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

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