BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 3 2014, pages 431–433
doi:10.1093/bioinformatics/btt693

Genome analysis

Advance Access publication November 29, 2013

AISAIC: a software suite for accurate identification of significant aberrations in cancers

Bai Zhang[1,2,†], Xuchu Hou[1,†], Xiguo Yuan[3,†], Ie-Ming Shih[2,4,5], Zhen Zhang[2], Robert Clarke[6,7,8], Roger R. Wang[9], Yi Fu[1], Subha Madhavan[6], Yue Wang[1] and Guoqiang Yu[1,*]

[1]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA, [2]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21231, USA, [3]School of Computer Science and Technology, Xidian University, Xi'an 710126, China, [4]Department of Oncology and [5]Department of Gynecology/Obstetrics, Johns Hopkins University School of Medicine, Baltimore, MD 21231, USA, [6]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20007, USA [7]Department of Oncology and [8]Department of Physiology and Biophysics, Georgetown University, Washington, DC 20057, USA and [9]Department of Electrical Engineering and Computer Science, University of Michigan, An Arbor, MI 48109, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Accurate identification of significant aberrations in cancers (AISAIC) is a systematic effort to discover potential cancer-driving genes such as oncogenes and tumor suppressors. Two major confounding factors against this goal are the normal cell contamination and random background aberrations in tumor samples. We describe a Java AISAIC package that provides comprehensive analytic functions and graphic user interface for integrating two statistically principled *in silico* approaches to address the aforementioned challenges in DNA copy number analyses. In addition, the package provides a command-line interface for users with scripting and programming needs to incorporate or extend AISAIC to their customized analysis pipelines. This open-source multiplatform software offers several attractive features: (i) it implements a user friendly complete pipeline from processing raw data to reporting analytic results; (ii) it detects deletion types directly from copy number signals using a Bayes hypothesis test; (iii) it estimates the fraction of normal contamination for each sample; (iv) it produces unbiased null distribution of random background alterations by iterative aberration-exclusive permutations; and (v) it identifies significant consensus regions and the percentage of homozygous/hemizygous deletions across multiple samples. AISAIC also provides users with a parallel computing option to leverage ubiquitous multicore machines.

Availability and implementation: AISAIC is available as a Java application, with a user's guide and source code, at https://code.google.com/p/aisaic/.

Contact: yug@vt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 18, 2013; revised on October 26, 2013; accepted on November 21, 2013

1 INTRODUCTION

Somatic DNA copy number alterations (CNAs) are important structural variants present in almost all human cancers. Oligonucleotide-based single nucleotide polymorphism (SNP) arrays provide a cost-effective technique to acquire high-density and allelic-specific copy number data for large-scale cancer studies. The Cancer Genome Atlas (TCGA) project represents an integrated attempt to acquire multiplatform genomic and molecular profiles on 20–25 cancer types each with hundreds of samples, where the Affymetrix Genome-Wide Human SNP Array 6.0 serves as the common platform for acquiring SNP and CNA data on these cancer types (Beroukhim *et al.*, 2010; Carter *et al.*, 2012).

Accurate identification of significant aberrations in cancers (AISAIC) is a systematic effort to identify potential cancer-driving genes by locating significant consensus CNA regions and detecting deletion types. Two major confounding factors against this goal are the normal cell contamination and random background aberrations in tumor samples. Moreover, simple permutation of CNAs would produce a biased null distribution of random background aberrations due to the contamination by the unknown yet true consensus CNAs. To address these challenges, we developed AISAIC by integrating and streamlining two statistically principled approaches: Bayesian Analysis of COpy number Mixtures (BACOM) for detecting deletion types and estimating normal cell proportions directly from copy number data (Yu *et al.*, 2011), and Significant Aberration in Cancer (SAIC) for identifying significant consensus CNA regions (Yuan *et al.*, 2012). In addition to the unique functionalities summarized in the abstract, AISAIC further provides the following unique new features: (i) AISAIC is a self-contained DNA copy number analysis tool implementing the entire analysis pipeline, from preprocessing raw data to reporting significant consensus aberrations (SCAs) results with gene annotations and visualization, and it provides a user-friendly 'one-click' solution without dependence on other software packages;

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

(ii) AISAIC naturally integrates BACOM and SAIC and significantly enhances the accuracy of identifying SCAs when compared with competing methods (even when compared with SAIC alone); (iii) AISAIC adopts concurrent computing techniques to take advantage of multicore/multiprocessor architecture found in most modern computers, which greatly reduces computation time, especially when handling large datasets.

The AISAIC package was tested using both simulation and real datasets, and the experimental results show that the integrated application of BACOM and SAIC to clinical samples can significantly improve the power to detect significant consensus CNA regions and produce biologically plausible findings warranting further studies.

Here, we describe a Java package that provides comprehensive analytical functions, a graphical user interface (GUI) and result visualization for disseminating the unified software of BACOM-SAIC algorithms to the cancer research community and beyond. In addition, this open-source and multiplatform software also provides a command-line interface for users with scripting and programming needs to incorporate or extend AISAIC to their customized analysis pipelines, and offers the users with a parallel computing option using ubiquitous multicore machines and concurrent fork/join framework in Java SE 7. We apply the AISAIC package to the glioblastoma multiforme (GBM) and lung adenocarcinoma (LUAD) datasets in TCGA and report various analysis results.

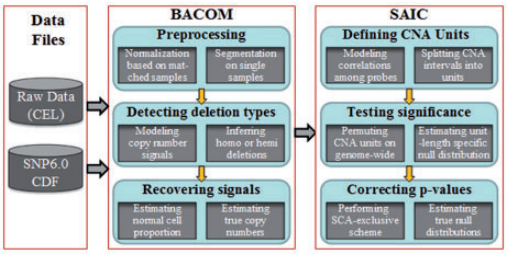

Fig. 1. Architecture and design of AISAIC software suite

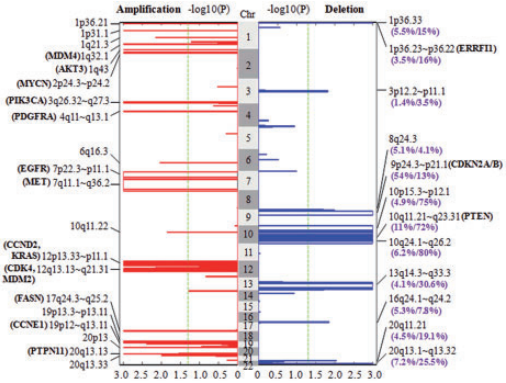

Fig. 2. Illustrative outcomes of the AISAIC analysis on the TCGA GBM case study (see Supplementary Material for more detailed discussions)

2 DESCRIPTION

2.1 Methods and software

BACOM first detects deletion segments using established allelic-specific analysis tools comprising genotype calling, signal normalization and deletion loci detection. A Bayes classifier is then used to identify each segment's deletion type (homozygous versus heterozygous deletions) directly from copy number signals. Finally, BACOM estimates both the fraction of normal cells in a tumor sample and the cancer-specific copy number profile. As a completely unsupervised approach, BACOM allows parameters of the underlying deletion-type conditioned probability models to be readily estimated from measured copy number signals, without any knowledge of the associated deletion type. First, SAIC defines CNA units by exploiting intrinsic correlation among consecutive probes and then assigns a score to each CNA unit instead of single probes based on both the amplitude and frequency of the signals within the unit. Next, to estimate the unbiased null distribution of random background aberrations, SAIC performs permutations on CNA units that iteratively exclude detected aberrations yet preserve correlations inherent in the copy number data. Finally, SAIC identifies significant consensus CNA regions based on adjusted *P*-values.

The AISAIC package is implemented entirely as a Java application. The illustrative flowchart given in Figure 1 provides an overview of the software architecture design. The AISAIC software contains two major analytic modules (BACOM and SAIC). Through a user-friendly graphical user interface, users can easily import data, assign proper parameter values and select their preferred analysis routes via the 'BACOM', 'SAIC' or 'BACOM+SAIC' options. AISAIC then takes raw data CEL files (e.g. Affymetrix SNP arrays) as its input and outputs various intermediate and exportable results to users, including normal cell fractions, deletion types (hemizygous/homozygous) and their percentages, significant consensus CNA regions and embedded gene lists.

The AISAIC package offers users parallel computing capabilities, taking full advantage of ubiquitous multicore CPUs and multiprocessor machines. To enable convenient concurrent computing, AISAIC adopts the widely embraced concurrent Fork/Join framework recently introduced in Java SE 7 and implements a hierarchy of multithreads. Specifically, two threads are used to analyze two chromosomes concurrently; amplification and deletion CNAs in each chromosome are analyzed via two threads simultaneously; and multiple threads are invoked to perform parallel permutations with each focused on a subgroup of the samples.

2.2 Case study on TCGA datasets

The AISAIC package has been tested on multiple TCGA datasets and runs successfully on both Microsoft Windows and Linux platforms. On the GBM dataset that contains 513 paired samples, the AISAIC analysis reveals biologically relevant regions that encompass many well-known oncogenes or tumor suppressor genes including MDM4, AKT3, MYCN, PIK3CA, PDGFRA, EGFR, KRAS, CDKN2A/B and PTEN, and novel consensus regions that may warrant further studies. To assess biological and clinical implications of significant consensus deletions, AISAIC also provides the frequencies of homo- and hemi-deletions in the sample population, e.g. CDKN2A/B exhibits high homo-deletion rates among others. Figure 2 displays the *P*-values (in −log10 scale) of aberrations, where the dashed lines indicate the significance cutoff at the adjusted *P*-value of 0.05, and (%|%) are the rates of homo- and hemi-deletions, respectively. On the LUAD dataset (405 paired samples), the AISAIC analysis, again is highly consistent with previous reports, confirming many known cancer-related genes such as TERT, EGFR, KRAS, CDK4, CCNE1, CDKN2A/B, RB1, TP53 and CCBE1. In a comparative study on the GBM and LUAD datasets, AISAIC identifies several common genes such as EGFR, KRAS, CDK4, MDM2, CCNE1 and CDKN2A/B, as well as some

cancer type-specific genes such as a high AKT3 amplification rate in GBM and a high TP53 deletion rate in LUAD.

Detailed descriptions of method, software and more case studies on real copy number data are included in the Supplementary Material.

3 DISCUSSION

AISAIC presents a comprehensive and unsupervised approach to analyze DNA copy number aberrations in the cancer genome. AISAIC is supported by a well-grounded statistical framework and can detect homo/hemi-deletion types and rates, estimate and correct normal cell contamination, and identify significant consensus CNA regions. Tested on both simulations and TCGA datasets, AISAIC is effective at revealing novel consensus regions that harbor potential cancer 'driver' genes and enhancers. We expect that, with further development, AISAIC's methodology can be applied to other forms of genomic and epigenetic data such as DNA methylations (de Assis *et al*., 2012).

Funding: *In Silico* Research Centers of Excellence (ISRCE)/NCI (HHSN2612200800001E); National Institutes of Health (12ST1101, CA149147, CA160036, CA164384); and the Natural Science Foundation of China (61201312 and 61070137) (in part).

Conflict of Interest: none declared.

REFERENCES

Beroukhim,R. *et al*. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Carter,S.L. *et al*. (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol*., **30**, 413–421.

de Assis,S. *et al*. (2012) High-fat or ethinyl-oestradiol intake during pregnancy increases mammary cancer risk in several generations of offspring. *Nat. Commun*., **3**, 1053.

Yu,G. *et al*. (2011) BACOM: *in silico* detection of genomic deletion types and correction of normal cell contamination in copy number data. *Bioinformatics*, **27**, 1473–1480.

Yuan,X. *et al*. (2012) Genome-wide identification of significant aberrations in cancer genome. *BMC Genomics*, **13**, 342.