BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 2 2013, pages 262–263
doi:10.1093/bioinformatics/bts650

Genome analysis

Advance Access publication November 4, 2012

HD-CNV: hotspot detector for copy number variants

Jenna L. Butler[1,*], Marjorie Elizabeth Osborne Locke[1], Kathleen A. Hill[1,2] and Mark Daley[1,2]

[1]Department of Computer Science, The University of Western Ontario, London, ON, Canada N6A 3K7 and [2]Department of Biology, The University of Western Ontario, London, ON, Canada N6A 5B7

Associate Editor: Alex Bateman

ABSTRACT

Summary: Copy number variants (CNVs) are a major source of genetic variation. Comparing CNVs between samples is important in elucidating their potential effects in a wide variety of biological contexts. HD-CNV (hotspot detector for copy number variants) is a tool for downstream analysis of previously identified CNV regions from multiple samples, and it detects recurrent regions by finding cliques in an interval graph generated from the input. It creates a unique graphical representation of the data, as well as summary spreadsheets and UCSC (University of California, Santa Cruz) Genome Browser track files. The interval graph, when viewed with other software or by automated graph analysis, is useful in identifying genomic regions of interest for further study.

Availability and implementation: HD-CNV is an open source Java code and is freely available, with tutorials and sample data from http://daleylab.org.

Contact: jcamer7@uwo.ca

Received on July 2, 2012; revised on October 23, 2012; accepted on October 29, 2012

1 INTRODUCTION

A copy number variant (CNV) refers to a large (≥1 kb) segment of the genome that is either duplicated or deleted. CNVs are found between individuals and between tissues of the same organism because of germline mutation and/or somatic mosaicism, and can significantly impact genetic variability, gene expression, phenotypic variability, disease, cancer, evolution and adaptation (Zhang *et al*., 2009). They may also be referred to as copy number abberations, changes, differences or polymorphisms depending on the discipline, context and the level of specificity. At the genome scale, CNVs can be detected using a variety of methods, including microarray and next-generation sequencing analysis. After detection, finding CNVs present in many samples (recurrent CNV regions) can indicate commonly inherited CNVs or mutation hotspots generating CNVs. Finding unique CNV events can indicate potential candidate mutations relevant to disease or other phenotypic variants. We introduce HD-CNV, which takes any collection of CNV calls as input, detects recurrent regions based on percentage overlap and creates a unique graphical visualization of clusters of overlapping CNV regions across samples. Other software exists to analyse overlapping CNV calls and find recurrent copy number variable regions among samples (Cazier *et al*., 2012; Forer *et al*., 2010; Kim *et al*., 2012; Subirana *et al*., 2011; Wittig *et al*., 2010). HD-CNV is unique, in that it analyses CNV events using interval graphs (Lekkerkerker *et al*., 1962) and gives a visual karyotype for quick identification of recurrent and unique CNVs across the genome.

*To whom correspondence should be addressed.

2 METHODS

HD-CNV imports pre-formatted CSV files containing previously detected CNV events. CNV events are treated as nodes in an interval graph. Interval graphs, in general, are used to represent regions (intervals) on a real line, and edges are added where intervals overlap (Lekkerkerker *et al*., 1962). Here, edges are added between nodes that share the user-specified base pair overlap required to consider two CNV events part of a 'merged region' (default 40%) (Fig. 1). Users also indicate the overlap required for a 'family' (region with highly similar CNV events, default 99%).

Once the graphs have been built, the Bron Kerbosch Clique Finder algorithm (Bron *et al*., 1973) returns all cliques (groups of nodes completely interconnected), which are reported as merged regions. Merged regions, therefore, contain a collection of CNV events, which each overlap all others in the merged region by the minimum overlap specified and indicate the genomic location in which that group of overlapping CNVs is found. They can be of any size from two to the number of detected CNVs depending on the input, and they will be detected by HD-CNV with 100% accuracy.

Human sample data were collected from Complete Genomics (Drmanac *et al*., 2010) publicly available diversity dataset, which includes CNV calls from their assembly software platform (version 2.0.0.26) for 46 individuals from diverse genetic backgrounds (including individuals from China, Japan, Kenya, Italy, USA and Nigeria). HD-CNV was applied to the data, and the graph files generated were visualized using Gephi (Fig. 2). Within each circular graph (one per chromosome), clusters (merged regions) are visible as circular substructures, whose size is proportional to the number of CNV events in the merged region. With 12 100 unique CNV events, HD-CNV runs in 62 s on an 2.8 GHz Intel Core i7 with 16 GB of random access memory.

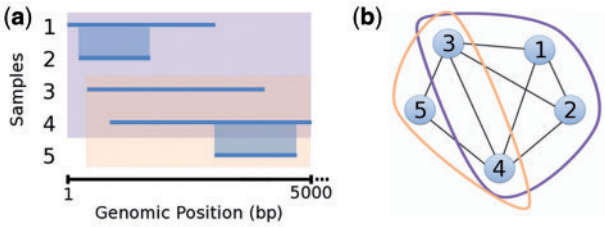

Fig. 1. (**a**) Called CNV events in five artificial samples. Two merged regions are indicated by light and dark shading. (**b**) Interval graph. Cliques are circled with shading corresponding to merged regions shown in (a)

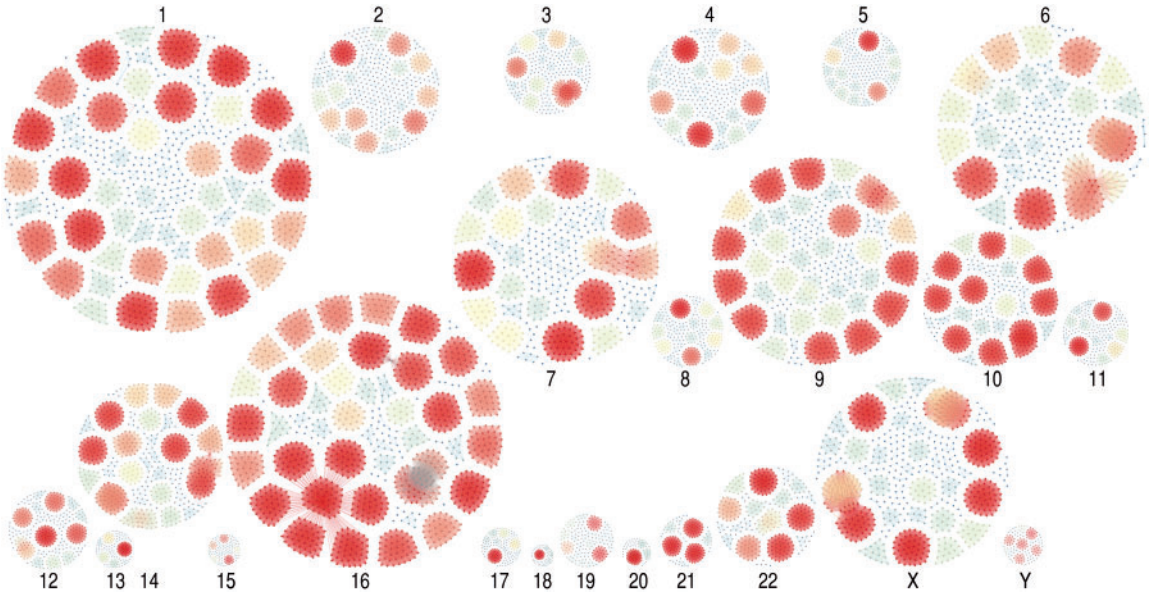

Fig. 2. Karyotype of CNV events in human chromosomes numbered 1–22, X and Y, as visualized from HD-CNV generated graph files. Graph: each graph represents events for one chromosome. The size of the graph is relative to the total events observed on that chromosome. Node: a dot indicates a CNV event (amplification or deletion), and all nodes are the same size. Edges: lines connect CNV events whose genomic regions overlap by at least 40%. Colour: dark indicates events with a high number of overlap with other events, while light indicates some level of overlap and medium is almost no overlap

3 APPLICATION

HD-CNV identifies recurrent CNV regions and allows for high-level visualization of the relative number of CNV events on each chromosome, clusters of overlapping CNV events and isolated CNV events. It can be used for qualitative exploratory analysis on data from any organism, and it does not require additional reference data or annotation. Statistical analysis can be done up- or downstream as appropriate for the dataset. It can be used to compare CNV calls from multiple programs for concordance or discordance as in, for example, tumour/normal sample comparisons. It has been tested on data from different species (human and mouse) and data from different CNV calling platforms. HD-CNV generates various output files, including:

'Graph files' can be used for large scale visualization, allowing quick identification of areas for further study. CNV events, or merged regions from the graph, can be found in the track files.

'UCSC Genome Browser track files' contain the merged regions and the original CNV events, including base pair locations, which allows cross referencing with other genomic data and features.

4 RESULTS

We have used our program to analyse CNVs detected in humans from a diverse genetic background (Fig. 2). The darker clusters indicate a CNV event that is recurrent in all samples, which may indicate CNVs common among many human populations. Individual nodes, and smaller clusters, show rare CNV events that may be limited to certain individuals and sub-populations. This genome-scale visualization gives a summary of copy number variability in this large sample set on each chromosome.

As an increasing number of investigators are able to detect CNVs in large numbers of samples, HD-CNV will facilitate analysis and visualization of the data and direct researchers to regions of interest for verification and further biological study.

Funding: Natural Sciences and Engineering Research Council (NSERC) of Canada.

Conflict of Interest: none declared.

REFERENCES

Bastian,M. *et al.* (2009) Gephi: an open source software exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*. AAAI Publications, San Jose, California.

Bron,C. *et al.* (1973) Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.

Cazier,J. *et al.* (2012) GREVE: Genomic Recurrent Event ViEwer to assist the identification of patterns across individual cancer samples. *Bioinformatics*, **28**, 2981–2982.

Drmanac,R. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.

Forer,L. *et al.* (2010) CONAN: copy number variation analysis software for genome-wide association studies. *BMC Bioinformatics*, **11**, 318.

Kim,J. *et al.* (2012) CNVRuler: a copy number variation-based case-control association analysis tool. *Bioinformatics*, **28**, 1790–1792.

Lekkerkerker,C.G. *et al.* (1962) Representation of a file graph by a set of intervals on the real line. *Fund. Math.*, **51**, 45–64.

Subirana,I. *et al.* (2011) CNVassoc: association analysis of CNV data using R. *BMC Med. Genomics*, **4**, 47.

Wittig,M. *et al.* (2010) CNVineta: a data mining tool for large case-control copy number variation datasets. *Bioinformatics*, **26**, 17.

Zhang,F. *et al.* (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet*, **10**, 451–481.