Genomics & Informatics Vol. 6(3) 126-129, September 2008

CGHscape: A Software Framework for the Detection and Visualization of Copy Number Alterations

Yong-Bok Jeong[1,2], Tae-Min Kim[1,2] and Yeun-Jun Chung[1,2]*

[1]Integrated Research Center for Genome Polymorphism, [2]Department of Microbiology, The Catholic University of Korea College of Medicine, Seoul 137-701, Korea

Abstract

The robust identification and comprehensive profiling of copy number alterations (CNAs) is highly challenging. The amount of data obtained from high-throughput technologies such as array-based comparative genomic hybridization is often too large and it is required to develop a comprehensive and versatile tool for the detection and visualization of CNAs in a genome-wide scale. With this respective, we introduce a software framework, CGHscape that was originally developed to explore the CNAs for the study of copy number variation (CNV) or tumor biology. As a standalone program, CGHscape can be easily installed and run in Microsoft Windows platform. With a user-friendly interface, CGHscape provides a method for data smoothing to cope with the intrinsic noise of array data and CNA detection based on SW-ARRAY algorithm. The analysis results can be demonstrated as $\log_2$ plots for individual chromosomes or genomic distribution of identified CNAs. With extended applicability, CGHscape can be used for the initial screening and visualization of CNAs facilitating the cataloguing and characterizing chromosomal alterations of a cohort of samples.

Availability: CGHscape installation package with online manual is freely available in our website, http://www.ircgp.com/software/CGHscape.

Keywords: array-CGH, copy number alteration (CNA), copy number variation (CNV)

Introduction

The advancement of high-throughput technologies such as microarray-based platforms has opened unprecedented challenges in a biological field. In case of genomic analysis, array-based comparative genomic hybridization (array-CGH) has enabled the detection of copy number alterations (CNAs) in a high resolution and has been used for the analysis of cancer or congenital disorders (Pinkel *et al*., 2005; Yim *et al*., 2004). It has also facilitated the discovery of large-scaled structural variations including copy number variation (CNV) that comprises a substantial amount of genomes in normal individuals (Freeman *et al*., 2006). In spite of the technological advancement, the identification of CNAs is often not straightforward largely due to the intrinsic noise of array-based dataset. A number of algorithms have been proposed to detect the CNAs (Lai *et al*., 2005), however, a majority of them require considerable experiences on handling of large-scale data or specific knowledge (i.e., R-package). In addition, the use of vendor-provided software is often limited to certain types of used array platforms limiting the general use. Thus, it is highly challenging to develop software that can be used with relative ease and those equipped with versatile methods that can be applied for common array-CGH platforms to ensure the extended compatibility.

Here, we propose a software framework designed for the identification and visual representation of CNAs using genome-wide array-CGH profiles. CNAs can be directly identified from log2 ratio profiles that can be obtained from array-CGH datasets with minimal modifications. Data smoothing option is also provided to cope with the noise level of data for reliable detection of CNAs. The identification of CNAs is based on SW-ARRAY algorithm that ensures fast and robust detection of chromosomal alterations. The identified CNAs are exported into Excel-compatible outputs or graphically illustrated with graphic-user interface. Relatively easy operability as well as the fast processing of overall procedures is the major advantage of our software over the conventional ones. CGHscape software package is freely available and provides the comprehensive environments for investigation of tumor genome and genomic variants.

Major Functionalities of CGHscape

(1) CGHscape was designed as a standalone program compatible in Microsoft Windows environments. Compiled codes of CGHscape can be easily installed. The interpreter- or web-based methods have the advantage

*Corresponding author: E-mail yejun@catholic.ac.kr
Tel +82-2-590-1214, Fax +82-2-596-8969
Accepted 7 September 2008

Table 1. Comparison with copy number analysis softwares

Feature	Available for Platforms	Requirements	CNA Detection algorithm	Scatter plot export	CNA region view	Run time	Data smoothing	Web site	Reference
CGHscape	Windows	Standalone	SW-Array	Yes	Yes	Fast (<1 sec)	Moving average, Gaussian	www.ircgp.com/software/CGHscape	
aCGH Smooth	Windows	Excel	Heuristic algorithm, regularized maximum likelihood, Threshold	Yes	Yes	Slow (4 min 40 sec)	User-defined	www.few.vu.nl/~vumarray	(Jong *et al.*, 2004)
CGH Analyzer	Windows, Unix, Mac	JRE	T-test (multiple test adjustment)	Yes	No	Fast (<1 sec)	N/A	www.genomics.upenn.edu/people/faculty/weberb/CGH/html/software.htm	(Margolin *et al.*, 2005)
ChARM View	Windows, Linux, Mac	JRE	EM (Expectation Maximization	Yes	Yes	Moderate (27 sec)	N/A	function.princeton.edu/ChARM/	(Myers *et al.*, 2004)
CGH-Explorer	Windows, Linux, Mac	JRE	ACE (Analysis of copy errors)	Yes	Yes	Fast (<1 sec)	Moving average	www.ifi.uio.no/forskning/grupper/bioinf/Papers/CGH/	(Lingjaerde *et al.*, 2005)

Sample data: Copy number analysis data using Phalanx Human OneArray(32K) chip.

of being independent of platforms, however, they often sacrifice the overall performance. As compared with previous softwares with similar purposes, CGHscape works relatively fast and it is one of the major advantages of our software (Table 1).

(2) To ensure the versatility of the methods, $\log_2$ ratio profile can be directly uploaded and analyzed in CGHscape. Currently available softwares for copy number detection are often limited to the use of vendor-specific array platforms or data types. In addition, publicly available algorithms from non-profit investigators often require extensive experiences or professional knowledge for large-scaled data handling. The log2 ratio profiles can be relatively easily obtained from array-CGH datasets ensuring the compatibility regardless of the array or data types and data flexibility.

(3) The inherent noise in microarray data must be considered in subsequent analysis and this is especially the case of oligonucleotide-based array CGH (Ylstra *et al.*, 2006). In case of genomic log2 profiles, the probes can be sorted according to the genomic coordinates and data smoothing can be applied to reduce the noise levels. For the robust detection of CNAs, CGHscape implements the Gaussian smoothing algorithm as well as the moving median or moving average options for data smoothing.

(4) A number of algorithms have been proposed for the identification of CNAs while each of them has own merits and drawbacks. SW-ARRAY algorithm that adopts Smith-Waterman algorithm in scanning and detecting CNAs, has been recently proposed and notable for its high sensitivity and robustness (Price *et al.*, 2005). The relatively fast performance of the algorithm is also one of its advantages and has been used in large-scale screening of genomic variants (Komura *et al.*, 2006). CGHscape adopts this algorithm for the detection of CNAs also providing a number of options to adjust the parameters such as threshold levels, permutation levels and cutoff of scores.

(5) Graphic user interface is implemented in the software for the users to easily handle large-scale data or access the analysis results (Fig. 1). The CNVs identified can be also visually demonstrated as individual chromosomal plots or genome-wide distribution map of scattered CNAs. The CNA regions can be also either exported into tab-delimited plain text or Excel-compatible format for further investigation.

In conclusion, CGHscape can be a useful tool with multi-functionalities containing the initial screening, smoothing, detecting and visualization of CNAs which facilitate cataloguing and characterizing chromosomal alterations of disease samples.

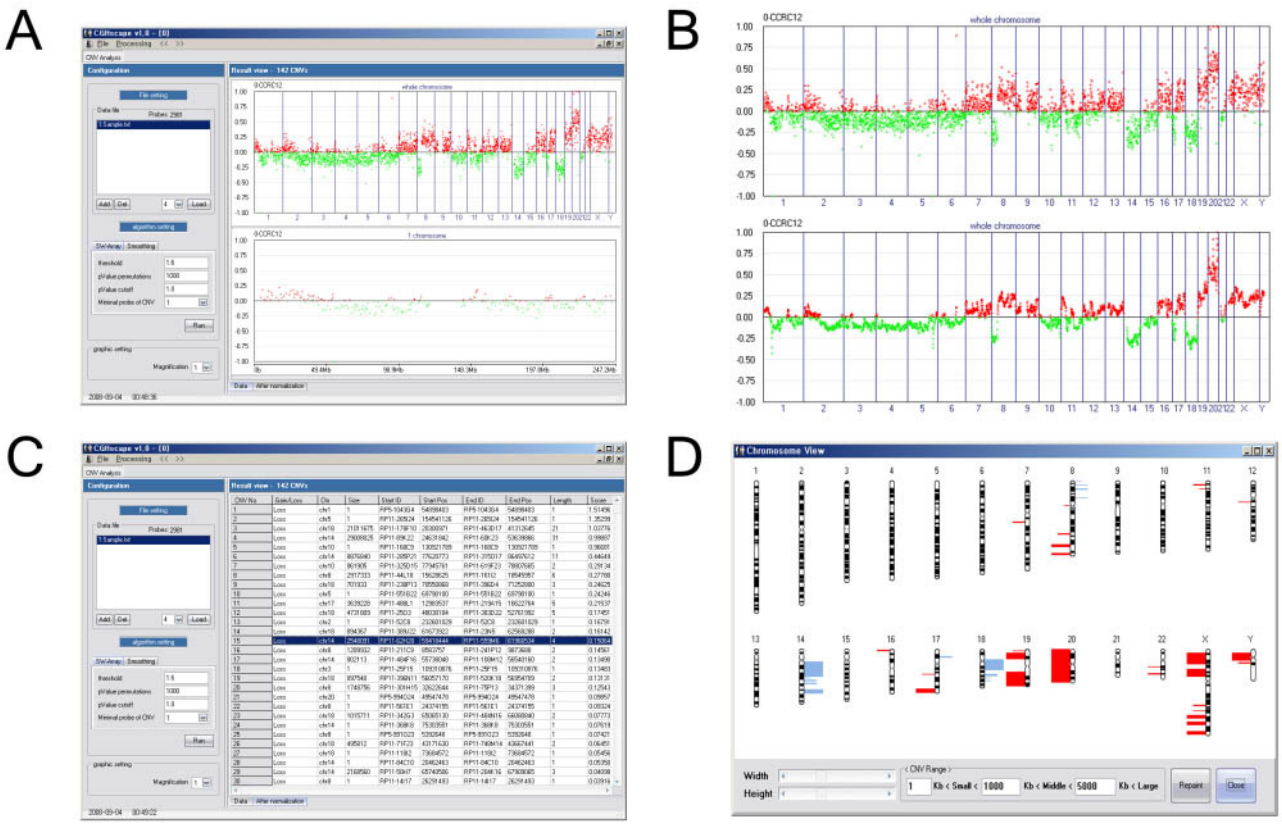

Fig. 1. Screenshots of CGHscape. (A) The main frame of CGHscape for data uploading and visualization is shown. The data uploading and the parameter setting of the subsequent analyses can be done with graphic user interface. The genome-wide $\log_2$ ratio plots are shown in order of $1p_{ter}$ to Yq_{ter} for all chromosomes demonstrating the chromosomal gains and losses in red and green, respectively. The selected chromosome can be further shown as individual chromosomal plot (below). (B) Log_2 ratio can be processed for Gaussian smoothing to reduce the noise level. Above (before Gaussian smoothing) and below (after) $\log_2$ plots clearly shows the benefits of data smoothing. (C) The CNAs identified using SW-ARRAY algorithm can be listed in table formats. The list can be exported into tab-delimited plain text or Microsoft Excel format. (D) The genome-wide distribution of identified CNAs is illustrated with respect to individual chromosomes. Left (red) and right (green) bars indicate the relative genomic gains and losses, respectively and the length of bars indicate the length of CNAs.

Acknowledgement

This work was supported by the FG06-12-01 of the 21C Frontier Functional Human Genome Project in Korea and YB Jeong was supported from Korea Research Foundation Grant (MOEHRD, Basic Research Promotion Fund) (KRF-2007-511-C00051).

References

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., Carter, N.P., Scherer, S.W., and Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Res.* 16, 949-961.

Komura, D., Shen, F., Ishikawa, S., Fitch, K.R., Chen, W., Zhang, J., Liu, G., Ihara, S., Nakamura, H., Hurles, M.E., Lee, C., Scherer, S.W., Jones, K.W., Shapero, M.H., Huang, J., and Aburatani, H. (2006). Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* 16, 1575-1584.

Jong, K., Marchiori, E., Meijer, G., Vaart, A.V., and Ylstra, B. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics* 20, 3636-3637.

Lai, W.R., Johnson, M.D., Kucherlapati, R., and Park, P.J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21, 3763-3770.

Lingjaerde, O.C., Baumbusch, L.O., Liestol, K., Glad, I.K., and Borresen-Dale, A.L. (2005). CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* 21, 821-822.

Margolin, A.A., Greshock, J., Naylor, T.L., Mosse, Y., Maris,

J.M., Bignell, G., Saeed, A.I., Quackenbush, J., and Weber, B.L. (2005). CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics* 21, 3308-3311.

Myers, C.L., Dunham, M.J., Kung, S.Y., and Troyanskaya, O.G. (2004). Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics* 20, 3533-3543.

Pinkel, D., and Albertson, D.G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* 37, Suppl, S11-S17.

Price, T.S., Regan, R., Mott, R., Hedman, A., Honey, B., Daniels, R.J., Smith, L., Greenfield, A., Tiganescu, A., Buckle, V., Ventress, N., Ayyub, H., Salhan, A., Pedraza-Diaz, S., Broxholme, J., Ragoussis, J., Higgs, D.R., Flint, J., and Knight, S.J. (2005). SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res.* 33, 3455-3464.

Yim, S.H., and Chung, Y.J. (2004). Current status and future clinical applications of Array based Comparative Genomic Hybridization. *Genomics & Informatics* 2, 113-120.

Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R.H., and Meijer, G.A. (2006). BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*. 34, 445-450.