BIOINFORMATICS APPLICATIONS NOTE
Vol. 31 no. 3 2015, pages 429–431
doi:10.1093/bioinformatics/btu665

*Sequence analysis*
Advance Access publication October 8, 2014

# ExomeAI: detection of recurrent allelic imbalance in tumors using whole-exome sequencing data

Javad Nadaf*, Jacek Majewski and Somayyeh Fahiminiya*
Department of Human Genetics, Faculty of Medicine, McGill University and Genome Quebec Innovation Center, Montreal, Quebec, Canada

Associate Editor: Michael Brudno

## ABSTRACT

**Summary:** Whole-exome sequencing (WES) has extensively been used in cancer genome studies; however, the use of WES data in the study of loss of heterozygosity or more generally allelic imbalance (AI) has so far been very limited, which highlights the need for user-friendly and flexible software that can handle low-quality datasets. We have developed a statistical approach, ExomeAI, for the detection of recurrent AI events using WES datasets, specifically where matched normal samples are not available.

**Availability:** ExomeAI is a web-based application, publicly available at: http://genomequebec.mcgill.ca/exomeai.

**Contact:** JavadNadaf@gmail.com or somayyeh.fahiminiya@mcgill.ca

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

Received on June 17, 2014; revised on August 31, 2014; accepted on October 6, 2014

## 1 INTRODUCTION

In cancer, recurrent genomic aberrations are highly likely to affect oncogenes or tumor suppressor genes that drive cancer progression. One efficient way to identify such aberrations is to investigate genomic profile of allelic imbalance (AI) simultaneously across the genomes of a number of similar tumor samples. Detection of AI is based on the study of the relative proportion of the two alleles (A and B) at heterozygous sites (Wong *et al.*, 2004). In a diploid normal heterozygous locus, the expected frequency of the B allele is 0.5 (1:1 ratio). AI is defined as a significant deviation from this proportion. The genotypes, B, AAB, BB, which show loss of heterozygosity (LOH), duplication and copy-neutral LOH, respectively, are examples of AI. The first two cases are also examples of Copy Number Aberrations (CNA).

To date, investigation of AI/LOH in cancer studies has mainly been based on single nucleotide polymorphism (SNP) genotyping or comparative genomic hybridization arrays and the potential of WES data has not been fully exploited. To our knowledge, there is no publicly available software for identification of recurrent genomic AI segments, using WES data, shared across multiple tumor-only samples (Liu *et al.*, 2013). To address the limitation, we developed a novel software, ExomeAI, which can detect recurrent AI across cancer genomes by analyzing batches of WES data, and specifically in the absence of matched normal samples. We recently applied our approach to different cancer types [e.g. Small-Cell Carcinoma of the Ovary Hypercalcemic Type, (Witkowski *et al.*, 2014), gliomas and renal cell carcinoma (our unpublished data)] and successfully identified recurrent AI regions. To facilitate the analysis for non-computational cancer researchers, using Galaxy platform (Goecks *et al.*, 2010), we implemented our approach and developed a user-friendly web application.

## 2 METHODS

As shown in Supplementary Figure S1, ExomeAI gets a batch of tumor Variant Call Format files, as input (see Supplementary information for an alternative format). Using preset quality filters on variants, it converts read counts for each variant to B-allele frequency (BAF), where BAF is the number of reads with the non-reference base at the variant site divided by the total read count. Since only heterozygous variants are informative in AI detection, only variants with BAF values from 0.05 to 0.95 (default values) will be used as heterozygous variants for further analysis. The expected BAF in a normal sample is 0.5 (one copy of B and one copy of A). The absolute deviation of BAF values (dBAF) from the expected value ($\text{dBAF} = |\text{BAF} - 0.5|$) is used for segmentation and segment-wise calling. Segments of similar dBAF values are detected using circular binary segmentation (CBS) algorithm (Venkatraman and Olshen, 2007). The CBS algorithm is applied to each arm of all chromosomes. For segment-wise calling, where matched samples are not available, the dBAF values of each segment can be compared with a fixed threshold. Although a fixed cutoff approach has been used successfully for SNP arrays (Staaf *et al.*, 2008), the proper threshold may vary for WES datasets with different qualities (e.g. varying sequencing depths). We propose a two-step approach: We first call AI at each variant over the genome using a binomial test (Sathirapongsasuti *et al.*, 2011) and we calculate the mean dBAF values for non-AI variants. In the second step, a Wilcoxon signed rank test is performed to evaluate the distribution of dBAF in each segment (see Supplementary information for details).

One of the main issues in AI or CNA calling is false positives, which gets even more challenging when the matched normal sample is not available. In cancer analysis, it is desirable to remove candidate segments that may either represent common Copy Number Variations (CNVs) or technology or alignment artifacts. In order to remove false positives calls, we created a control database of 500 non-cancer WES samples (the database is kept updated). The statistically significant AI segments will then be compared with the control database to find the number of hits within the database. By default, ExomeAI counts each overlap of the control database with more than 50% of the query segment as one hit. In the next step, the significant segments that were not seen (or seen less than a given

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

frequency) in the control database will be used to find the recurrent AI. The output of this final step will be a list of identified segments with names and number of tumor samples that share the segment (Supplementary Table S2) along with the plots of the segments for each chromosome (Supplementary Figs S6 and S7).

## 3 IMPLEMENTATION AND EXAMPLES

The development of ExomeAI method has been motivated by the need to detect the recurrent AI using WES dataset of tumor-only SCCOHT samples. By applying this method, we showed that 19p were deleted in all patients with SMARCA4 mutations (Witkowski *et al.*, 2014). Since then, we have successfully applied our method to several other cancer types such as glioma, renal cell carcinoma (unpublished data) and ETMR (Kleinman *et al.*, 2014) (see Supplementary information) and showed that the approach is sensitive to clearly detect the recurrent AI across cancer genomes, in the absence of paired normal samples.

As we believe that the approach has a wide application in cancer genome research and in order to make it publically available to non-computational cancer researchers, we developed a web-based application facilitated by Galaxy platform (Goecks *et al.*, 2010) which is platform for biomedical genomic research. ExomeAI server is equipped with two 3.4-GHz Intel processors (a total of 8 cores/16 threads), 32 GB of RAM, and 10 TB of disc space, and is accessible through http://genomequebec.mcgill.ca/exomeai.

## 4 EVALUATION USING SIMULATED DATA

In order to evaluate the accuracy of the method, we created a simulated dataset using 10 real non-cancer Exomes (not included in the control database) and we added over 600 CNAs on 22 autosomal chromosomes (hg19). Ranging in length, size and copy number, CNVs were randomly located on genome. Simulated aberrations included all combinations of: length (Mb) = [1,3,5,10,15,20], copy number = [1,3] and non-aberrant/normal DNA fraction (%) = [5,10,20,30,40,50].

We analyzed the simulated dataset using ExomeAI and four other softwares: SomatiCA, ExomeCNV (LOH analysis), XHMM and CoNIFER. For the methods that required matched normal samples (ExomeCNV and SomatiCA), samples before adding simulated CNAs were used as matched normal samples. For all softwares, default parameters were used (see supplementary information for details).

Sensitivity was calculated as the overlap of identified CNAs with true simulated ones divided by the total length of simulated CNAs. Specificity was calculated as TN/(TN + FP), where TN is the length of True Negative and FP is the length of False Positive regions at single base resolution.

Supplementary Figure S8 shows the sensitivity and specificity of all five softwares averaged over all Exomes ($n$ = 10). The Circos plot (Supplementary Fig. S9) depicts the results for all methods across the genome of one simulated Exome. As shown in Supplementary Table S3, ExomeAI achieved a sensitivity of 0.76 with specificity >0.99. Using matched normal samples, SomatiCA could achieve a higher sensitivity (0.82); however, its specificity was lower (0.97).

In general, the softwares that were primarily developed for CNA detection in cancer (ExomeAI, ExomeCNV and SomatiCA) showed higher power to detect CNA. It was at least partially because ExomeCNV and SomatiCA used matched normal samples and ExomeAI used the internal control database. CoNIFER and XHMM had very similar results. The sensitivities of the two software were low in the simulated dataset, which is consistent with previous studies (Tan *et al.*, 2014); however, they may perform better when CNVs are more homogeneous and are present only in a low proportion of samples.

## 5 DISCUSSION AND CONCLUSION

We present a novel web application for detection of recurrent LOH or, more generally, AI events across batches of tumor-only WES datasets. We show that ExomeAI can be effectively applied in various cancer types. Using a control database of non-cancer samples, ExomeAI overcomes the limitation of the ‘obligatory’ usage of matched normal samples and efficiently reduces the rate of false positive calls. Working on multiple samples and looking for recurrent events further reduce the chance of false positives. In case where some of the samples have matched-normal counterparts, the recurrent aberrations can be further studied in those samples using other softwares which have been well reviewed elsewhere (Liu *et al.*, 2013; Alkodsi *et al.*, 2014).

## ACKNOWLEDGEMENTS

We wish to thank Information Technology (IT) group at Genome Quebec Innovation Center, specifically, Terrance Mcquilkin, Alexandru Guja and Marc-Andre Labonte for providing infrastructure facilities and hosting ExomeAI server.

*Funding*: This work was supported by the Génome Québec, Genome Canada and the Canadian Institutes of Health Research [grant number 77764]; Tier II Canada Research Chair award (to J.M.).

*Conflict of interest*: none declared.

## REFERENCES

Alkodsi,A. *et al.* (2014) Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief. Bioinform*, [Epub ahead of print, doi: 10.1093/bib/bbu004, March 5, 2014].

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Kleinman,C.L. *et al.* (2014) Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR. *Nat. Genet.*, **46**, 39–44.

Liu,B. *et al.* (2013) Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*, **4**, 1868–1881.

Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, **27**, 2648–2654.

Staaf,J. *et al.* (2008) Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.*, **9**, R136.

Tan,R. *et al*. (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat*., **35**, 899–907.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Witkowski,L. *et al*. (2014) Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. *Nat*. *Genet*., **46**, 438–443.

Wong,K.-K. *et al*. (2004) Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. *Nucleic Acids Res*., **32**, e69.