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

Ho Jang 1 and Hyunju Lee 1 *

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

Received on XXXX; revised on XXXX; accepted on XXXX

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: R source code is available at http://gcancer.org/wifax.

Contact: hyunjulee@gist.ac.kr

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) 1 . We found many GBM and LUAD driver genes. Our method can be widely used for identifying cancer driver genes in WES datasets.

* To whom correspondence should be addressed.

1 Authorization was obtained from the database of Genotypes and Phenotypes (accession No. phs000178.v8.p7)
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, in ascending order of absolute log2 values of copy number aberrations, 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 1 (c) 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 cancer 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 (Tables S3–S6; Figures S13 and S14).

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 (Tables S7–S14 and Figures 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 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 1 (c) 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 cancer 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 (Tables S3–S6; Figures 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 (Tables S7–S14 and Figures 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 (Tables S15–S17; Figures 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 (Tables S18–S20; Figures S15 and S16).

FUNDING
This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2016R1A2B2013855).

REFERENCES