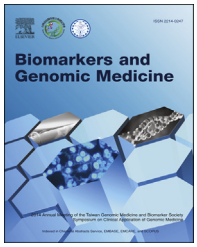
Biomarkers and Genomic Medicine (2014) 6, 133–143

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.j-bgm.com



ORIGINAL ARTICLE

Comprehensive analysis of common coding sequence variants in Taiwanese Han population

Ya-Chi Lin [a,e,f], Joseph T. Tseng [b,e], Shuen-Lin Jeng [c], H. Sunny Sun [d,e,f,*]

[a] Institute of Cancer Research, National Health Research Institutes, Miaoli, Taiwan
[b] Institute of Bioinformatics and Biosignal Transduction, College of Bioscience and Biotechnology, National Cheng Kung University, Tainan, Taiwan
[c] Department of Statistics, National Cheng Kung University, Tainan, Taiwan
[d] Institute of Molecular Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[e] Center for Genomic Medicine, National Cheng Kung University, Tainan, Taiwan
[f] Bioinformatics Center, National Cheng Kung University, Tainan, Taiwan

Received 7 March 2014; received in revised form 12 May 2014; accepted 16 May 2014
Available online 24 June 2014

KEYWORDS
exome sequencing;
loss-of-function;
molecular medicine;
population genetics;
sequence variants

Abstract The diversity of genomic variations exists among different ethnic populations. Information on population-specific genomic variants provides important insights to link between genotypes and phenotypes. To facilitate genomic medicine research, this study aims to detect and characterize sequence variations enriched in the coding regions of the genome in the Chinese population residing in Taiwan. DNAs from 11 unrelated Taiwanese individuals were enriched for coding regions (i.e., exome) and followed by deep sequencing. Approximately 30 Gb of high-quality data from massively parallel sequencing was obtained. On average, ~60% of the total reads were uniquely mapped to the human reference genome and overall 97% of the target regions were covered by sequence reads, resulting in an average enrichment fold relative to target size of ~50-fold. Comprehensive variant detection and analysis were performed with various in-house established bioinformatics pipelines, and information for different types of variations including single nucleotide variants, short insertions and deletions, and copy number variations was collected. The sequence variations were crossed with variants in the public databases to identify ethnic-specific variants. To study the impact of sequence variations that are enriched in the Taiwanese Han population, variants that are present in at least two exomes (i.e., minor allele frequency >9%) were further annotated. Overall, we detected 308 loss-of-function variants that belong to 291 genes in the Taiwanese Han Exome Sequencing dataset. Functional annotation revealed a significant pathological influence

* Corresponding author. Institute of Molecular Medicine, College of Medicine, National Cheng Kung University, 1 University Road, Tainan 70101, Taiwan.
E-mail address: hssun@mail.ncku.edu.tw (H.S. Sun).

http://dx.doi.org/10.1016/j.bgm.2014.05.001

of these loss-of-function-associated genes in the risk of various human diseases including lung cancer. This is the first NGS (next-generation sequencing)-generating dataset to comprehensively report coding sequence variants in the Taiwanese Han population. Given that the Taiwanese Han population is the Han Chinese residing in Taiwan, it is normally underrepresented in population-genetics studies. We believe the study will contribute valuable information that will have an impact on medical as well as population genetics.

Introduction

The completion of the Human Genome Project has identified an entire human reference sequence. As our knowledge of this sequence grows, the need to explore the levels of natural variation relative to this reference sequence also increases, among individuals and between human populations. Knowledge of DNA sequence variation among individuals plays an essential role in understanding the genetic background of the population. In addition, sequence information are also extremely important for decoding the impact of genetic variations—including single nucleotide variants (SNVs), short insertions and deletions (InDels), and copy number variations (CNVs)—on complex human diseases, conferring susceptibility or resistance, or influencing interaction with environmental factors.

The advances of DNA sequencing technologies lead to ultrahigh throughput detection of DNA sequence variants and shorten the duration from years to days for finding causal mutations in medical research. With next-generation sequencing (NGS) technology, whole genome sequencing (WGS) can detect DNA variants across the whole genome, whereas whole exome sequencing (WES or exome-seq) is focused on sequencing the protein coding exons and finding variants in the coding regions that usually result in a direct effect on protein functions and lead to many human diseases.[1–3] Compared to WGS, exome-seq is more cost-efficient, with enough depth of coverage to identify the variants especially in the coding regions.[4] Although WGS can discover more variants, it is currently not easy to systematically interpret the functional importance of the "noncoding variants." Moreover, other substantial drawbacks of WGS, such as the need for further analysis and storage of larger-scale data, make exome-seq currently more practical as a promising approach in molecular diagnostics for diseases caused by coding region variants.[5]

The variant profile established by exome-seq in a specific population is of fundamental importance for investigating population genetics and studying genomic medicine. Exome-seq has been applied in several large sequencing projects such as the 1000 Genomes Project (a collaboration of international research teams),[6,7] the National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project,[8] the National Institute of Environmental Health Sciences Environmental Genome Project (http://evs.gs.washington.edu/niehsExome/), and the UK10K Project (http://www.uk10k.org/). These multiple efforts indicate that elucidating the roles of coding variants in population genomics is an ongoing processes. However, only a limited number of ethnic populations are covered in the large-scale exome-seq projects. Thus, several sequencing-based studies have been carried out in specific populations, including Danes (Denmark) and Southeast Asian Malays, to profile genetic variations and identify functionally important variants.[9–11] These studies have strengthened the importance of population-specific variation in the exome to public health significance in human populations.

In this report, we applied deep exome sequencing and systematic workflows to profile genetic variations including SNVs, small InDels, and CNVs enriched in the coding regions of the Taiwanese Han population. The established pipeline is ready to be used in routine genomic study. This ethnic-specific information of Taiwanese Han coding variants will serve as a reference dataset to facilitate medical research in the Han Chinese population.

Materials and methods

Samples

A total of 11 unrelated individuals belonging to the Han Chinese population were randomly selected, and genomic DNA samples were obtained from the Taiwan Han Chinese Cell and Genome Bank.[12] Five male and six female individuals were included, and the available demographic information is listed in Table S1 in the supplementary material online. The quality and quantity of each DNA sample were calibrated using an ND-1000 spectrophotometer (NanoDrop, Wilmington, DE, USA) to ensure adequate input DNA for further experiments.

Library preparation, exome enrichment, and sequencing

Three micrograms of each DNA sample was used to prepare the fragment library for the SOLiD5500xl sequencer (Life Technologies, Grand Island, NY, USA). Targeted exomes were captured and enriched using the SureSelect Target Enrichment Kit (Agilent, Santa Clara, CA, USA) or the TargetSeq Exome Enrichment Kit (Life Technologies), both of which targeted ~20,000 genes covering 37–45 Mb genomic sequences. The captured exomes were subsequently used for emulsion polymerase chain reaction to amplify the DNA fragments onto the beads following the protocols provided