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

Supplementary information: [Supplementary data are available at Bioinformatics online.]

Received on January 28, 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.

A number of tools supporting CNV-disease association analysis have been developed [Supplementary Table S1]. However, most of these tools do not offer methods for defining the CNVRs and require additional manual processes for converting the input CNV calls into files suitable for statistical analysis. In order to support the steps from merging CNVs into CNVRs to CNVR-based association analysis in a single software program, we developed a user-friendly tool for CNV-GWAS, called CNVRuler.

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 used directly 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
The third definition involves splitting the overlapping regions
Figure S2).
methods of CNVRuler were defined by CONAN (Supplementary
et al. 2010). Nearly 100% of the CNVRs defined by the three
defined CNVRs using CNVRuler and CONAN software (Forer
fragments, the calculation time is longer than that in other methods.
into fragments. Since this definition generates a larger number of
by 0.01 with Bae et al.’s significant CNVs. Two significant CNVs identified by Bae et al.
(copy number loss in 4q31.3 and copy number gain in 10p15.1) were detected by all three algorithms of CNVRuler. In association analyses,
for CNVRuler, we applied the data of 4574 CNVs identified in 500 cases of subarachnoid
aneurysmal hemorrhage using Illumina HumanHap300 BeadChip
et al. 2010). The three different CNVR-defining algorithms
overlap between any two CNV calls defined as the overlap of
be selected by the user. Second, common CNV regions can also
be removed by default. The density threshold for trimming can
by guest on February 1, 2016 http://bioinformatics.oxfordjournals.org/Downloaded from
Funding: This study was supported by a grant from the Korea Healthcare Technology R&D Project (A0902258) 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

2.4 Association analysis
CNVRuler supports chi-squared and Fisher’s exact tests in addition
to logistic and linear regression analyses using defined CNVRs and
clinical information. Clinical information can be easily coded into
a simple tab-delimited text file format. Both the false discovery rate and Bonferroni correction can be used for multiple testing
with this software. There is an option to remove the CNVRs from
the association analysis based on the frequency (see user manual).
CNVRuler supports the likelihood ratio test (LRT), which can be
used to assess the goodness-of-fit of logistic regression models.
For population stratification, CNVRuler uses principal component
analysis. It calculates eigenvectors and uses up to 3 principal
components as covariates of regression.
In order to validate the performance of CNVRuler, we applied
the data of 4574 CNVs identified in 500 cases of subarachnoid
aneurysmal hemorrhage using Illumina HumanHap300 BeadChip
et al. 2010). The three different CNVR-defining algorithms
identified different numbers of CNVRs: 1843 CNVRs, 2211 ROs
and 2797 fragments (Supplementary Table S3). We compared the
lists of our CNVRs with a raw P-value <0.01 with Bae et al.’s significant CNVs. Two significant CNVs identified by Bae et al.
(copy number loss in 4q31.3 and copy number gain in 10p15.1) were detected by all three algorithms of CNVRuler. In association analyses,
the two CNVRs were consistently significant in univariate
models regardless of the CNVR-defining algorithm (Supplementary
Table S4). However, our CNVRs were not significant in logistic
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
particularly 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.

Flowcharts of each building process are presented in Supplementary
Figure S1. First, individual CNVs are merged into CNVRs, which
are genomic regions covering CNVs overlapping by at least 1 bp
(Redon et al. 2006). This process is simple and straightforward,
but it may overestimate the size of CNVRs when any of the
overlapping CNVs are extremely long (see user manual). In order
to minimize this possibility, CNVRuler provides the option to
assess the regional density of the participating CNVs base-by-base
and trim the low-density areas. For example, the area covered
by <10% of the total contributing CNVs within a CNVR will
be removed by default. The density threshold for trimming can
be determined by reciprocal overlap (RO). RO is the degree of
overlap between any two CNV calls defined as the overlap of
one CNV with another over a predefined threshold value (Conrad
et al. 2010). In CNVRuler, we set the RO threshold to 0.5.
The third definition involves splitting the overlapping regions
into fragments. Since this definition generates a larger number of
fragments, the calculation time is longer than that in other methods.
To validate the CNVRuler, we used the CNVs identified from
Affymetrix SNP array 6.0 genotyping data of 10 individuals and
defined CNVRs using CNVRuler and CONAN software (Forer
et al. 2010). Nearly 100% of the CNVRs defined by the three
methods of CNVRuler were defined by CONAN
Supplementary
Figure S3).