CNVRuler: a copy number variation-based case-control association analysis tool

Ji-Hong Kim1,†, Hae-Jin Hu1,†, Seon-Hee Yim1, Joon Seol Bae2, Seon-Young Kim3 and Yeun-Jun Chung1,*

1Integrated Research Center for Genome Polymorphism, Department of Microbiology, School of Medicine, Catholic University of Korea, Seoul 137-701, Korea, 2Laboratory of Genomic Diversity, Department of Life Science, Sogang University, Seoul 121-742, Korea and 3Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea

ABSTRACT
Summary: The method for genome-wide association study (GWAS) based on copy number variation (CNV) is not as well established as that for single nucleotide polymorphism (SNP)-GWAS. Although there are several tools for CNV association studies, most of them do not provide appropriate definitions of CNV regions (CNVRs), which are essential for CNV-association studies. Here we present a user-friendly program called CNVRuler for CNV-association studies. Outputs from the 10 most common CNV defining algorithms can be directly used as input files for determining the three different definitions of CNVRs. Once CNVRs are defined, CNVRuler supports four kinds of statistical association tests and options for population stratification. CNVRuler is based on the open-source programs R and Java from Sun Microsystems.

Availability: CNVRuler software is available with an online manual at the website, www.ircgp.com/CNVRuler/index.html

Contact: yeun@catholic.ac.kr.

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on January 29, 2012; revised on April 18, 2012; accepted on April 19, 2012

1 INTRODUCTION
Copy number variation (CNV) is thought to contribute to inter-individual differences in phenotypes including disease susceptibility (Feuk et al. 2006; McCarroll and Altshuler, 2007; Yim et al. 2010). Single nucleotide polymorphism (SNP) genotypes are always represented as categorical values, while copy number variations (CNVs) are often represented as regions consisting of continuous values for consecutive probes. For this reason, the following three steps are required to perform a CNV-based genome-wide association study (GWAS): (i) calling CNVs, (ii) merging CNVs into common CNV regions (CNVRs), and (iii) statistical analysis of the associations. Despite the importance and popularity of CNV-phenotype association studies, there are not many algorithms that provide all three key steps mentioned above.

2 DESCRIPTION
CNVRuler was designed to define three different types of CNVRs from the predefined CNVs and provides four statistical methods for CNV-based association studies. The overall analysis flow in CNVRuler is illustrated in Figure 1. All forms of major CNV call outputs from different segmentation tools such as Genotyping Console, Genome Studio, Genomic Workbench, PennCNV and Nexus can be processed without additional conversion steps. Details are described in the user manual.

2.1 Prerequisites
CNVRuler requires Java Runtime Environment from Sun Microsystems or the equivalent (JRE ≥ 1.6.0). The major functions of CNVRuler algorithms are implemented in R program as a calculation core.

2.2 Processing CNV data
Outputs from the 10 CNV defining algorithms can be directly used as input files for determining the CNVRs with CNVRuler [Supplementary Table S2]. Alternatively, CNVRuler can read a manually prepared custom tab-delimited text file of CNV information to build the CNVRs including next-generation sequencing data as a user-customized text file. Users can filter out the CNVs by size and signal intensity threshold. Details are available in the user manual.

2.3 Building CNVRs
For association analyses, each CNVR should have the same boundaries among subjects such that each subject will be coded as CNVR-gain positive, CNVR-loss positive, or diploid. CNVRuler supports three different definitions of common CNV regions.
Affymetrix SNP array 6.0 genotyping data of 10 individuals and
To validate the CNVRuler, we used the CNVs identified from
methods of CNVRuler were defined by CONAN (Supplementary
et al) defined CNVRs using CNVRuler and CONAN software (Forer
fragments, the calculation time is longer than that in other methods.
assess the regional density of the participating CNVs base-by-base
overlapping CNVs are extremely long (see user manual). In order
to minimize this possibility, CNVRuler provides the option to
particular CNVR aggregates in cases, but it is not found at all
ratios cannot be calculated or the approximation of significance
regression models adjusted for age and sex and for multiple
comparisons by the FDR method. This discrepancy may be partly
due to the different correction methods applied for multiple
comparisons and the difference between region-based and probe-
based analyses (Bae et al. 2010).
CNVRuler can handle both common and rare CNVs once CNVs
are called. Different from common CNPs, rare CNVs can cause
complete or quasi ‘separation’ in 2×2 tables, where the odds
ratios cannot be calculated or the approximation of significance
is inadequate. For example, complete separation occurs when a
particular CNVR aggregates in cases, but it is not found at all in
controls. In these situations, users can select the χ² test with
Yates’ continuity correction or Fisher’s exact test. There are different
opinions regarding which of these two methods to choose; so users
should use their own statistical knowledge and discretion. After this
step, users can perform exact logistic regression analysis or other
methods specialized for dealing with a small number of events or
small sample sizes, but CNVRuler does not provide these specialized
regression methods at the moment.

3 CONCLUSION
CNVRuler is a user-friendly program with multiple functions that
support all procedures of CNV–phenotype association analysis in a
single system without requiring any additional manual processes.

Funding: This study was supported by a grant from the Korea
Healthcare Technology R&D Project (A0902285) and Korea Health
21 R&D Project (A040002), Ministry of Health and Welfare,
Republic of Korea.

Conflict of Interest: none declared.

REFERENCES
Bae J.S. et al. (2010) Genome-wide association analysis of copy number variations in

