BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 11 2013, pages 1461–1462
doi:10.1093/bioinformatics/btt151

Sequence analysis

Advance Access publication March 28, 2013

FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data

Yuhao Shi[1,2,*] and Jacek Majewski[1,2]

[1]Department of Human Genetics, McGill University, Montreal, Quebec, Canada H3A 1B1 and [2]Genome Quebec Innovation Centre, Montreal, Quebec, Canada H3A 0G1

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Rare copy number variations (CNVs) are frequent causes of genetic diseases. We developed a graphical software package based on a novel approach that can consistently identify CNVs of all types (homozygous deletions, heterozygous deletions, heterozygous duplications) from exome-sequencing data without the need of a paired control. The algorithm compares coverage depth in a test sample against a background distribution of control samples and uses principal component analysis to remove batch effects. It is user friendly and can be run on a personal computer.

Availability and implementation: The main scripts are implemented in R (2.15), and the GUI is created using Java 1.6. It can be run on all major operating systems. A non-GUI version for pipeline implementation is also available. The program is freely available online: https://sourceforge.net/projects/fishingcnv/

Contact: yuhao.shi@mail.mcgill.ca

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

Received on December 20, 2012; revised on March 2, 2013; accepted on March 25, 2013

1 INTRODUCTION

Many sporadic and Mendelian disorders are caused by exonic mutations that alter the amino acid sequence of the affected gene. Exome sequencing has so far shown great utilty in elucidating single nucleotide mutations that contribute to these diseases. However, using exome sequencing to detect rare copy number variations (CNVs) that contribute to diseases remain a challenge.

We developed FishingCNV, a graphical software package for comprehensive analysis of CNVs in exome-sequencing data. We included a matched case/control analysis feature adapted from ExomeCNV, but more importantly, we implemented a new CNV detection algorithm that prioritizes rare variants for non-matched sample analysis. Many sequencing centers have by now accumulated large internal sets of exome samples that can be used as controls for CNV detection in new unmatched samples. Our method relies on pooling such a control set, comparing the exonic read depth of the test sample against the distribution of read depths in the control set. To remove batch-to-batch variations in the control set, we applied principal component analysis (PCA) and removed the top contributing principal components (~2–10) from the data. We bundled GATK (McKenna *et al.*, 2010) into our program to extract coverage information. Read depth information is then converted to Reads Per Kilobase per Million mapped reads (RPKM) (Mortazavi *et al.*, 2008), and the test sample is segmented into regions of similar RPKM ratios using circular binary segmentation (CBS) before comparison against the control distribution (Olshen *et al.*, 2004). The output is a file that contains segmented CNV calls ranked by statistical confidence (See Supplementary Materials for details on method). We tested our algorithm on simulated and actual exome data and show that it is effective in identifying homozygous as well as heterozygous CNVs of varying length. We also show that principal component removal is effective in reducing false positives. A summary of the program workflow is presented in Figure 1. We compared our algorithm against ExomeCNV (Sathirapongsasuti *et al.*, 2011) and Conifer (Krumm *et al.*, 2012) and show that FishingCNV consistently outperforms previous approaches (See Supplementary Materials for data).

2 RESULTS

2.1 Simulation and performance

We created test files with presumably no CNVs by averaging the RPKM values of a subset of samples from the control set. In each file, we simulated heterozygous deletions, heterozygous duplications and homozygous deletions of length 1 exon, 2 exons, 5 exons and 10 exons by changing the RPKM data within the file. An FDR-adjusted *P*-value of 0.05 was used as the threshold of detection. Our results show that our algorithm is effective in detecting homozygous deletions, as well as heterozygous deletions and duplications at the resolution of a single exon (See Supplementary Materials). Sensitivity increases as the size of the affected locus increases. Homozygous CNVs are detected with higher sensitivity.

We compared our algorithm against ExomeCNV (Sathirapongsasuti *et al.*, 2011) and Conifer (Krumm *et al.*, 2012) using the same simulated dataset. We show that ExomeCNV is not effective in detecting heterozygous variants or larger sized CNVs and Conifer is unsuccessful when CNVs are small (one to two exons long), but FishingCNV performed well for all CNV types and lengths. The performance of Conifer for larger sized CNVs is similar to FishingCNV. Overall, FishingCNV outperforms previous approaches to CNV detection.

*To whom correspondence should be addressed

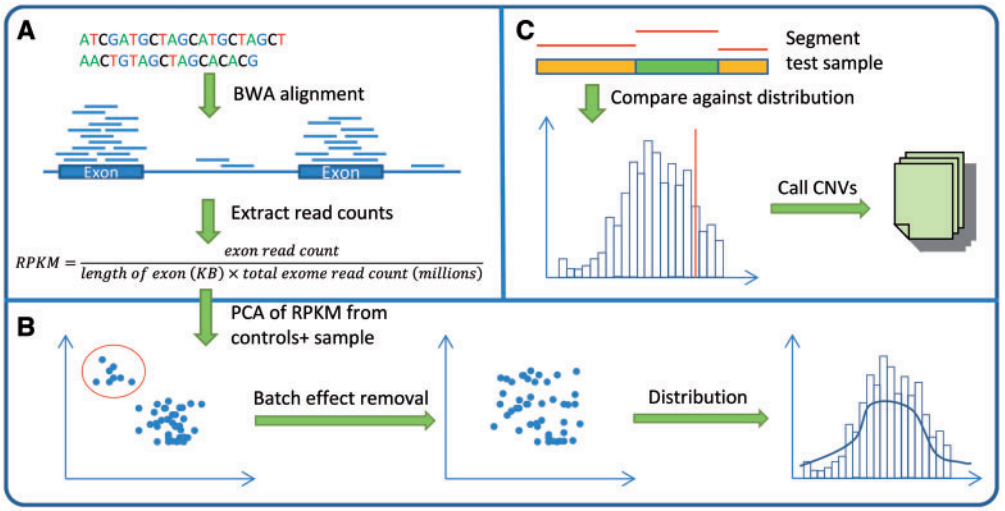

Fig. 1. Summary of the workflow of FishingCNV. **(A)** RPKM values are extracted from each test or control sample. **(B)** PCA-based method is used to remove batch-to-batch variations in the data. **(C)** The test sample is segmented into regions of similar read depth ratio (with respect to the average of the controls) and tested against the read depth distribution of all the controls

2.2 CNV detection—two validated examples

We tested our program on two patients suffering from metaphyseal dysplasia with maxillary hypoplasia and brachydactyly (MDMHB), and it identified a 4 exon amplification spanning *SUPT3H* and *RUNX2* as the sixth most confident call ($P < 1.06e\text{-}11$). This CNV was verified as the causal variant in a previous study (Moffatt *et al.*, 2013).

In addition, we analyzed a sample from a study on mandibulofacial dysostosis with microcephaly (MFDM), which was previously shown to harbor a deleterious heterozygous deletion in the last 9 exons of the *EFTUD2* gene (Lines *et al.*, 2012). The known deletion was the highest ranked call produced by our algorithm, with a *P*-value of 4.38e-12. FishingCNV determines it to be a 12 exons long deletion that spans the last 9 exons of *EFTUD2* and the first 3 exons of the neighboring gene *HIGD1B*.

The analysis of the MFDM illustrates some of the problems that may be encountered in real-life analysis. We note that the MFDM sample is an outlier, as identified by PCA, and produces many more CNV calls than is expected. However, our results demonstrate the power of our approach because after the ranking of CNV calls according to the *P*-values (even in cases where the normal approximation does not hold well), the true variants can be identified within the top ranking calls, despite the non-ideal input.

3 CONCLUSION

We believe that looking for rare CNVs in exome-sequencing data can be a powerful way of detecting new disease-causing mutations. Until now, efforts on this front have been largely limited by the computational tools available to us, but here we presented a new graphical software package for easy detection of CNVs and showed that it is effective in identifying causal CNVs in rare diseases. This program should have vast applications in genomics research, and will aid in the ever-expanding scientific efforts to identify the genetic causes of rare diseases, such as the Finding of Rare Disease Genes in Canada (FORGE) and Rare Disease Consortium for Autosomal Loci (RaDiCAL) projects.

ACKNOWLEDGEMENTS

We would like to thank the FORGE Canada Consortium for allowing us to use exome data to test the algorithm and Kevin Ha for providing the initial ideas that gave rise to this project.

Funding: This work was supported by the Canadian Institutes of Health Research (grant number 77764); Tier II Canada Research Chair Award (to J.M.).

Conflict of Interest: none declared

REFERENCES

Krumm,N. *et al.* (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res.*, **22**, 1525–1532.

Lines,M.A. *et al.* (2012) Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. *Am. J. Hum. Genet.*, **90**, 369–377.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Moffatt,P. *et al.* (2013) Metaphyseal dysplasia with maxillary hypoplasia and brachydactyly is caused by a duplication in RUNX2. *Am. J. Hum. Genet.*, **92**, 252–258.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

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