Genomics 105 (2015) 145–149

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Combinatorial approach to estimate copy number genotype using whole-exome sequencing data

Mi Yeong Hwang [1], Sanghoon Moon [1], Lyong Heo, Young Jin Kim, Ji Hee Oh, Yeon-Jung Kim, Yun Kyoung Kim, Juyoung Lee, Bok-Ghee Han, Bong-Jo Kim *

Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, 361-951, Republic of Korea

ARTICLE INFO

Article history:
Received 11 September 2014
Accepted 16 December 2014
Available online 20 December 2014

Keywords:
Copy number variation
Whole-exome sequencing
CNV discovery
Combinatorial approach
CNV genotyping

ABSTRACT

Copy number variations (CNVs) are known risk factors in complex diseases. Array-based approaches have been widely used to detect CNVs, but limitations of array-based CNV detection methods, such as noisy signal and low resolution, have hindered detection of small CNVs.
Recently, the development of next-generation sequencing techniques has increased rapidly owing to declines in cost. Particularly, whole-exome sequencing has proved useful for finding causal genes and variants in complex diseases. Because gene copy number may affect expression, CNV genotyping can be very valuable in disease association studies. However, almost all current CNV detection tools consider only two types of CNV genotypes. In this study, we propose a CNV genotype estimation approach using a combination of existing methods. Our approach was comprehensively compared with the customized Agilent array–comparative genomic hybridization. We found that our genotyping approach proved to be accurate, and reproducible, suggesting that it can complement existing CNV genotyping methods.

1. Introduction

Copy number variations (CNVs) are DNA deletions or duplications of ≥1 kb in the human genome [1,2]. Previous studies have reported that CNVs constitute risk factors that may predispose individuals to complex diseases including cancer, autism, and Parkinson's disease [3]. Array-based approaches such as array–comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) microarrays are widely used to identify CNVs [4]. Because the resolution and coverage of array-based approaches are strongly influenced by the number of probes and internal spacing between probes, these methods are limited in their precision in detection of CNVs with exact break points and their ability to detect small CNVs [5–9].

Recently, next-generation sequencing technologies have been applied to explore genetic variation because they are able to overcome some of the limitations of array-based approaches [4,6]. Because whole-exome sequencing (WES) focuses on protein coding regions rather than the entire genome, it is less biased and more cost-effective and time-efficient than whole-genome sequencing [10,11]. Therefore, WES is rapidly becoming a fundamental tool in functional genomic research and clinical diagnostics [7,12,13]. Four strategies, namely read depth, paired-end, split-read, and sequence assembly, are widely used in CNV discovery using WES data [12,14]. Of these algorithms, the read depth approach, which is based on mapping read counts to the reference genome at genomic regions, has been the most widely employed [7].

Diverse CNV analysis tools using WES data have been reported, such as ExomeDepth [15], cn.MOPs [16], GenomeSTRiP [17], and splitread [18,19]. Exact detection of CNV genotype is essential to determine associations with disease, but most of these CNV detection tools do not consider CNV genotypes but rather assign individuals to copy number classes, such as 0, 1, and 2 copies. Although the EXCAVATOR software is an exception, it is still unable to ascertain all CNVs in a dataset.

In this study, we propose a combinatorial approach for estimating CNV genotypes using the existing WES-based CNV detection and array-based CNV genotyping methods. To increase the reliability of detection of CNVs, we used two different WES-based CNV detection methods, ExomeDepth and cn.MOPs, to analyze data from 80 Korean individuals. The accuracy of the two methods was evaluated by comparison with the customized Agilent aCGH. Subsequently, CNVtools was used for CNV genotyping. Moreover, the CNV genotyping accuracy of our approach was evaluated by comparison with the customized Agilent aCGH and with results from EXCAVATOR.

* Corresponding author at: Division of Structural and Functional Genomics, Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, 361-951, Republic of Korea. Fax: +82 43 719 8908.
E-mail address: kbj6181@cdc.go.kr (B.-J. Kim).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ygeno.2014.12.003
0888-7543/

2. Materials and methods

2.1. Samples

A total of 80 healthy individuals (31 males and 49 females) were selected from a community-based cohort of an ongoing prospective study by the Korea Association REsource project (KARE). Peripheral blood samples were collected from the individuals with written informed consent. Genomic DNA was then extracted from the 80 blood samples, and 5 μg of each DNA sample was used for exome enrichment using the SureSelect v2 44 M system (Agilent Technologies, Santa Clara, CA). Raw FASTQ data files were generated with the Illumina HiSeq 2000 platform with a target average depth of coverage of 60× (Supplementary Table 1).

2.2. Data preprocessing

Data preprocessing consisted of five steps: trimming, alignment, removal of duplication, realignment, and recalibration (Supplementary Fig. 1). Specifically, Sickle was used to trim the FASTQ files based on quality values (https://github.com/najoshi/sickle), and trimmed reads were aligned with the Human Reference Genome (UCSC hg19) using the Burrows-Wheeler Aligner software package (http://bio-bwa.sourceforge.net/) for mapping low-divergence sequences. Duplicated reads were removed with Picard (http://picard.sourceforge.net). Misaligned reads around insertions/deletions (InDels) were realigned and base quality scores of sequencing-by-synthesis reads in the realigned BAM files were recalibrated using the Genome Analysis Toolkit software (http://www.broadinstitute.org/gatk/).

2.3. CNV calling

Read depth-based exome CNV-calling tools such as ExomeCNV [20], CONTRA [21], CoNIFER [22], ExomeDepth, and cn.MOPs can be categorized into three types according to the reference: single sample (no reference), paired case/control samples (matched reference), and a population of samples (the average value of the read depth from multiple samples) [18]. In this study, we selected ExomeDepth and cn.MOPs, both of which are of the population of samples reference type (Fig. 1). To call CNVs, we ran each tool with the author-recommended or default parameters on the data from the 80 individuals. Read counts were calculated using BAM files obtained from data preprocessing. cn.MOPs uses all of the test samples as the reference set, whereas in ExomeDepth a subset of samples needs to be designated as the reference. Therefore, we randomly selected ten individuals (five males and five females) as the reference set in ExomeDepth.

2.4. Determination of CNV regions

Among the CNVs identified in the calling process, we only considered CNVs of ≤5 Mb in length because CNVs are typically defined as genomic changes ≥1 kb but ≤5 Mb. Moreover, we excluded singleton CNVs from further study to enhance the reliability of our analysis. To determine CNV regions, we used a 50% reciprocal overlap criterion. If clusters of overlapping CNVs shared ≥50% of reciprocal overlap, we merged them into a CNV region (Supplementary Fig. 2).

2.5. CNV region validation

To evaluate the validity of the estimated CNVs, we used the Agilent SurePrint G3 Human 8 × 60 K aCGH microarray, which contains customized probes based on previously detected CNV regions in Koreans [23]. The same 80 DNA samples for WES were used for the customized Agilent aCGH experiment. CNV calling was performed with the Agilent Genomic Workbench software using the ADM-2 algorithm and default parameters. CNVs detected with Genomic Workbench contained at least three consecutive probes.

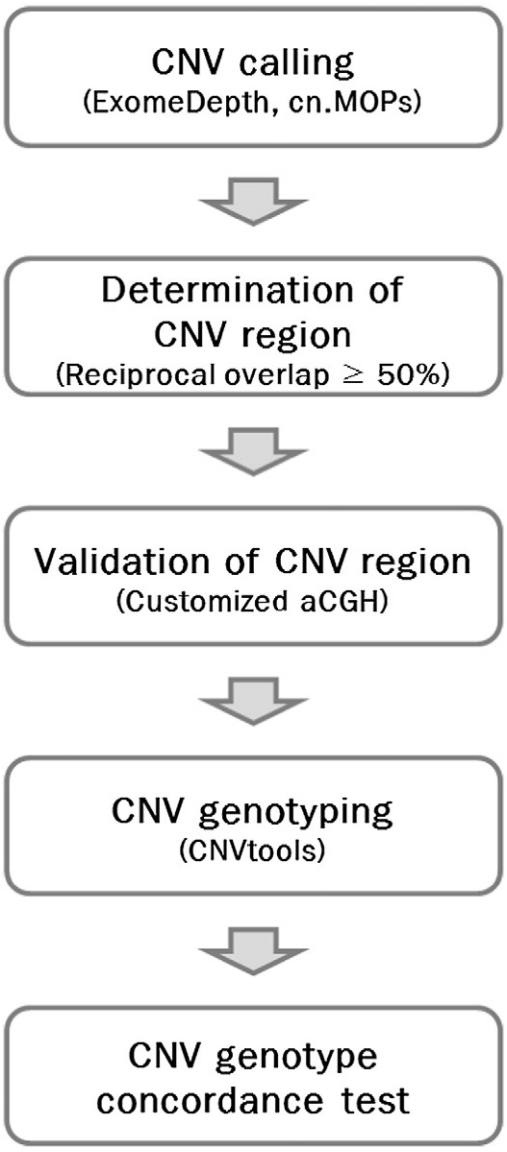

Fig. 1. Overall scheme of the combinatorial approach employed in our study. ExomeDepth and cn.MOPs were used for CNV detection in WES reads from 80 individuals. CNV regions were then defined according to a 50% reciprocal overlap criterion. The CNV regions were validated by comparison with the customized Agilent aCGH data from the same individuals. CNV genotypes were also compared to those of the customized Agilent aCGH, and the performance of our approach was compared with EXCAVATOR.

2.6. CNV genotype estimation and concordance test

Mathematically, read depth data from read depth-based next-generation sequencing are similar to the log ratio of signal intensity from aCGH [18]. Therefore, some classical algorithms for aCGH data are also applicable to CNV genotype estimation with read depth-based WES. In our study, we applied a combinatorial approach that combines CNV calling and genotyping software to estimate each CNV genotype. For CNV genotyping software, we selected CNVtools, which is an R package for CNV case–control and quantitative trait association [24]. In other words, the probe log2 ratio of each CNV call from ExomeDepth was used as an input for CNVtools. To evaluate the estimated CNV genotypes, we compared them to the customized Agilent aCGH results from the same sample. The CNVtools results and average log2 ratio values from the customized Agilent aCGH were used to assign individuals to CNV classes. We then conducted a concordance test of the CNV genotypes from ExomeDepth and CNVtools. In addition, to evaluate the performance