Bioinformatics, 34(3), 2018, 519–521
doi: 10.1093/bioinformatics/btx620
Advance Access Publication Date: 3 October 2017
Applications Note

OXFORD

Downloaded from https://academic.oup.com/bioinformatics/article/34/3/519/4265459 by Gwang Ju Institute of Science & Technology user on 06 September 2024

Genome analysis

Identification of cancer driver genes in focal genomic aberrations from whole-exome sequencing data

Ho Jang and Hyunju Lee*

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on July 5, 2017; revised on September 15, 2017; editorial decision on September 22, 2017; accepted on September 27, 2017

Abstract

Summary: Whole-exome sequencing (WES) data have been used for identifying copy number aberrations in cancer cells. Nonetheless, the use of WES is still challenging for identification of focal aberrant regions in multiple samples that may contain cancer driver genes. In this study, we developed a wavelet-based method for identifying focal genomic aberrant regions in the WES data from cancer cells (WIFA-X). When we applied WIFA-X to glioblastoma multiforme and lung adenocarcinoma datasets, WIFA-X outperformed other approaches on identifying cancer driver genes.
Availability and implementation: R source code is available at http://gcancer.org/wifax.
Contact: hyunjulee@gist.ac.kr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the availability of high-throughput sequencing data, we can detect copy number aberrations (CNAs) in cancers more precisely. Because it is important to identify focal aberrant regions that occur repeatedly across multiple patients with cancer and may contain cancer driver genes, several tools have been developed for single-nucleotide polymorphism (SNP) array data and whole-genome sequencing (WGS) data including GISTIC and our methods WIFA and WIFA-Seq (Beroukhim *et al.*, 2007; Hur and Lee, 2011; Jang *et al.*, 2016). Nevertheless, there are few computational tools that can be applied to whole-exome sequencing (WES) data although WES is more frequently used than WGS because it is less expensive than WGS and many studies usually focus on protein-coding regions. Thus, in the present study, we developed a wavelet-based method for identifying focal genomic aberrant regions in the WES data from cancer cells (WIFA-X) and applied our method to the WES data on glioblastoma multiforme (GBM) and lung adenocarcinomas (LUAD) from The Cancer Genome Atlas (https://tcga-data.nci.nih.gov) [Authorization was obtained from the database of Genotypes and Phenotypes (accession No. phs000178.v8.p7)]. We found many GBM and LUAD driver genes. Our method can be widely used for identifying cancer driver genes in WES datasets.

2 Materials and methods

Figure 1a shows an outline of the WIFA-X method. First, for a pair of tumorous and normal WES data, somatic copy number changes across genomic regions are quantified. Although many investigators (Amarasinghe *et al.*, 2013) consider only exon regions for quantifying copy numbers, a recent study (D'Aurizio *et al.*, 2016) took into account both exon regions and off-target regions as the markers to increase the detection power of CNAs. WIFA-X can take as input either BAM files or log2 ratios of copy numbers for both exon and off-target regions (Supplementary Figs S1–S2).

Next, focal aberrations in a single cancer sample are detected using Haar translation invariant discrete wavelet transform that is the same as in our existing methods WIFA and WIFA-Seq (Hur and Lee, 2011; Jang *et al.*, 2016). At this step, noise is removed from the raw data using hard thresholding of wavelet coefficients, and focal aberrations are calculated by reconstructing signals without scaling

© The Author 2017. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

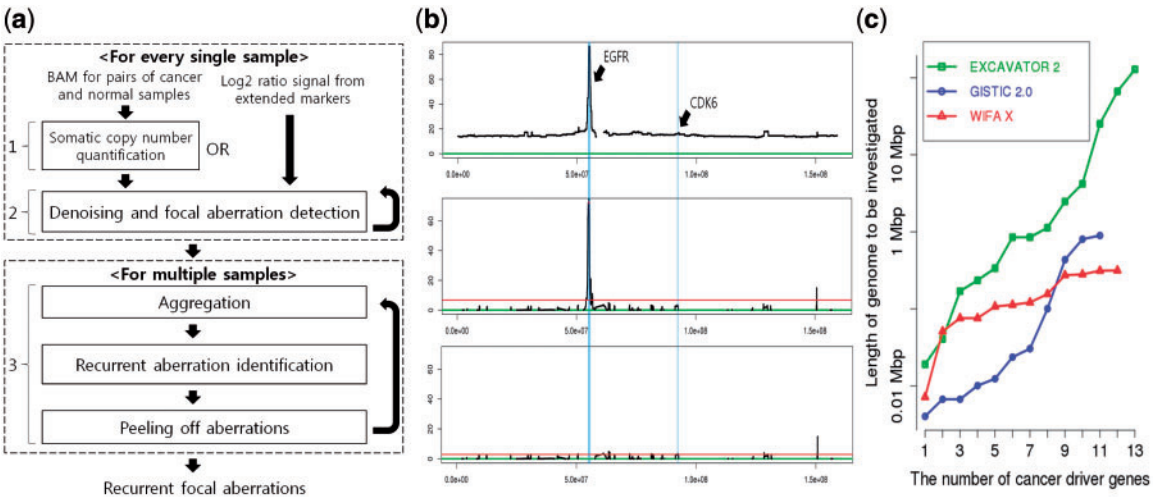

Fig. 1. (**a**) The procedure for the WIFA-X method. (**b**) (top) The aggregated profile of chromosome 7 from 35 WES GBM samples using EXCAVATOR2 segments is shown. (middle) $\bar{h}_j^{amp}$ signals from the same dataset are shown. (b) (bottom) $\bar{h}_j^{amp}$ signals after the peel-off step are shown. Green horizontal lines indicate zero log2 ratio values. Red horizontal lines indicate thresholds for identifying recurrent regions. (**c**) Performance comparison among the manual inspection of EXCAVATOR2 segments, GISTIC 2.0 and WIFA-X is presented. The *y*-axis represents the length of exon regions necessary to identify cancer driver genes on the *x*-axis

coefficients. The size of aberration is controlled by wavelet transform levels for wavelet coefficients. Because aberrant regions whose lengths are less than 25% of the chromosome arm are usually considered focal aberrations (Koboldt *et al.*, 2012), we assign the wavelet transform levels to ensure that the length of identified aberrant regions are less than 25% of the chromosome arm. In the case of WES data, some locations may have abnormally high or low copy number values owing to the sequencing or mappability bias. Thus, we control these abnormalities by repetitively applying wavelet transform to log2 ratio copy number, which generates y_{HIGH} that is used for locating abnormalities and correcting abnormal values by considering their neighbouring markers. The final focal aberration signal is named y^*_{HIGH}. Signal y^*_{HIGH} is produced for every single WES cancer sample (Supplementary Figs S3–S4).

Finally, WIFA-X identifies recurrent aberrations in multiple samples. Recurrently amplified regions and recurrently deleted regions are separately identified. For identifying recurrently amplified regions, an aggregated profile $\bar{h}_j^{amp} = \sum_{i=1}^{M} h_{ij}^{amp}$ is calculated, where $h_{ij}^{amp} = h_{ij} \times I(h_{ij} > 0)$, h_{ij} is y^*_{HIGH} value of marker j for a patient i, and M is the number of patients. Figure 1b (middle) shows an example of aggregated profile $\bar{h}_j^{amp}$. Then, WIFA-X conducts a statistical test based on a cyclic permutation test. The P-value of each location marker j in the aggregated profile $\bar{h}_j^{amp}$ is calculated as $p\text{-value}_j = \frac{\sum_{k=1}^{N} I\left(quantile(k,0.99) > \bar{h}_j^{amp}\right)}{N}$, where $quantile(k, 0.99)$ is 0.99-quantile of the kth randomly aggregated profile, and N denotes the total number of cyclic permutations. The randomly aggregated profiles are calculated by cyclically shifting markers in h_{ij}^{amp} for each individual patient i independently and aggregating them into a single profile. In the $\bar{h}_j^{amp}$ signal, we select consecutively significant markers with $P\text{-value} < P\text{-value}_{thres}$ and regard these regions as the recurrent aberrations. To identify further recurrent focal aberrations, we adapted the peel-off step used in other methods (Beroukhim *et al.*, 2007; Walter *et al.*, 2011). WIFA-X removes focal aberrations overlapping with the identified recurrent aberration, re-estimates the null distribution based on the remaining focal aberrations, and finds a new recurrent region based on the new null distribution. This procedure continues until WIFA-X cannot identify any further significantly recurrent regions. Figure 1b (middle) shows identification of *EGFR* regions and Figure 1b (bottom) shows identification of *CDK6* after the peel-off step for chromosome 7 from 35 GBM samples. For comparison, we used a simple approach [Figure 1b (top)], where each single sample in the same GBM samples is segmented using EXCAVATOR2 (D'Aurizio *et al.*, 2016), and segmentation data from the 35 samples are summated. The comparison between Figure 1b (top) and (middle) shows that the recurrent aberrations are better distinguished by WIFA-X than by the simple approach. For identifying recurrently deleted regions, $h_{ij}^{del} = -h_{ij} \times I(h_{ij} < 0)$ and $\bar{h}_j^{del} = \sum_{i=1}^{M} h_{ij}^{del}$ are calculated, and $\bar{h}_j^{del}$ is used instead of $\bar{h}_j^{amp}$. See Supplementary Material and Supplementary Figures S1–S6 for more details about WIFA-X.

3 Results

We applied WIFA-X to 35 pairs of tumorous and normal WES data for GBM, 27 pairs of WES data for LUAD, and another 293 pairs of WES data for GBM, where the 35 GBM and 27 LUAD datasets have matching WGS data. An exome capture kit, Agilent SureSelect V2 (931070), was used to produce BAM files from these datasets. The total number of exons provided by this kit for 22 chromosomes is 182 568. For evaluating the performance of WIFA-X [Fig. 1c], we used normalized log2 ratios for exon and off-target regions obtained by the EXCAVATOR2 method. Because EXCAVATOR2 exploits CNAs by considering off-target regions together, we can use additional 33 663 markers. In WIFA-X, copy number differences between neighboring genomic regions up to 3 megabases were considered for identifying focal aberrations (Supplementary Material; Supplementary Tables S1 and S2).

We identified recurrently amplified or deleted aberrations in 22 autosomes and compared the performance among GISTIC 2.0, the manual inspection of segments detected by EXCAVATOR2 and WIFA-X. Although GISTIC 2.0 was originally developed for SNP array data, it can be used for WES data as well if markers having copy numbers across genomic regions are provided as input. Here, segmented log2 ratios from EXCAVATOR2 were used as input data

Downloaded from https://academic.oup.com/bioinformatics/article/34/3/519/4265459 by Gwang Ju Institute of Science & Technology user on 06 September 2024

for GISTIC 2.0. For the manual inspection, segments containing CNAs were detected by EXCAVATOR2 for each sample, and if these segments were detected in more than one sample, we considered them recurrent regions. For performance evaluation, we used 13 previously known cancer driver genes as silver standard genes, which were collected from the GBM WGS data from our previous study (Jang *et al.*, 2016). We compared genomic lengths required to identify the silver standard genes by sorting aberrant regions in descending order of absolute scores of recurrent regions for WIFA-X, in ascending order of *q*-values of the peaks for GISTIC 2.0, and in descending order of absolute log2 values of copy number segments in EXCAVATOR2.

Figure 1c shows that WIFA-X can identify more known GBM genes at lesser inspection length than GISTIC 2.0 can in the 35 GBM WES dataset, suggesting a higher coverage of WIFA-X with a lower false positive rate than GISTIC 2.0. Both methods identified seven driver genes including *EGFR*, *CDK4*, *MDM4*, *MDM2*, *PDGFRA*, *CCND2* and *CDK6* in the recurrently amplified regions and four driver genes including *CDKN2A/B*, *QKI* and *PTEN* in the recurrently deleted regions, while WIFA-X identified one more gene *FGFR3* in the amplified region (Supplementary Tables S3–S6; Supplementary Figs S7 and S8).

WIFA-X consistently identified more cancer genes at lesser inspection lengths than GISTIC 2.0 did for the 27 WES LUAD dataset and the 293 WES GBM dataset (Supplementary Tables S7–S14 and Supplementary Figs S9–S12). In the case of identifying either amplified regions or deleted regions only, the identification performance of WIFA-X is better than the performance of GISTIC 2.0 for all the datasets (Supplementary Tables S15–S17; Supplementary Figs S13 and S14). In addition, when we compared performance between the use of both exon and off-target regions and the use of only exon regions, we found that by means of both exon and off-target regions, we can identify more silver standard genes (Supplementary Tables S18–S20; Supplementary Figs S15 and S16).

Funding

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (The Ministry of Science, ICT and Future Planning) (NRF-2016R1A2B2013855).

Conflict of Interest: none declared.

References

Amarasinghe,K.C. *et al*. (2013) Convex: copy number variation estimation in exome sequencing data using hmm. *BMC Bioinformatics*, **14**, (2), S2.

Beroukhim,R. *et al*. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA*, **104**, 20007–20012.

D'aurizio,R. *et al*. (2016) Enhanced copy number variants detection from whole-exome sequencing data using excavator2. *Nucleic Acids Res*., **44**, e154–e154.

Hur,Y. and Lee,H. (2011) Wavelet-based identification of DNA focal genomic aberrations from single nucleotide polymorphism arrays. *BMC Bioinformatics*, **12**, 146.

Jang,H. *et al*. (2016) Identification of cancer-driver genes in focal genomic alterations from whole genome sequencing data. *Scientific Reports*, **6**, 25582.

Koboldt,D.C. *et al*. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*., **22**, 568–576.

Walter,V. *et al*. (2011) Dinamic: a method to identify recurrent dna copy number aberrations in tumors. *Bioinformatics*, **27**, 678–685.

Downloaded from https://academic.oup.com/bioinformatics/article/34/3/519/4265459 by Gwang Ju Institute of Science & Technology user on 06 September 2024