CNV information that were mostly collected from European populations with phenotype information. DGV database was used to explore a common CNV in Thai population and also in other populations. In summary, although CNV is one type of genetic variation, but the common CNV among the Thais was also common in other populations (MacDonald et al., 2014). This result reconfirmed that we achieved the most possibly accurate CNV calling.

5.3 HAPMAP3 characteristics

5.3.1 Characteristics of CNV and CNVR

Previously analyzed and published CNVs from 11 global ancestry groups: ASW (African ancestry in Southwest USA); CEU (Utah residents with Northern and Western European ancestry from the CEPH collection); CHD (Chinese in Metropolitan Denver, Colorado); GIH (Gujarati Indians in Houston, Texas); LWK (Luhya in Webuye, Kenya); MEX (Mexican ancestry in Los Angeles, California); MKK (Maasai in Kinyawa, Kenya); TSI (Tuscans in Italy); YRI (Yoruba in Ibadan, Nigeria) on the HAPMAP3 project were downloaded from the HAPMAP project website (http://HapMap.ncbi.nlm.nih.gov/downloads/cnv_data/hm3_cnv_submission.txt).

Table 5.5 The numbers of individuals in each HAPMAP3 populations.

Population	Total numberin each population
CHB	84
CHD	85
JPT	86
LWK	90
YRI	113
ASW	47
MKK	142
CEU	165
TSI	88
GIH	88
MEX	50

We assigned 11 global ancestry groups into 3 continents: Asian populations (CHB, CHD, JPT) African populations (ASW, LWK, MKK, YRI) and European populations (CEU, MEX, GIH, TSI). We obtained 1,038 individuals and 79,517 CNVs, which passed the same filtering criteria with the Thai CNVs from this study. The numbers of individuals in each group was shown in Table 5.5.

Table 5.6 HAPMAP3 CNVs and their CNVR characteristics.

	HAPMAP3 CNVs	HAPMAP3 CNVRs
Total count	79,517	506
CN-gain count	10,990	50
CNV-loss count	68,527	426
Multiplex		30
Average number per genome	76.6	70.33
Median size (range) (Kbp)	14.46(5.00-456.89)	11.89(5.00-456.89)
Median size of CN-gains	57.89(10.05-456.89)	26.52(10.05-304.50)
Median size of CN-losses	11.95(5.00-456.89)	10.34(5.00-231.38)
Genome coverage		16.257Mb (0.5 %)

The characteristics of HAPMAP3 CNVs were obtained in order to make a comparison with the Thai CNVs. The median size of HAPMAP3 CNVs was 14.46 Kbp (ranging from 5.00 Kbp to 456.89 Kbp). We found approximately 76.6 CNVs per individual (ranging from 54 to 138), higher than the Thai CNVs. The number of deletion CNVs was higher than duplication CNV: 86% and 14%, respectively. The median size of CNVs was 11,950 bp for deletions and 57,890 bp for duplications.

We identified 506 CNVRs in total, 426 deletion CNVRs, 50 duplication CNVRs and 30 complex CNVRs, which contains both duplication and deletion, deletion CNVRs were more common than duplication. The median size of HAPMAP3 CNVR was approximately 11.89 Kbp (ranging from 5.00 Kbp to 456.89 Kbp), and CNVR regions of the HAPMAP3 accounted for roughly 0.5% of the genome, which was lower than that of our Thai CNV result. A pie chart was used to represent the number of HAPMAP3 CNVs (Fig 5.7). In summary, 45.79 % of the CNVs overlap with genes, and 54.21% do not overlap with any known

genes.

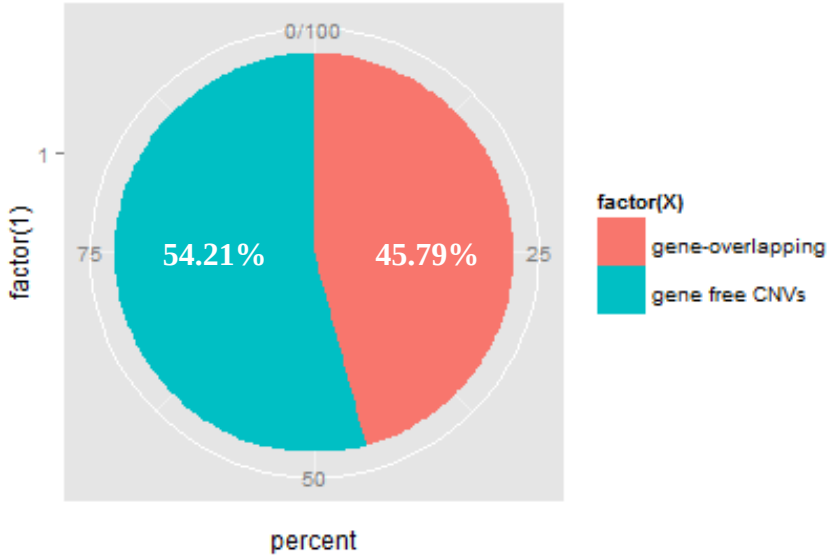


Fig 5. 6 The percentage of HAPMAP3 CNV, which overlap or without overlap with known genes

5.3.2 Principle Component Analysis of HAPMAP3

Principle Component Analysis was used to analyze the similarity of CNVs annotated in the HAPMAP3 CNV data in order to evaluate its population structure. The plot of the first two principal components (PC1 vs' PC2) was generated using the states of CNVs in each HAPMAP3 CNVR. A scatter plot of the first two components do not clearly distinguish between Asian populations (CHB, CHD, JPT), African populations (ASW, LWK, MKK, YRI) or European population (CEU, MEX, GIH, TSI). In this scatter plot, each color represents a population from HAPMAP3. Surprisingly, the majority of African, Asian and Europeans were grouped separately, but some populations were clustered into continent, which is not consistent to their ancestry. For example, some individuals of the African populations (e.g. ASW) and Asian populations (e.g. CHB) were clustered with European populations (e.g. CEU). Based on these results, the copy state of CNV was not enough to distinguish patterns among population.

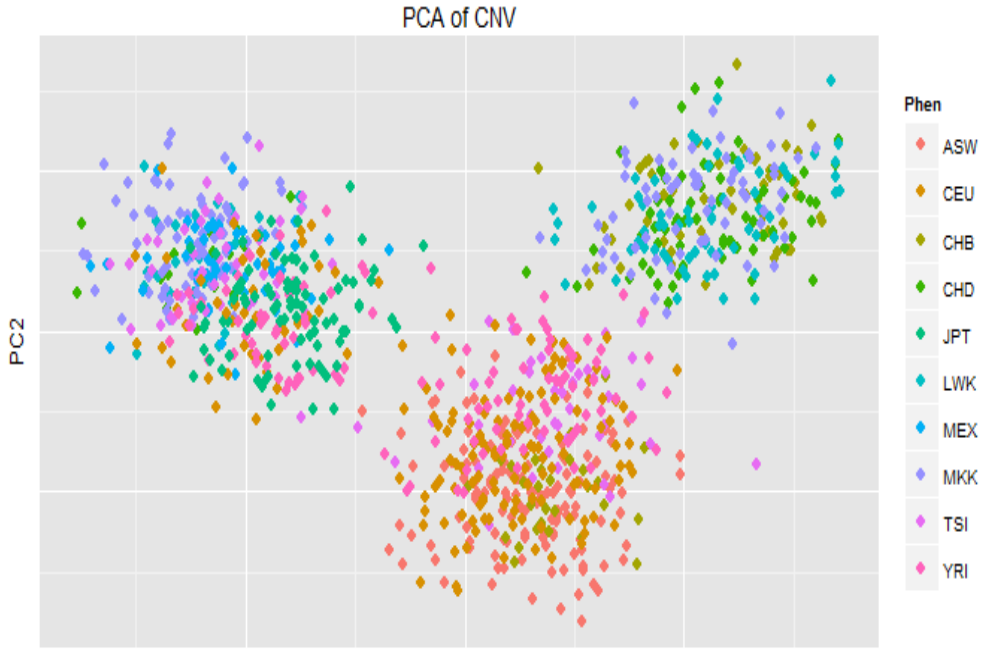


Fig 5.7 Principle Components Analysis using the copy number state of each CNV in HAPMAP3 CNVR in each population group of HAPMAP3.

5.4 Comparison between Thai and Hapmap CNVs

5.4.1 Size of CNV distribution comparison between Thais and HAPMAP3

Having analysed the Thai and Hapmap CNV characteristics, we addressed one of the main hypotheses of this project: By comparing CNV in Thais and HAPMAP3 populations, is the CNV frequency in the Thai population different from other Asian populations? Frequency and size distribution were two parameters used as a first round of comparison. Sizes of CNVs were plotted in a wide range to capture all size distribution ranges. The smallest size of CNV in the graph was in 1-10 Kbp and the largest size of CNV in the graph was more than 1 Mbp. After comparing the size distribution of CNV between Thai and Hapmap populations, the results showed small CNVs (1-50 Kbps)were common in HAPMAP3, but slightly less common in the Thai population.

In contrast, the larger size CNVs (> 50 Kbp) were found more commonly in the Thai population than the HAPMAP3 populations. Although, the number of CNVs in each range was different in the two populations, the pattern of the size distribution was similar. The proportions of small size CNV from the two populations were higher than the large size CNVs. The highest number of CNV was in the 10-50 Kbp range in both populations.

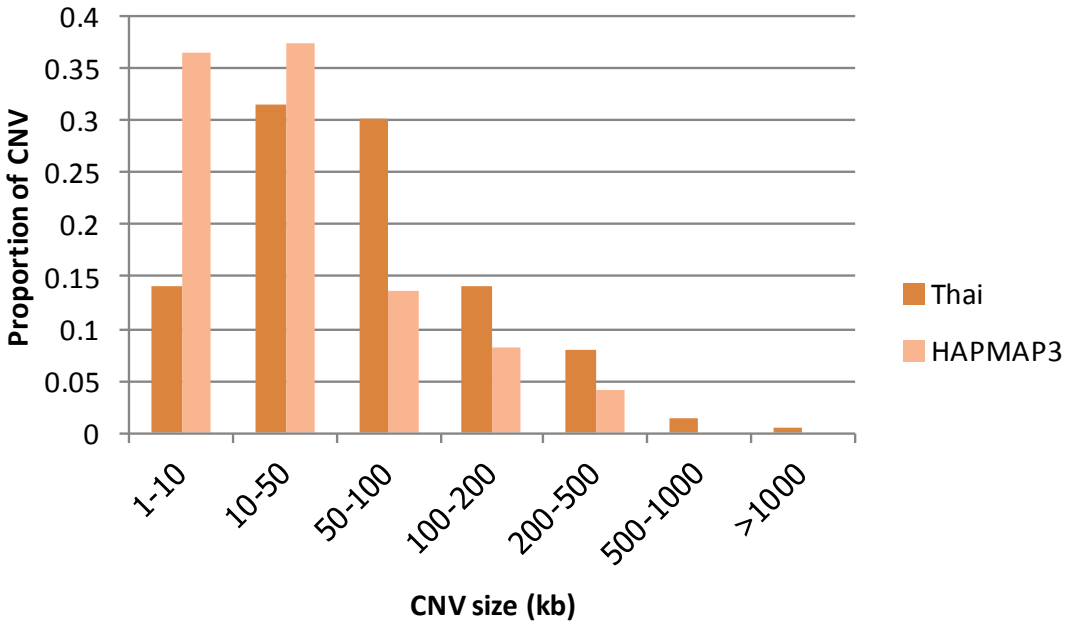


Fig 5.8. Size distributions of Thai and HAPMAP3 CNVs.

5.4.2 Degree of match between the Thai and HAPMAP3 CNVR

The CNVR terms were used to represent the uniqueness of overlapping CNV regions, the Thai and HAPMAP3 CNVR were created separately to explore common CNV between them. The frequency of Thai CNVR was calculated from the number of Thai individuals whose CNVs fall into each CNVR divided by the total number of Thai individuals. The frequency of Thai CNVRs was divided into ranges, with the lowest less than 1 percent frequency and the highest more than 50 % frequency. In Figure 5.10, each frequency range was compared separately with the HAPMAP3 CNVR to identify the degree of match between Thais and HAPMAP3 CNVR. When the frequency was less than 1%, the graph shows less than 20% were overlapped while more than 20% frequency shows nearly 100% overlap between the Thai and Hapmap CNVR. These results suggest that the degree of match increases when higher frequency CNVRs were used in the comparison.

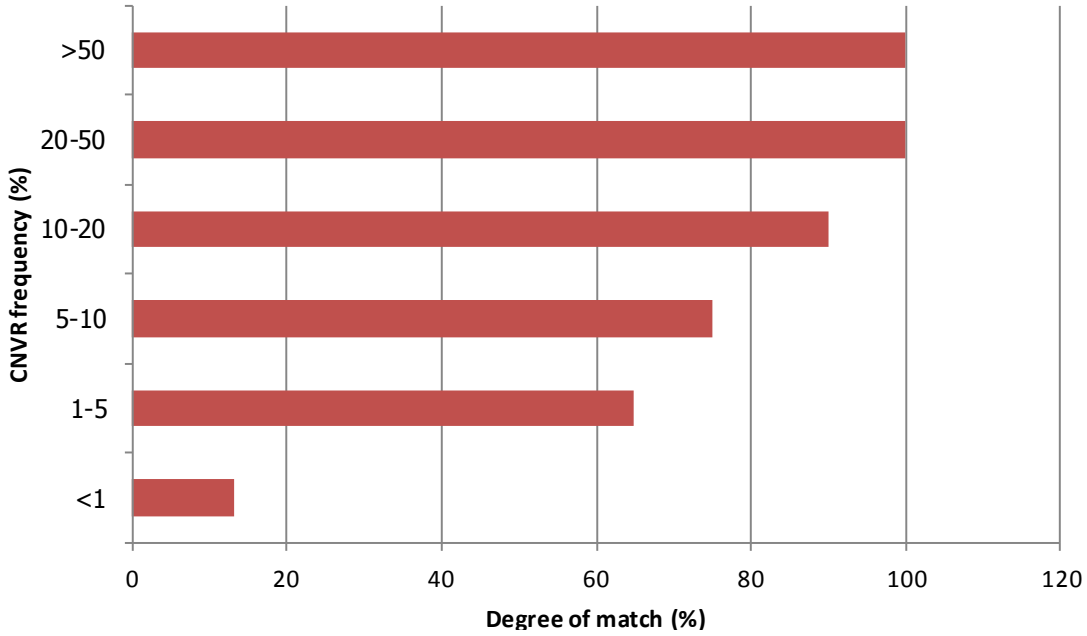


Fig 5.9 Degrees of match between Thais and HAPMAP3 CNVRs.

Considering only the Thai CNVRs, 822 of 1,014 CNVRs (81.14%) did not overlap with any HAPMAP3 CNVRs, while only 192 CNVRs were common in both Thai and HAPMAP3 populations. The median size of common CNVRs was 174.1 Kbp, with a mean allele frequency of 2.65%. On the other hand, the median size of unique CNVRs found only in the Thai population was 83.8 Kbp, with a mean allele frequency of 0.26%.

5.4.3 Frequency pattern comparison between Thais and HAPMAP3

All population CNVRs were created to represent a unique CNV region between the Thai and HAPMAP3 populations. After combining CNV data from the Thai and HAPMAP3 populations together, 2,560 CNVRs were generated. Allele frequency spectrum of CNVs with at least 1% across the Thai and HAPMAP3 CNVRs were selected to represent a pattern of frequency in each population as shown in Figure 5.11. The frequency pattern told us the same result with the degree of match pattern between the Thai and HAPMAP3 CNVRs. The most frequent range in the Thai

and HAPMAP3 populations displayed the same frequency pattern. However, there were slightly differences in the less than 0.1 and more than 0.5 frequency ranges. The Thais show the highest number of CNVs in the less than 0.1 range and lowest number of CNVs in more than 0.5 frequency range.

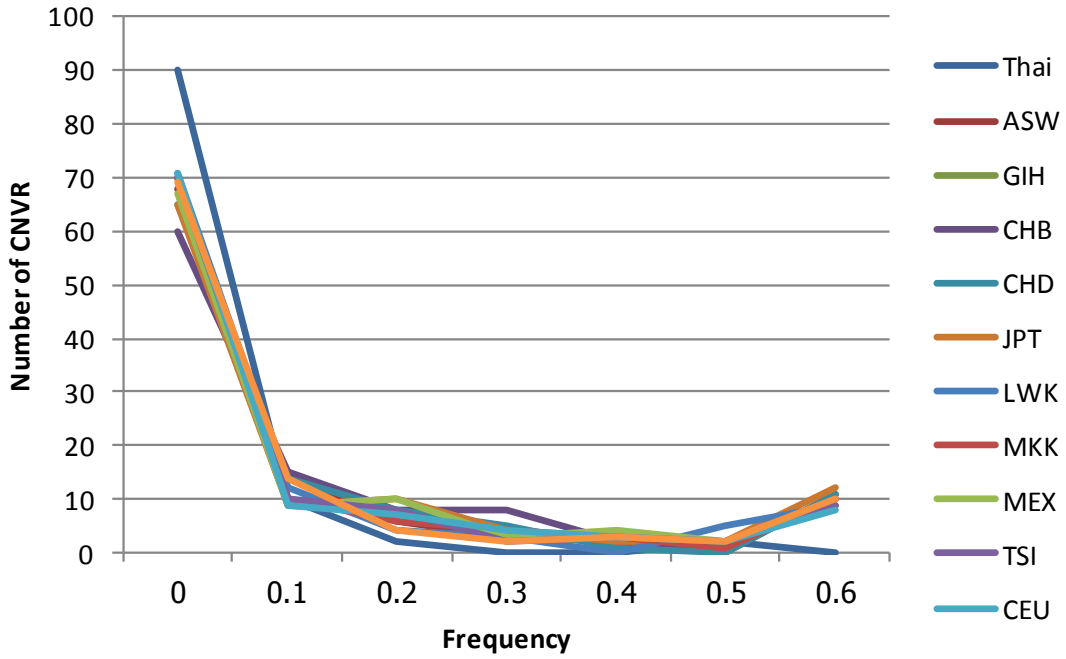


Fig 5.10 Frequency patterns of the Thai and HAPMAP3 CNVRs.

The CNVRs containing at least 5 % frequency was defined as common CNVRs. Common Thai CNVR were summarized and compared with HAPMAP3 populations as shown in Table 5.7. Some common CNVR regions were found only in the Thai populations, such as chr15 : 32459510-32626310 and some common CNVR regions were found only in the Asian populations such as chr1 : 1870130199-187847262. This result shows that the analyzed CNVs contain variation between populations.

5.4.4 Common Thai CNV region compares to HAPMAP3 population

The most common CNVR was located in chr4: 69,045,672–69,258,302. This common region overlapped with the UGT2B15 and UGT2B17, encoding Uridine

diphospho-glucuronosyltransferase genes, in more than half of the Thai population (1,564, 52%). In the majority of the Thais, the CNVs overlapping UGT2B17 were found as homozygous deleted (92.3%), similar to Japanese from Tokyo (JPT: 78.8%), Chinese from Beijing (CHB: 75.0%), and Chinese from Denver (CHD: 73.8%) as shown in Table 5.9, with very high frequency when compare to European and Africa populations. The lowest proportion of people containing a homozygous deletion of UGT2B17 was the African populations, with 9.8 % and 9.5 % in Yoruban in Ibadan, Nigeria (YRI) and African ancestry in Southwest USA (ASW), respectively.

Table 5.7 Common CNVRs with at least 5 % allele frequency in Thai population and their frequencies across HAPMAP3

ID	Chr	Start	Stop	Genes	THAI	CHB	CHD	JPT	ASW	LWK	MKK	YRI	GIH	MEX	TSI	CEU
1	1	187013019	187847262		0.06	0.04	0.05	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	34554235	35281044		0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	3	163995351	164108689		0.52	1.00	0.95	0.98	0.66	0.86	0.89	0.78	0.42	0.40	0.56	0.49
4	3	163690547	163719579		0.17	0.37	0.28	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
5	3	65163493	65190844		0.06	0.05	0.07	0.13	0.00	0.00	0.01	0.00	0.09	0.14	0.06	0.13
6	3	116125098	116154405	ZBTB20	0.05	0.02	0.01	0.03	0.43	0.59	0.32	0.65	0.23	0.04	0.01	0.02
7	3	53001754	53021256	SFMBT1	0.05	0.14	0.11	0.05	0.06	0.00	0.00	0.00	0.06	0.30	0.18	0.16
8	4	69045672	69258302	TMPRSS11E2, TMPRSS11E, UGT2B17, UGT2B15	0.52	0.95	0.99	0.99	0.45	0.63	0.68	0.36	0.80	0.56	0.58	0.57
9	4	63352170	63377531		0.08	0.64	0.72	0.71	0.21	0.18	0.09	0.13	0.56	0.36	0.26	0.34
10	4	64328367	64483913		0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	8	39350791	39509376		0.20	0.24	0.24	0.26	0.15	0.14	0.42	0.14	0.67	0.72	0.68	0.72
12	8	15444945	15580087	TUSC3	0.08	0.12	0.08	0.08	0.00	0.02	0.13	0.00	0.06	0.00	0.06	0.07
13	8	115595696	115932676		0.07	0.27	0.24	0.53	0.21	0.30	0.32	0.27	0.25	0.40	0.31	0.22
14	11	81181640	81203793		0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	13	56604813	56850408	FLJ40296	0.10	0.39	0.32	0.41	0.87	0.91	0.81	0.91	0.07	0.22	0.17	0.21
16	14	40671757	40744653		0.14	0.24	0.32	0.28	0.04	0.02	0.04	0.00	0.19	0.50	0.30	0.38
17	14	105069589	105997070		0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	14	42420358	44320168	FSCB	0.07	0.17	0.13	0.09	0.00	0.00	0.00	0.00	0.00	0.10	0.01	0.00
19	15	18294933	22368232	A26B1, CYFIP1, GOLGA8E, LOC283755, LOC283767, MAGEL2, MKRN3, NDN, NIPA1, NIPA2, OR4M2, OR4N4, TUBGCP5	0.12	0.20	0.11	0.10	0.11	0.08	0.08	0.11	0.24	0.22	0.06	0.16
20	15	32459510	32626301	GOLGA8A, GOLGA8B	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	16	31909325	33867424	LOC729355, LOC729355	0.20	0.75	0.79	0.71	0.36	0.38	0.29	0.31	0.50	0.90	0.75	0.67
22	17	41519743	42137359	ARL17, KIAA1267, LRRC37A2, LRRC37A, NSF	0.22	0.74	0.71	0.70	0.70	0.51	0.63	0.56	0.84	0.78	0.91	0.91
23	17	14030694	15533487	CDRT15, CDRT1, CDRT4, COX10, FAM18B2, FLJ45831, HS3ST3B1, PMP22, TEKT3, TRIM16	0.07	0.36	0.38	0.31	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.00
24	18	64862553	64919011	CCDC102B	0.24	0.25	0.26	0.43	0.00	0.01	0.05	0.00	0.02	0.06	0.05	0.05
25	19	20385941	20559157	ZNF826	0.11	0.21	0.18	0.23	0.17	0.11	0.24	0.19	0.19	0.10	0.15	0.09

Table 5.7 Common CNVRs with at least 5% allele frequency in Thai population and their frequencies across HAPMAP3.

	ASW	YRI	MKK	LWK	TSI	CEU	MEX	GIH	CHD	CHB	JPT	THAI
Total population frequency	0.447	0.363	0.676	0.633	0.580	0.570	0.560	0.795	0.988	0.952	0.988	0.522
Homozygous deletion frequency	0.042	0.036	0.225	0.122	0.125	0.158	0.100	0.341	0.729	0.714	0.779	0.482
Homozygous deletion proportion	0.095	0.098	0.333	0.193	0.216	0.277	0.179	0.429	0.738	0.750	0.788	0.923

5.4.5 Hierarchical clustering analysis (HCA) of most common gene among Thais and HAPMAP3

To confirm the result from CNVR suggesting that CNVs were variable between different groups of populations. Next, the frequency of the Thai CNVs was calculated versus the frequency of the HAPMAP3 populations by pairwise comparison. The cut off p- value < 0.01 was used as a criteria to select a CNV, which is different in frequency between the Thai and HAPMAP3 populations. By this method, we discovered 173 genes overlapping with CNVs as in Table 5.9.

Table 5.9 173 genes list which are overlap to difference frequency CNV between Thai and HAPMAP3.

GENE	CHR	START	END	THAI	CHB	CHD	JPT	ASW	LWK	MKK	YRI	GIH	MEX	TSI	CEU
KIF1B	1	10193350	10364248	0.000	0.000	0.000	0.000	0.085	0.067	0.035	0.221	0.000	0.020	0.000	0.000
PRAMEF1	1	12774132	12778810	0.010	0.095	0.106	0.116	0.043	0.056	0.085	0.027	0.068	0.100	0.045	0.085
PRAMEF2	1	12839527	12844351	0.009	0.095	0.106	0.116	0.043	0.056	0.085	0.027	0.068	0.100	0.045	0.085
CROCC	1	17121031	17172061	0.003	0.833	0.812	0.826	0.617	0.733	0.761	0.788	0.511	0.500	0.750	0.636
ELA3A	1	22200735	22211622	0.001	0.107	0.118	0.163	0.106	0.178	0.106	0.080	0.057	0.020	0.057	0.018
FGGY	1	59535212	60000990	0.000	0.012	0.000	0.116	0.255	0.211	0.254	0.239	0.125	0.100	0.068	0.030
INADL	1	61980736	62402179	0.000	0.000	0.000	0.000	0.021	0.000	0.007	0.080	0.000	0.000	0.000	0.000
FCGR1C	1	147635917	147644921	0.006	0.000	0.000	0.012	0.000	0.011	0.000	0.009	0.045	0.020	0.034	0.024
GBA	1	153470866	153481112	0.000	0.000	0.000	0.000	0.106	0.100	0.077	0.097	0.000	0.020	0.011	0.000
NME7	1	167368392	167603810	0.000	0.298	0.212	0.186	0.277	0.244	0.465	0.133	0.375	0.380	0.545	0.576
CFHR3	1	195010552	195029496	0.005	0.202	0.153	0.105	0.596	0.589	0.451	0.796	0.602	0.220	0.420	0.394
CFHR1	1	195055483	195067942	0.006	0.202	0.153	0.105	0.596	0.589	0.451	0.796	0.602	0.220	0.420	0.394
CFHR4	1	195123834	195154386	0.018	0.083	0.035	0.023	0.043	0.056	0.007	0.071	0.057	0.020	0.011	0.000
HHAT	1	208568228	208916255	0.000	0.000	0.000	0.000	0.043	0.056	0.014	0.062	0.000	0.000	0.000	0.000
LOC149643	1	211070107	211087614	0.002	0.036	0.047	0.058	0.021	0.000	0.042	0.000	0.023	0.020	0.080	0.103
EGLN1	1	229568053	229627413	0.000	0.036	0.071	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GALM	2	38746742	38815031	0.001	0.000	0.000	0.000	0.043	0.011	0.028	0.000	0.000	0.020	0.011	0.121
SFRS7	2	38824244	38832005	0.000	0.000	0.000	0.000	0.043	0.011	0.028	0.000	0.000	0.020	0.011	0.121
ANKRD36B	2	97487692	97572761	0.000	0.714	0.612	0.558	0.681	0.567	0.634	0.460	0.420	0.660	0.807	0.861
DPP10	2	114916368	116318406	0.044	0.000	0.000	0.000	0.064	0.056	0.007	0.080	0.000	0.000	0.000	0.000
CNTNAP5	2	124499333	125389333	0.001	0.000	0.000	0.000	0.128	0.078	0.021	0.080	0.000	0.020	0.000	0.000
GPR39	2	132890616	133120639	0.000	0.012	0.000	0.000	0.085	0.067	0.000	0.124	0.000	0.000	0.000	0.006
SESTD1	2	179674663	179837595	0.000	0.179	0.176	0.291	0.660	0.600	0.725	0.699	0.682	0.480	0.557	0.352
ZNF385B	2	180014955	180434477	0.005	0.000	0.000	0.000	0.298	0.378	0.155	0.292	0.114	0.060	0.068	0.121
BMPR2	2	202949294	203140719	0.000	0.000	0.000	0.000	0.255	0.211	0.120	0.416	0.000	0.020	0.000	0.000
ERBB4	2	211948686	213111597	0.001	0.000	0.000	0.000	0.021	0.000	0.000	0.000	0.170	0.000	0.057	0.139
TM4SF20	2	227935117	227952266	0.000	0.000	0.000	0.000	0.064	0.000	0.000	0.000	0.023	0.060	0.159	0.133
UGT1A8	2	234191029	234346684	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
UGT1A10	2	234209861	234346690	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
UGT1A9	2	234245282	234346690	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
UGT1A7	2	234255322	234346684	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
UGT1A6	2	234265059	234346690	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
UGT1A5	2	234286376	234346684	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
UGT1A4	2	234292176	234346684	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
UGT1A3	2	234302511	234346684	0.000	0.000	0.000	0.000	0.021	0.078	0.099	0.044	0.000	0.000	0.000	0.000
CTDSPL	3	37878672	38000964	0.045	0.155	0.141	0.035	0.064	0.011	0.042	0.027	0.216	0.080	0.273	0.121

Table 5.9 173 genes list which are overlap to difference frequency CNV between the Thai and HAPMAP3.

ULK4	3	41263093	41978664	0.005	0.000	0.000	0.000	0.043	0.033	0.049	0.062	0.000	0.020	0.000	0.000
SFMBT1	3	52913666	53055110	0.052	0.143	0.106	0.047	0.064	0.000	0.000	0.000	0.057	0.300	0.182	0.158
EPHA3	3	89239363	89613974	0.015	0.024	0.024	0.012	0.043	0.000	0.000	0.000	0.023	0.100	0.102	0.036
ZBTB20	3	115540212	116348817	0.053	0.024	0.012	0.035	0.426	0.589	0.317	0.655	0.227	0.040	0.011	0.018
NLGN1	3	174598937	175483810	0.000	0.000	0.000	0.000	0.085	0.067	0.021	0.080	0.000	0.000	0.000	0.000
LEPREL1	3	191157315	191321412	0.001	0.798	0.694	0.570	0.404	0.344	0.211	0.372	0.614	0.380	0.489	0.497
SLIT2	4	19864332	20229886	0.000	0.000	0.000	0.000	0.021	0.011	0.035	0.080	0.000	0.020	0.011	0.006
KCNIP4	4	20339336	21559472	0.001	0.179	0.200	0.198	0.468	0.544	0.408	0.513	0.068	0.060	0.000	0.000
SCFD2	4	53433907	53926999	0.000	0.000	0.000	0.012	0.043	0.000	0.021	0.053	0.000	0.000	0.000	0.000
UGT2B17	4	69085497	69116840	0.518	0.952	0.988	0.988	0.447	0.633	0.676	0.363	0.795	0.560	0.580	0.570
GPR103	4	122469246	122521631	0.000	0.310	0.235	0.140	0.106	0.122	0.120	0.204	0.023	0.060	0.148	0.085
GYPB	4	145136706	145159946	0.007	0.000	0.000	0.000	0.064	0.067	0.021	0.124	0.000	0.020	0.000	0.012
LRBA	4	151405045	152156329	0.000	0.000	0.000	0.000	0.468	0.311	0.246	0.469	0.000	0.040	0.000	0.000
TRIM61	4	166095047	166118268	0.000	0.000	0.000	0.000	0.106	0.078	0.211	0.159	0.023	0.000	0.000	0.000
GALNT17	4	172971149	174198133	0.000	0.298	0.447	0.314	0.766	0.778	0.690	0.726	0.591	0.740	0.784	0.776
LOC285501	4	178886900	179148663	0.001	0.000	0.000	0.000	0.021	0.044	0.042	0.053	0.000	0.020	0.000	0.000
COMMD10	5	115448625	115656877	0.002	0.000	0.000	0.012	0.000	0.011	0.014	0.071	0.011	0.040	0.000	0.000
PCDHA1	5	140146059	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA2	5	140154627	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA3	5	140160966	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA4	5	140166855	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA5	5	140181544	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA6	5	140187833	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA7	5	140194152	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA8	5	140201090	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA9	5	140207540	140372113	0.001	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
PCDHA10	5	140215817	140372113	0.000	0.060	0.059	0.058	0.128	0.256	0.289	0.106	0.068	0.160	0.125	0.206
BTNL8	5	180258734	180310512	0.000	0.548	0.400	0.523	0.234	0.200	0.113	0.150	0.193	0.520	0.614	0.533
BTNL3	5	180348506	180366333	0.000	0.548	0.400	0.523	0.234	0.200	0.113	0.150	0.193	0.520	0.614	0.533
DUSP22	6	237100	296355	0.002	0.952	0.894	0.919	1.000	1.000	0.993	0.991	0.989	0.940	0.966	0.976
HLA-A29.1	6	30018304	30085130	0.014	0.000	0.000	0.000	0.000	0.078	0.000	0.053	0.000	0.000	0.000	0.000
HSPA1A	6	31891269	31893698	0.001	0.000	0.000	0.023	0.021	0.033	0.014	0.009	0.000	0.020	0.034	0.006
HSPA1B	6	31903490	31906010	0.001	0.000	0.000	0.023	0.021	0.033	0.014	0.009	0.000	0.020	0.034	0.006
HLA-DRB5	6	32593131	32605984	0.004	0.988	0.988	0.953	1.000	0.989	1.000	0.973	0.932	1.000	0.977	0.958
HLA-DRB1	6	32654524	32665540	0.001	0.988	0.988	0.953	1.000	0.989	1.000	0.973	0.932	1.000	0.977	0.958
HLA-DQA1	6	32713160	32719407	0.000	0.893	0.965	0.930	0.936	0.833	0.944	0.894	0.898	1.000	0.943	0.915
HLA-DQB1	6	32735634	32742444	0.000	0.810	0.918	0.814	0.787	0.644	0.817	0.770	0.750	0.980	0.864	0.824
FKBP5	6	35649344	35764692	0.000	0.000	0.000	0.000	0.021	0.011	0.007	0.053	0.000	0.000	0.000	0.000
BTBD9	6	38244204	38715902	0.000	0.000	0.000	0.000	0.043	0.067	0.028	0.062	0.011	0.000	0.000	0.000
NKAIN2	6	124166767	125188485	0.001	0.000	0.000	0.000	0.000	0.000	0.035	0.000	0.023	0.020	0.011	0.073
SLC22A2	6	160557781	160599949	0.000	0.000	0.000	0.000	0.043	0.100	0.092	0.035	0.000	0.000	0.011	0.000
PARK2	6	161688579	163068824	0.009	0.000	0.000	0.000	0.000	0.011	0.000	0.071	0.000	0.000	0.000	0.000
C7orf28A	7	5904866	5932129	0.000	0.071	0.024	0.047	0.000	0.000	0.014	0.000	0.000	0.000	0.000	0.000
NPSR1	7	34664421	34884469	0.000	0.000	0.000	0.000	0.064	0.033	0.021	0.142	0.000	0.000	0.000	0.000
ABCA13	7	48208388	48657637	0.000	0.000	0.000	0.000	0.000	0.000	0.035	0.000	0.045	0.020	0.034	0.036
TYW1	7	66099251	66341933	0.001	0.155	0.141	0.221	0.596	0.656	0.669	0.575	0.534	0.500	0.477	0.479
HIP1	7	75001344	75206215	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.109
MAGI2	7	77484309	78920826	0.000	0.000	0.000	0.000	0.021	0.011	0.000	0.071	0.000	0.000	0.000	0.000

Table 5.9 173 genes list which are overlap to difference frequency CNV between the Thai and HAPMAP3.

LHFPL3	7	103756339	104336239	0.000	0.000	0.000	0.000	0.043	0.056	0.021	0.071	0.000	0.020	0.000	0.000
IMMP2L	7	110090345	110989583	0.007	0.012	0.000	0.012	0.043	0.033	0.028	0.080	0.034	0.000	0.000	0.012
C7orf60	7	112246437	112367168	0.000	0.083	0.118	0.128	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MGAM	7	141342147	141453016	0.033	0.333	0.294	0.256	0.128	0.033	0.127	0.000	0.227	0.220	0.375	0.345
CNTNAP2	7	145444385	147749019	0.003	0.000	0.000	0.000	0.021	0.044	0.007	0.115	0.000	0.000	0.000	0.000
CSMD1	8	2780281	4839736	0.012	0.012	0.000	0.000	0.085	0.178	0.176	0.150	0.000	0.000	0.000	0.006
SGCZ	8	13991743	15140163	0.001	0.000	0.000	0.000	0.000	0.022	0.049	0.035	0.000	0.000	0.000	0.000
TUSC3	8	15442100	15666366	0.080	0.119	0.082	0.081	0.000	0.022	0.134	0.000	0.057	0.000	0.057	0.067
PSD3	8	18429092	18915476	0.001	0.000	0.000	0.000	0.021	0.089	0.063	0.009	0.034	0.020	0.023	0.012
NKAIN3	8	63324054	64066182	0.001	0.000	0.000	0.000	0.255	0.322	0.063	0.381	0.000	0.020	0.000	0.000
RALYL	8	85258007	85996633	0.000	0.000	0.000	0.000	0.277	0.178	0.077	0.142	0.011	0.000	0.045	0.055
GRHL2	8	102573843	102751128	0.000	0.000	0.000	0.000	0.106	0.111	0.077	0.248	0.000	0.020	0.011	0.006
KANK1	9	494702	736103	0.008	0.000	0.000	0.000	0.149	0.322	0.225	0.239	0.000	0.000	0.023	0.000
RLN1	9	5324968	5329873	0.015	0.036	0.071	0.000	0.064	0.044	0.021	0.009	0.023	0.020	0.000	0.024
APBA1	9	71235021	71477042	0.006	0.000	0.047	0.035	0.213	0.111	0.176	0.150	0.023	0.040	0.011	0.012
OR13C5	9	106400558	106401515	0.001	0.000	0.000	0.058	0.064	0.111	0.092	0.142	0.000	0.000	0.000	0.000
MPP7	10	28379928	28611073	0.000	0.000	0.000	0.000	0.021	0.044	0.000	0.035	0.000	0.000	0.000	0.000
PARG	10	50696330	51041337	0.000	0.024	0.012	0.035	0.000	0.011	0.000	0.018	0.023	0.020	0.000	0.000
DMBT1	10	124310170	124393242	0.000	0.274	0.200	0.209	0.000	0.033	0.077	0.027	0.284	0.080	0.182	0.158
CYP2E1	10	135190856	135202610	0.022	0.024	0.035	0.035	0.191	0.133	0.162	0.168	0.000	0.040	0.034	0.030
SYCE1	10	135217394	135232866	0.021	0.024	0.035	0.035	0.191	0.133	0.162	0.168	0.000	0.040	0.034	0.030
OR52N1	11	5765659	5766622	0.000	0.369	0.412	0.267	0.319	0.367	0.570	0.451	0.466	0.460	0.455	0.388
NELL1	11	20647711	21553577	0.001	0.000	0.000	0.000	0.000	0.156	0.035	0.097	0.000	0.000	0.000	0.000
TRIM48	11	54786233	54795171	0.000	0.417	0.518	0.465	0.106	0.078	0.134	0.195	0.534	0.200	0.227	0.158
OR8U8	11	55899675	56265136	0.000	0.000	0.000	0.000	0.106	0.022	0.035	0.018	0.000	0.000	0.000	0.000
CNTN5	11	98397080	99732683	0.027	0.095	0.082	0.058	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006
SLC35F2	11	107166926	107234864	0.000	0.012	0.000	0.000	0.021	0.000	0.000	0.000	0.034	0.000	0.034	0.048
CACNA1C	12	2032676	2677376	0.001	0.012	0.000	0.000	0.021	0.000	0.000	0.000	0.034	0.180	0.011	0.145
SLC2A14	12	7857663	7916762	0.028	0.036	0.024	0.047	0.106	0.000	0.035	0.062	0.045	0.020	0.068	0.067
SLC2A3	12	7963093	7980138	0.028	0.036	0.024	0.047	0.106	0.000	0.035	0.062	0.045	0.020	0.068	0.067
PRR4	12	10889714	11215480	0.000	0.417	0.329	0.349	0.681	0.600	0.683	0.637	0.409	0.680	0.602	0.770
PRH1	12	10924826	11215477	0.000	0.417	0.329	0.349	0.681	0.600	0.683	0.637	0.409	0.680	0.602	0.770
PRB1	12	11396023	11399791	0.016	0.119	0.106	0.023	0.340	0.233	0.232	0.274	0.170	0.100	0.182	0.188
PRB2	12	11435740	11439768	0.012	0.119	0.106	0.023	0.340	0.233	0.232	0.274	0.170	0.100	0.182	0.188
LOH12CR1	12	12401321	12511105	0.000	0.012	0.012	0.000	0.064	0.022	0.035	0.071	0.068	0.040	0.102	0.055
PTPRO	12	15366753	15641602	0.000	0.000	0.000	0.000	0.064	0.089	0.000	0.186	0.000	0.000	0.000	0.000
SLC2A13	12	38435089	38785928	0.000	0.000	0.000	0.000	0.021	0.044	0.000	0.062	0.000	0.000	0.000	0.000
FAM19A2	12	60388307	60872818	0.000	0.000	0.000	0.000	0.128	0.089	0.021	0.124	0.000	0.000	0.000	0.000
MGAT4C	12	84897167	85756812	0.001	0.048	0.071	0.058	0.149	0.189	0.056	0.292	0.023	0.020	0.011	0.000
ANKS1B	12	97653201	98902563	0.001	0.012	0.000	0.000	0.000	0.011	0.106	0.000	0.000	0.000	0.000	0.000
LHFP	13	38815028	39075356	0.000	0.000	0.000	0.000	0.021	0.056	0.056	0.009	0.000	0.000	0.000	0.000
FLJ40296	13	56619622	56622644	0.003	0.155	0.082	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
RNASE3	14	20429401	20430347	0.006	0.036	0.024	0.047	0.106	0.156	0.099	0.150	0.000	0.040	0.011	0.018
DHRS4	14	23492804	23508328	0.000	0.024	0.000	0.035	0.043	0.067	0.021	0.027	0.023	0.000	0.023	0.024
DHRS4L2	14	23527866	23545459	0.001	0.024	0.000	0.035	0.043	0.067	0.021	0.027	0.023	0.000	0.023	0.024
KIAA0391	14	34661526	34813021	0.000	0.655	0.659	0.605	0.149	0.022	0.077	0.018	0.295	0.320	0.284	0.248
PRKCH	14	60858267	61087451	0.000	0.000	0.000	0.000	0.000	0.022	0.021	0.053	0.000	0.000	0.000	0.000
HEATR4	14	73014944	73095404	0.000	0.952	0.976	0.965	0.532	0.544	0.634	0.522	0.557	0.620	0.693	0.733
CHRFAM7A	15	28440734	28473156	0.001	0.988	1.000	0.977	0.915	0.978	0.979	0.956	1.000	1.000	0.955	0.982
ARHGAP11B	15	28706170	28718305	0.000	0.000	0.000	0.128	0.000	0.000	0.000	0.009	0.000	0.000	0.000	0.000
CKMT1B	15	41672543	41678896	0.000	0.012	0.012	0.023	0.021	0.056	0.007	0.035	0.023	0.000	0.011	0.012
STRC	15	41679052	41698290	0.002	0.012	0.012	0.023	0.021	0.056	0.007	0.035	0.023	0.000	0.011	0.012
CATSPER2	15	41710063	41728331	0.002	0.012	0.012	0.023	0.021	0.056	0.007	0.035	0.023	0.000	0.011	0.012
CKMT1A	15	41772375	41778712	0.000	0.012	0.012	0.023	0.021	0.056	0.007	0.035	0.023	0.000	0.011	0.012
KCTD5	16	2672498	2699032	0.001	0.024	0.071	0.140	0.106	0.078	0.225	0.035	0.261	0.380	0.352	0.248
C16orf63	16	15867077	15889948	0.002	0.000	0.000	0.000	0.043	0.067	0.028	0.053	0.000	0.000	0.000	0.000
NOMO2	16	18418682	18480935	0.000	0.083	0.106	0.070	0.043	0.022	0.042	0.000	0.011	0.020	0.000	0.000
HPR	16	70654625	70668646	0.001	0.000	0.000	0.000	0.149	0.167	0.056	0.292	0.000	0.020	0.000	0.000
TBC1D26	17	15576315	15588823	0.005	0.000	0.000	0.012	0.000	0.022	0.049	0.106	0.000	0.000	0.000	0.000
KRT33B	17	36773271	36779573	0.000	0.000	0.000	0.000	0.191	0.144	0.113	0.239	0.000	0.000	0.000	0.000
KIAA1267	17	41463128	41605371	0.005	0.000	0.000	0.012	0.255	0.011	0.077	0.000	0.682	0.360	0.659	0.727
ARL17	17	41732689	41794876	0.178	0.738	0.706	0.698	0.617	0.500	0.592	0.558	0.568	0.620	0.580	0.576
TRIM37	17	54414781	54539048	0.000	0.000	0.000	0.174	0.000	0.022	0.014	0.000	0.000	0.000	0.011	0.000
DOK6	18	65219270	65660359	0.013	0.024	0.059	0.035	0.000	0.000	0.000	0.000	0.068	0.060	0.091	0.079

Table 5.9 173 genes list which are overlap to difference frequency CNV between the Thai and HAPMAP3.

ZNF799	19	12361829	12373034	0.000	0.012	0.000	0.000	0.043	0.000	0.014	0.035	0.000	0.000	0.000	0.006
ZNF826	19	20366359	20399602	0.000	0.214	0.176	0.233	0.170	0.111	0.239	0.195	0.193	0.100	0.148	0.091
PSG1	19	48063197	48075711	0.025	0.060	0.035	0.035	0.064	0.089	0.042	0.071	0.114	0.160	0.080	0.067
PSG11	19	48203648	48222471	0.025	0.083	0.035	0.070	0.170	0.089	0.106	0.133	0.136	0.200	0.080	0.073
PSG2	19	48260201	48278654	0.012	0.060	0.000	0.035	0.106	0.056	0.063	0.080	0.034	0.140	0.034	0.006
PSG5	19	48363734	48382528	0.009	0.060	0.000	0.035	0.106	0.056	0.063	0.080	0.034	0.140	0.034	0.006
PSG4	19	48388693	48401630	0.004	0.060	0.000	0.035	0.106	0.056	0.063	0.080	0.034	0.140	0.034	0.006
ZNF285A	19	49581647	49597617	0.000	0.000	0.000	0.000	0.000	0.011	0.014	0.035	0.000	0.000	0.000	0.000
ZNF229	19	49622265	49644505	0.000	0.000	0.000	0.000	0.000	0.011	0.014	0.035	0.000	0.000	0.000	0.000
SIGLEC14	19	56837617	56841944	0.000	0.738	0.671	0.756	0.404	0.522	0.451	0.451	0.523	0.340	0.420	0.352
ZNF468	19	58033596	58052714	0.002	0.024	0.082	0.128	0.021	0.000	0.014	0.027	0.057	0.000	0.000	0.018
LILRA6	19	59432280	59438536	0.001	0.131	0.118	0.209	0.574	0.567	0.387	0.540	0.295	0.240	0.409	0.412
KIR3DP1	19	59927812	60001550	0.003	0.071	0.035	0.012	0.128	0.200	0.099	0.150	0.125	0.020	0.034	0.018
KIR2DL3	19	59941785	59956316	0.000	0.071	0.035	0.012	0.128	0.200	0.099	0.150	0.125	0.020	0.034	0.018
KIR2DS4	19	60035985	60051835	0.000	0.071	0.035	0.012	0.128	0.200	0.099	0.150	0.125	0.020	0.034	0.018
SIRPB1	20	1493028	1548689	0.000	0.405	0.365	0.512	0.128	0.133	0.063	0.115	0.284	0.560	0.307	0.455
MACROD2	20	13924145	15981841	0.025	0.000	0.000	0.000	0.064	0.078	0.021	0.097	0.000	0.020	0.000	0.000
EYA2	20	44956915	45250899	0.000	0.119	0.153	0.070	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.000
BCAS1	20	51993485	52120711	0.002	0.000	0.000	0.000	0.043	0.011	0.070	0.009	0.068	0.100	0.125	0.067
CDH4	20	59260953	59945694	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.136	0.000	0.000	0.000
SNF1LK	21	43658826	43671430	0.000	0.000	0.000	0.000	0.021	0.111	0.134	0.000	0.034	0.020	0.034	0.048
DDTL	22	22639025	22644748	0.005	0.440	0.412	0.419	0.532	0.378	0.493	0.398	0.477	0.500	0.545	0.545
GSTT2	22	22652313	22656106	0.005	0.440	0.412	0.419	0.532	0.378	0.493	0.398	0.477	0.500	0.545	0.545
LRP5L	22	24077426	24088524	0.041	0.095	0.082	0.023	0.043	0.089	0.120	0.088	0.057	0.080	0.080	0.109
MKL1	22	39136237	39362636	0.000	0.000	0.000	0.000	0.000	0.044	0.085	0.009	0.000	0.000	0.000	0.000

Finally, we selected the top 20 genes with specific CNVs to the Thai population, that is, the genes showing the greatest difference in frequency between each Thai and HAPMAP3 popuation. The top 10 genes with highest frequencies and the bottom 10 genes with lowest frequencies in the Thai population were selected in each pairs. After we combined each 20 genes set of 11 pairs together then the redundant gene were ignored. In summary, 35 of non redundant gene were chosen. Hierarchical clustering analysis (HCA) was performed on the frequency of these overlapping genes in each population to group the most similar in term of frequency together. The Thai population was grouped with other Asian populations from HAPMAP3 (JPT, CHB and CHD), as shown in Figure 5.12, suggesting that the Thais are closest to Asian population, and more than the African and Europe populations. Other Europe (GIH , TSI , MEX) and Africa populations (LWK, YRI, ASW, MKK) in HAPMAP3 were grouped together with Europe and Africa clades.

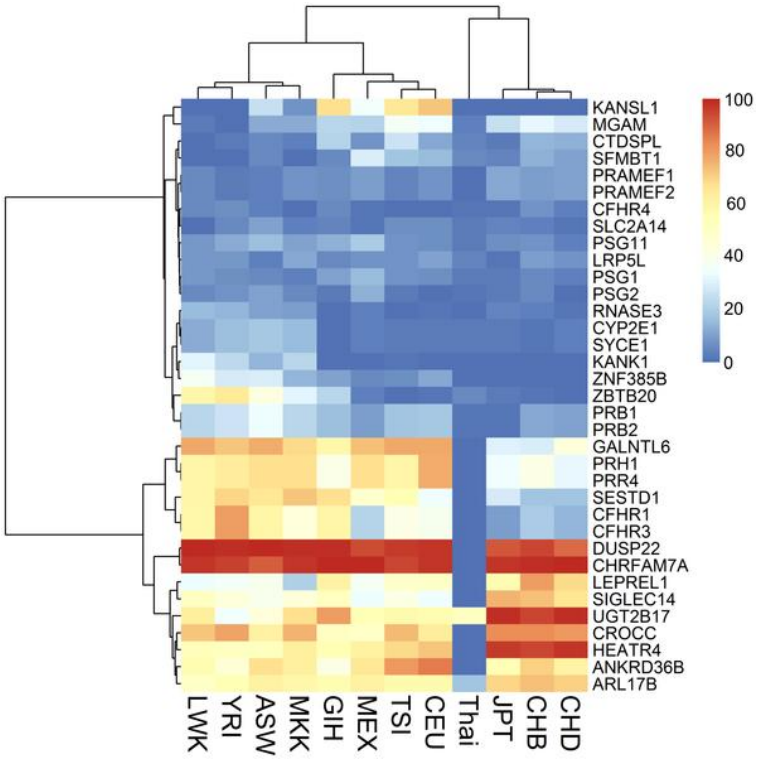


Fig 5.11 Hierarchical clustering analysis (HCA) of the 35 genes overlapping CNVs with statistically significantly different in allele frequencies between HAPMAP3 populations Thais (permutation P-value <0.01).

Table 5.10 35 non-redundant genes as used in the hierarchical clustering plot (Figure 5.11)

Gene	Function	Chr	Start	End	F_{Thai}	F_{CHB}	F_{CHD}	F_{JPT}	F_{ASW}	F_{LWK}	F_{MKK}	F_{YRI}	F_{GIH}	F_{MEX}	F_{TSI}	F_{CEU}
ANKRD36B	*ankyrin repeat domain containing protein*	2	97487692	97572761	0.000	0.71	0.61	0.56	0.68	0.57	0.63	0.46	0.42	0.66	0.81	0.86
ARL17B	*GTP binding protein*	17	41732689	41794876	0.178	0.74	0.71	0.70	0.62	0.50	0.59	0.56	0.57	0.62	0.58	0.58
CFHR1	*a secreted protein belonging to the complement factor H protein family*	1	195055483	195067942	0.006	0.20	0.15	0.10	0.60	0.59	0.45	0.80	0.60	0.22	0.42	0.39
CFHR3	*a secreted protein that binds to heparin, and may be involved in complement regulation*	1	195010552	195029496	0.005	0.20	0.15	0.10	0.60	0.59	0.45	0.80	0.60	0.22	0.42	0.39
CFHR4	*enhances the cofactor activity of CFH, and is involved in complement regulation*	1	195123834	195154386	0.018	0.08	0.04	0.02	0.04	0.06	0.01	0.07	0.06	0.02	0.01	0.00
CROCC	*major structural component of the ciliary rootlet*	1	17121031	17172061	0.003	0.83	0.81	0.83	0.62	0.73	0.76	0.79	0.51	0.50	0.75	0.64
CTDSPL	*a tumor suppressor gene involved in many types of cancer*	3	37878672	38000964	0.045	0.15	0.14	0.03	0.06	0.01	0.04	0.03	0.22	0.08	0.27	0.12
CYP2E1	*a member of the cytochrome P450 superfamily that is induced by ethanol, the diabetic state, and starvation*	10	135190856	135202610	0.022	0.02	0.04	0.03	0.19	0.13	0.16	0.17	0.00	0.04	0.03	0.03
DUSP22	*acts as a negative regulator of the ERalpha-mediated signaling pathway*	6	237100	296355	0.002	0.95	0.89	0.92	1.00	1.00	0.99	0.99	0.99	0.94	0.97	0.98
CHRFAM7A	*mediates fast signal transmission at synapses*	15	28440734	28473156	0.001	0.99	1.00	0.98	0.91	0.98	0.98	0.96	1.00	1.00	0.95	0.98
GALNTL6	*metal ion binding and polypeptide N-acetylgalactosaminyltransferase activity*	4	172971149	174198133	0.000	0.30	0.45	0.31	0.77	0.78	0.69	0.73	0.59	0.74	0.78	0.78
HEATR4	*HEAT repeat containing protein*	14	73014944	73095404	0.000	0.95	0.98	0.97	0.53	0.54	0.63	0.52	0.56	0.62	0.69	0.73
KANK1	*functions in cytoskeleton formation by regulating actin polymerization*	9	494702	736103	0.008	0.00	0.00	0.00	0.15	0.32	0.23	0.24	0.00	0.00	0.02	0.00
KANSL1	*nuclear protein that is a subunit of two protein complexes involved with histone acetylation*	17	41463128	41605371	0.005	0.00	0.00	0.01	0.26	0.01	0.08	0.00	0.68	0.36	0.66	0.73
LEPREL1	*plays a critical role in collagen chain assembly stability and cross-linking by catalyzing post-translational 3-hydroxylation of proline residues*	3	191157315	191321412	0.001	0.80	0.69	0.57	0.40	0.34	0.21	0.37	0.61	0.38	0.49	0.50
LRP5L	*Wnt-activated receptor activity and Wnt-protein binding*	22	24077426	24088524	0.041	0.10	0.08	0.02	0.04	0.09	0.12	0.09	0.06	0.08	0.08	0.11
MGAM	*plays a role in the final steps of digestion of starch*	7	141342147	141453016	0.033	0.33	0.29	0.26	0.13	0.03	0.13	0.00	0.23	0.22	0.38	0.35
PRAMEF1	*may function in reproductive tissues during development*	1	12774132	12778810	0.010	0.10	0.11	0.12	0.04	0.06	0.08	0.03	0.07	0.10	0.05	0.08
PRAMEF2	*may function in reproductive tissues during development*	1	12839527	12844351	0.009	0.10	0.11	0.12	0.04	0.06	0.08	0.03	0.07	0.10	0.05	0.08
PRB1	*encodes salivary proline-rich proteins (PRPs)*	12	11396023	11399791	0.016	0.12	0.11	0.02	0.34	0.23	0.23	0.27	0.17	0.10	0.18	0.19
PRB2	*encodes salivary proline-rich proteins (PRPs)*	12	11435740	11439768	0.012	0.12	0.11	0.02	0.34	0.23	0.23	0.27	0.17	0.10	0.18	0.19
PRH1	*proline-rich protein found in saliva*	12	10924826	11215477	0.000	0.42	0.33	0.35	0.68	0.60	0.68	0.64	0.41	0.68	0.60	0.77
PRR4	*proline-rich protein found in saliva*	12	10889714	11215480	0.000	0.42	0.33	0.35	0.68	0.60	0.68	0.64	0.41	0.68	0.60	0.77
PSG1	*a member of the immunoglobulin (Ig) superfamily produced by the placental syncytiotrophoblasts during pregnancy*	19	48063197	48075711	0.025	0.06	0.04	0.03	0.06	0.09	0.04	0.07	0.11	0.16	0.08	0.07
PSG11	*a member of the immunoglobulin (Ig) superfamily produced by the placental syncytiotrophoblasts during pregnancy*	19	48203648	48222471	0.025	0.08	0.04	0.07	0.17	0.09	0.11	0.13	0.14	0.20	0.08	0.07
PSG2	*a member of the immunoglobulin (Ig) superfamily produced by the placental syncytiotrophoblasts during pregnancy*	19	48260201	48278654	0.012	0.06	0.00	0.03	0.11	0.06	0.06	0.08	0.03	0.14	0.03	0.01
RNASE3	*major granule-derived protein with cytotoxic activity found in eosinophils and has been known as a useful marker of allergic inflammation*	14	20429401	20430347	0.006	0.04	0.02	0.05	0.11	0.16	0.10	0.15	0.00	0.04	0.01	0.02
SESTD1	*may act as a primary docking protein directing membrane turnover*	2	179674663	179837595	0.000	0.18	0.18	0.29	0.66	0.60	0.73	0.70	0.68	0.48	0.56	0.35
SFMBT1	*may be involved in antigen recognition*	3	52913666	53055110	0.052	0.14	0.11	0.05	0.06	0.00	0.00	0.00	0.06	0.30	0.18	0.16
SIGLEC14	*a member of sialic acid-binding lectins*	19	56837617	56841944	0.000	0.74	0.67	0.76	0.40	0.52	0.45	0.45	0.52	0.34	0.42	0.35
SLC2A14	*a member of the glucose transporter (GLUT) family that transport hexoses such as glucose and fructose into cells*	12	7857663	7916762	0.028	0.04	0.02	0.05	0.11	0.00	0.04	0.06	0.05	0.02	0.07	0.07
SYCE1	*involved in meiotic synapsis, cell cycle, and mitosis*	10	135217394	135232866	0.021	0.02	0.04	0.03	0.19	0.13	0.16	0.17	0.00	0.04	0.03	0.03
UGT2B17	*catalyzes the transfer of glucuronic acid from uridine diphosphoglucuronic acid to a diverse array of substrates*	4	69085497	69116840	0.518	0.95	0.99	0.99	0.45	0.63	0.68	0.36	0.80	0.56	0.58	0.57
ZBTB20	*negative regulation of transcription and DNA binding*	3	115540212	116348817	0.053	0.02	0.01	0.03	0.43	0.59	0.32	0.65	0.23	0.04	0.01	0.02
ZNF385B	*zinc ion binding and nucleic acid binding protein*	2	180014955	180434477	0.005	0.00	0.00	0.00	0.30	0.38	0.15	0.29	0.11	0.06	0.07	0.12

Fig 5.12 Screenshot of the Thai CNV database website and example search results.

5.4.6 Thai CNV database

As part of a team mentioned previously, I have identified specific CNVs that were found only the Thai population. We have catalogued common and unique Thai CNVs in a database that can be used to help genetic diagnostics. This database can be accessed at http://thaicnv.icbs.mahidol.ac.th/thaicnv, which is freely available and contains user-friendly interface for clinicians. The MySQL database schema employed was shown in Figure 5.13. The location CNVs in each chromosome can be used as a query to search more specific CNV. The searching result is shown in table and a list of CNV in a graphical interface. User can select the type of CNVs (deletion, duplication or both) and algorithms used to call CNV (PennCNV, CNVworkshop or both) shown in the result table. Moreover, the gene which overlapped with each CNV can be linked directly to RefGene commonly used in the UCSC genome browser, Ensembl, DGV, DECIPHER, and NCBI dbVar.

CHAPTER VI

Discussion and Conclusions

The Thai CNV database was set up with an aim to serve as a reference for the interpretation of future CNV results for the Thai population. The total Thai samples were collected from seven GWAS studies, all of which have already been published. Subjects in each GWAS study contained patients with different infectious diseases: tuberculosis, leprosy, Thyrotoxic Hypokalemic Periodic Paralysis (THPP), HIV/AIDS, and the patients with well-characterised genetic diseases Hb E/β-thalassemia, but no medical report for any subjects containing other genetic disorders at any diagnostic state. In this study, we also included β-thalassemia as subjects, even though this disease is a genetic disease, because the compound heterozygous mutations in the HBB gene is already well-studies and thus should not mask the CNV dectection.

From this information, we can assume that most of the CNVs detected in the Thai subjects are benign. To confirm this finding, we also used the DECIPHER database (v7.0) to evaluate to what extent the Thai CNVs annotated overlap with known chromosomal disorders. Indeed, only 3.27 % of the CNVs overlap with genetic disorder CNVs from DECIPHER. Therefore, this database is the largest representation of reference Thai CNVs so far, and is a suitable database for clinical interpretation of CNVs in the Thai samples

In summary, a total of 3,017 subjects were used in the CNV call as described in the results. In autosome, we combined the CNV results from two different algorithms: PennCNV and CNVWorkshop, to reduce possible algorithm-specific biases. We employed the most possible stringent criteria to filter the Thai CNV results. However, research studies reported many CNVs found on chromosome X are correlated with the neurodevelopmental diseases such as Intellectual Disability (Whibley et al., 2010) and Autism spectrum disorder (ASD) (Vissers, de Vries, & Veltman, 2010). Taking this point into account, a reference database for CNVs on

chromosome X is a valuable resource for clinical diagnosis. For CNV calling on chromosome X, only PennCNV contains this option. However, the challenge still remains due to the difference in neutral copy number between males and females on chromosome X. When calling CNVs with sex information by PennCNV, LRR value in HMM file will be changed to zero at copy state one in male, while in female copy state two is set as zero. (Kai Wang et al., 2007). When comparing CNV characteristics with autosomes, we found on average 4.79 CNVs per X chromosome. On the other hand, we found approximately 8 CNVs per individual in autosomes. Both autosomal and X chromosomes show comparable characteristic in terms of both CNVs and CNVRs. In both types of chromosomes analyzed, the number of deletion CNVs is higher than the number of duplication CNVs and the size of duplication CNVs is larger than deletion CNVs. The higher number of deletion CNVs might reflect a limitation of CNV calling algorithms, where it is easier to discover deletion CNVs than duplication CNVs (Kai Wang et al. 2007)

Principle Component Analysis (PCA) was used to identify the association between the CNV states and the phenotype groups in the Thai population analyzed. Both the state of CNV and the presence/absence pattern in each CNVR were used to perform PCA. We did not find evidence of any correlation between infectious disease state and CNV genotype. We have also performed a cluster analysis using the same dataset, and it also confirmed a lack of correlation. That is, there was no association between the states of CNVs and disease susceptibility for these particular diseases. This, in turn, confirms that the CNVs from this study are a suitable reference dataset for CNV analysis in the Thai population.

In addition to this, we addressed the possibility of bias from the two calling methods and found that the methods did not cause any observable difference in CNV calling. We also applied similar PCA to the HAPMAP3 populations and comapared with the Thai population. The result from the HAPAMP3 populations demonstrated similar results with that of the Thais, as in each cluster of the HAPMAP3 PCAs, there was a mixture of subjects from Asian, European and Africa populations. This pattern showed that state of CNVs in CNVRs alone are not sufficient to distinguish these clades.

We also investigated the details of the HAPMAP3 cases which showed a higher than average number of CNVs per individual (76.6). This could be because HAPMAP3 used two denser SNP genotyping arrays: Illumina 1M and Affymetrix 6.0, which allow superior detection of small CNVs. However, the estimated genome coverage of HAPMAP3 is 0.5% which is lower than in the Thai population (8.72 %). However, this figure is comparable with what has been estimated before (Altshuler et al., 2010), as 0.1% genome coverage in each individual genome was described. This difference in coverage might be due to different filtration criteria, after the filtration, 1,038 HAPMAP3 samples passed the criteria.

We further compared the CNVs found in the Thai and HAPMAP3 populations by focusing on the degree of matches between the CNVRs in the two populations, and the frequency patterns in each population. Only 20% of the Thai CNVRs are common when compared with the HAPMAP3 population. This large amount of Thai-specific CNVRs was expected, as there is no Thai population in HAPMAP3. Saying that, the most common CNVRs among the Thai (>5% frequency) still showed a high degree of match between with the HAPMAP3 population. These shared CNVs between populations are probably conserved in general, regardless of the ethnicity, which is consistent with a previous report (Yim et al., 2010b)

Even though the similarity in allele frequency and linkage disequilibrium between the Thais and East Asians is high, it has been shown that more than 5% of drug-related alleles in the Thais might not be captured by East Asian-derived haplotype-tagging SNPs (Mahasirimongkol et al., 2006). From our results, using only the state of CNV can not distinguish between population, so hierarchical clustering was created from the top ten genes with highest frequencies and the bottom ten genes with lowest frequencies in the Thai population. As expected, these group of genes can successfully separate 12 study populations into three clusters which are consistant to their ancestral origin. Thai population was clustered with other Asian populations (CHB, CHD and JPT), while Africans (LWK, ASW, MKK, YRI) and Europeans (GIH, MEX, TSI, CEU) were grouped separately according to their ancestry.

We have identified the gene Uridine diphospho-glucuronosyltransferase 2B17 (UGT2B17) as one of the genes found to be common in terms of CNVR, and it was one of the top 35 genes identified from pairwise comparison between the Thai and

HAPMAP3 populations. UGT2B17 encodes a protein involved in catalyzing transfer of glucuronic acid from uridine diphosphoglucuronic acid to diverse substrates, including steroid hormones. This enzyme is key to glucuronidation of androgens, a major source for estrogen. Both enzymes have been described to help stimulating bone formation. We have found a duplication in the UGT2B17 gene, which might be associated with increased risk of osteoporotic hip fracture in both the Chinese and Caucasian populations. In contrast, homozygous deletion of UGT2B17 might be a protective factor (Sambrook & Cooper, 2006; Yang et al., 2008). Our result also showed a higher frequency of UGT2B17 homozygous deletions in the East Asian populations than in the European and African populations, consistent with higher hip fracture rates among the Europeans. The number of UGT2B17 homozygous deletions in the Thai population is even lower than in the East Asians and Caucasians. However, more detailed studies are required to confirm the relationship between hip fracture in the Thai population and osteoporosis risk

ARL17 is among the top 35 genes and also overlapped with common Thai CNVR. *ARL17* gene encodes GTP binding protein that functions as an allosteric activator of the cholera toxin catalytic subunit, an ADP-ribosyltransferase (Pasqualato, Renault, & Cherfils, 2002). In one study, a CNV on chromosome 17 containing *ARL17* gene was reportedly associated with increasing antibody response to anthrax vaccine in both European Americans and African Americans (GenomicFalola et al., 2013). The frequency of a duplication CNV containing *ARL17* gene was different between African American and European population. European has a significantly higher frequency of this duplication than that of the African American population, (McElroy, Nelson, Caillier, & Oksenberg, 2009). In line with McElroy et al, in our study the duplication frequency of European is also higher than that of African, although not as striking (0.9 and 0.5 respectively). Our Thai data showed lower frequency of this duplication CNV as compared to European population (0.21 and 0.9 respectively). However, so far no study on the effect of this CNV on anthrax vaccine response has been conducted in any Asian population. Immuno-molecular biological study will be required to identify whether the same low antibody reponse after receiving anthrax vaccine will be observed in the Thai population.

In summary, we envisage that the Thai CNV database will contribute to the more accurate interpretation of uncertain clinical significance CNVs among the Thais, and serve as one of the most informative population-specific CNV reference databases for population geneticists.

REFERENCES

Abel, H. J., & Duncavage, E. J. (2013). Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genetics*, *206*(12), 432–40.

Almal, S. H., & Padh, H. (2012). Implications of gene copy-number variation in health and diseases. *Journal of Human Genetics*, *57*(1), 6–13.

Altshuler, D. M., Gibbs, R. a, Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–8.

Bailey, J. a, & Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews. Genetics*, *7*(7), 552–64.

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, *39*(7 Suppl), S16–21.

Chantarangsu, S., Mushiroda, T., Mahasirimongkol, S., Kiertiburanakul, S., Sungkanuparph, S., Manosuthi, W., … Nakamura, Y. (2011). Genome-wide association study identifies variations in 6p21.3 associated with nevirapine-induced rash. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *53*(4), 341–8.

Chen, W., Hayward, C., Wright, A. F., Hicks, A. a, Vitart, V., Knott, S., … Porteous, D. J. (2011). Copy number variation across European populations. *PloS One*, *6*(8), e23087.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*, *38*(1), 75–81.

Day, I. N. M. (2010). dbSNP in the detail and copy number complexities. *Human Mutation*, *31*(1), 2–4.

Falola, M. I., Wiener, H. W., Wineinger, N. E., Cutter, G. R., Kimberly, R. P., Edberg, J. C., … Shrestha, S. (2013). Copy Number Variants : Evidence for Association with Antibody Response to Anthrax Vaccine Adsorbed, *8*(5).

Fanciulli, M., Petretto, E., & Aitman, T. J. (2010). Gene copy number variation and common human disease. *Clinical Genetics*, *77*(3), 201–13.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews. Genetics*, *7*(2), 85–97.

Firth, H. V, Richards, S. M., Bevan, a P., Clayton, S., Corpas, M., Rajan, D., … Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, *84*(4), 524–33.

Fusté, B. (2012). “ Next-generation ” Sequencing (NGS): The new genomic revolution.

Gai, X., Perin, J. C., Murphy, K., O’Hara, R., D’arcy, M., Wenocur, A., … White, P. S. (2010). CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. *BMC Bioinformatics*, *11*(1), 74.

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human copy number variation and complex genetic disease. *Annual Review of Genetics*, *45*, 203– 26.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews. Genetics*, *10*(8), 551–64.

Hughes, S., Capper, R., Lam, S., & Sparkes, N. (2011). Sequencing and microarrays for genome analysis : complementary rather than competing ?

Iafrate, a J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., … Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36*(9), 949–51.

Institute NHGRI. (2012). A Brief History of the Human Genome Project. Retrieved from https://www.genome.gov/12011239

Jongjaroenprasert, W., Phusantisampan, T., Mahasirimongkol, S., Mushiroda, T., Hirankarn, N., Snabboon, T., … Nakamura, Y. (2012). A genome-wide association study identifies novel susceptibility genetic variation for

thyrotoxic hypokalemic periodic paralysis. *Journal of Human Genetics*, *57*(5), 301–4.

Karimpour-Fard, A., Dumas, L., Phang, T., Sikela, J. M., & Hunter, L. E. (2010). A survey of analysis software for array-comparative genomic hybridisation studies to detect copy number variation. *Human Genomics*, *4*(6), 421. doi:10.1186/1479-7364-4-6-421

Kearney, H. M., Thorland, E. C., Brown, K. K., Quintero-Rivera, F., & South, S. T. (2011). American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *13*(7), 680–5.

Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y., & Chia, K. S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics*, *55*(7), 403–15.

Kurotaki, N., Shen, J. J., Touyama, M., Kondoh, T., Visser, R., Ozaki, T., … Lupski, J. R. (2005). Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency. *Genetics in Medicine*, *7*(7), 479–483.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, *37*(13), 4181–93. 2

Lee, C., Iafrate, a J., & Brothman, A. R. (2007). Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature Genetics*, *39*(7 Suppl), S48–54.

Li, W., & Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiological Genomics*, *45*(1), 1–16.

LIFTON, R. P., DLUHY, R. G., POWERS, M., RICH, G. M., COOK, S., ULICK, S., & LALOUEL, J.-M. (1992). A chimaeric llβ-hydroxylasealdosterone synthase gene causes glucocorticoid-remediable aldosteronism and human. *Nature*, *355*, 262–265.

Lou, H., Li, S., Yang, Y., Kang, L., Zhang, X., Jin, W., … Xu, S. (2011). A map of copy number variations in Chinese populations. *PloS One*, *6*(11), e27341.

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, *42*(Database issue), D986–92.

Mahasirimongkol, S., Chantratita, W., Promso, S., Pasomsab, E., Jinawath, N., Jongjaroenprasert, W., … Sura, T. (2006). Similarity of the allele frequency and linkage disequilibrium pattern of single nucleotide polymorphisms in drug-related gene loci between Thai and northern East Asian populations: implications for tagging SNP selection in Thais. *Journal of Human Genetics*, *51*(10), 896–904.

Mahasirimongkol, S., Yanai, H., Mushiroda, T., Promphittayarat, W., Wattanapokayakit, S., Phromjai, J., … Tokunaga, K. (2012). Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. *Journal of Human Genetics*, *57*(6), 363–7.

Major Changes in Our DNA Lead to Major Changes in Our Thinking. (2012). Answering Big Questions 2012. Retrieved from http://www.genome.gov/27553258

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402.

McElroy, J. P., Nelson, M. R., Caillier, S. J., & Oksenberg, J. R. (2009). Copy number variation in African Americans. *BMC Genetics*, *10*, 15. doi:10.1186/1471-2156-10-15

Mitra, R. D., Shendure, J., Olejnik, J., & Church, G. M. (2003). Fluorescent in situ on polymerase colonies. *Analytical Biochemistry*, *320*(1), 55– 65.

NHGRI. (2012).International HapMap Project. Retrieved from https://www.genome.gov/HapMap/

Nowak, D., Hofmann, W.-K., & Koeffler, H. P. (2009). Genome-wide Mapping of Copy Number Variations Using SNP Arrays. *Transfusion Medicine and Hemotherapy : Offizielles Organ Der Deutschen Gesellschaft Fur Transfusionsmedizin Und Immunhamatologie*, *36*(4), 246–251.

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)*, *5*(4), 557–72.

Osiriphun, Y., Wongtrakoongate, P., Sanongkiet, S., Suriyaphol, P., Thongboonkerd, V., & Tungpradabkul, S. (2009). Identification and Characterization of RpoS Regulon and RpoS-Dependent Promoters in Burkholderia pseudomallei. *Enzyme*, 3118–3131.

Pasqualato, S., Renault, L., & Cherfils, J. (2002). Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for "front-back" communication. *EMBO Reports*, *3*(11), 1035–41.

Perry, G. H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A. S., … Redon, R. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Research*, *18*(11), 1698–710.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–75.

R Development Core Team. (2013). R : A Language and Environment for Statistical Computing, *1*.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–54.

Sambrook, P., & Cooper, C. (2006). Osteoporosis. *Lancet*, *367*(9527), 2010–8.

Sebat, J. (2007). Major changes in our DNA lead to major changes in our thinking. *Nature Genetics*, *39*(7 Suppl), S3–5. doi:10.1038/ng2095

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P. Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, *305*(5683), 525–8.

Stankiewicz, P., & Lupski, J. R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, *61*, 437–55.

Valsesia, A., Macé, A., Jacquemont, S., Beckmann, J. S., & Kutalik, Z. (2013). The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. *Frontiers in Genetics*, *4*(May), 92. d

Valsesia, A., Stevenson, B. J., Waterworth, D., Mooser, V., Vollenweider, P., Waeber, G.,… Bergmann, S. (2012). Identification and validation of copy number variants using SNP genotyping arrays from a large clinical cohort. *BMC Genomics*, *13*(1), 241.

Vandeweyer, G., & Kooy, R. F. (2013). Detection and interpretation of genomic structural variation in health and disease. *Expert Review of Molecular Diagnostics*, *13*(1), 61–82.

Velagaleti, G. V. N., Bien-Willner, G. a, Northup, J. K., Lockhart, L. H., Hawkins, J. C., Jalal, S. M., … Stankiewicz, P. (2005). Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. *American Journal of Human Genetics*, *76*(4), 652–62.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–51.

Visscher, P. M., Brown, M. a, McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24.

Vissers, L. E. L. M., de Vries, B. B. a, & Veltman, J. a. (2010). Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. *Journal of Medical Genetics*, *47*(5), 289–97.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., … Sebat, J. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, N.Y.)*, *320*(5875), 539–43.

Wang, K., & Bucan, M. (2010). Copy Number Variation Detection via High-Density SNP Genotyping. *Cold Spring Harbor Protocols*, *2008*(6)

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. a, … Bucan, M. (2007a). PennCNV: an integrated hidden Markov model designed for

high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17*(11), 1665–74.

Watson, J., & Crick, F. (1953). Molecular structure of nucleic acids. *Nature*. Retrieved from http://www.nature.com/physics/looking-back/crick/

Whibley, A. C., Plagnol, V., Tarpey, P. S., Abidi, F., Fullston, T., Choma, M. K., … Raymond, F. L. (2010). Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability. *American Journal of Human Genetics*, *87*(2), 173–88.

Winchester, L., Yau, C., & Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics & Proteomics*, *8*(5), 353–66.

Yang, T.-L., Chen, X.-D., Guo, Y., Lei, S.-F., Wang, J.-T., Zhou, Q., … Deng, H.-W. (2008). Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *American Journal of Human Genetics*, *83*(6), 663–74.

Yim, S.-H., Kim, T.-M., Hu, H.-J., Kim, J.-H., Kim, B.-J., Lee, J.-Y., … Chung, Y.-J. (2010). Copy number variations in East-Asian population and their evolutionary and functional implications. *Human Molecular Genetics*, *19*(6), 1001–8. doi:10.1093/hmg/ddp564

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–81.

APPENDIX

CNV Workshop default parameters

	Duplication	Homozygous Deletion	Heterozygous Deletion
Illumina 550k	Mean LRR≥0.2 and b2.sd>b3.sd	Mean LRR≤-2 and b2.sd≥b3.sd	-2<Mean LRR<=-0.3 and BAF_{40-60}<=4%
Illumina 610k	(Mean LRR≥0.2 and b2.sd>b3.sd) OR (0.2>Mean LRR≥0.12 and b2.sd>1.5* b3.sd)	Mean LRR≤-2 and b2.sd≥b3.sd	-2<Mean LRR<=-0.3 and BAF_{40-60}<=4%
Illumina Omni	(Mean LRR≥0.19 and b2.sd>b3.sd) OR (0.19>Mean LRR≥0.12 and b2.sd>1.4* b3.sd)	Mean LRR≤-1.7 and b2.sd≥b3.sd	-1.7<Mean LRR<=-0.24 and BAF_{40-60}<=6%
Affy 6.0	Mean LRR≥0.3 and b2.sd>b3.sd	Mean LRR≤-2 and b2.sd≥b3.sd	-2<Mean LRR<=-0.45 and BAF_{40-60}<=10%

BIOGRAPHY

NAME	Chaiwat Naktang
DATE OF BIRTH	15 May, 1990
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTION ATTENDED	Mahidol University, Thailand, 2008-2012 Bachelor of Science Mahidol University, Thailand, 2012-2014 Master of Biochemistry
SCHOLARSHIP	Full scholarship, Master of Science Program in Biochemistry
HOME ADDRESS	50/3 Moo 6 Bangkhuntain, Samaedam Bangkok, Thailand, 10150 Tell: (+66) 4167397 E-mail: naktang1@hotmail.com