Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

FocalCall: An R Package for the Annotation of Focal Copy Number Aberrations

Oscar Krijgsman[1,†], Christian Benner[1,‡], Gerrit A. Meijer[1], Mark A. van de Wiel[2,3] and Bauke Ylstra[1]

[1]Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands. [2]Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands. [3]Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands. [†]Current address: Department of Molecular Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands. [‡]Current address: Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland.

ABSTRACT: In order to identify somatic focal copy number aberrations (CNAs) in cancer specimens and to distinguish them from germ-line copy number variations (CNVs), we developed the software package FocalCall. FocalCall enables user-defined size cutoffs to recognize focal aberrations and builds on established array comparative genomic hybridization segmentation and calling algorithms. To distinguish CNAs from CNVs, the algorithm uses matched patient normal signals as references or, if this is not available, a list with known CNVs in a population. Furthermore, FocalCall differentiates between homozygous and heterozygous deletions as well as between gains and amplifications and is applicable to high-resolution array and sequencing data.

AVAILABILITY AND IMPLEMENTATION: FocalCall is available as an R-package from: https://github.com/OscarKrijgsman/focalCall. The R-package will be available in Bioconductor.org as of release 3.0.

KEYWORDS: R-package, focal CNAs, DNA copy number, sequencing, aCGH

CITATION: Krijgsman et al. FocalCall: An R Package for the Annotation of Focal Copy Number Aberrations. *Cancer Informatics* 2014:13 153–156 doi: 10.4137/CIN.S19519.

RECEIVED: August 19, 2014. **RESUBMITTED:** September 22, 2014. **ACCEPTED FOR PUBLICATION:** September 24, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Technical Advance

FUNDING: This study was supported by the VUmc Cancer Center Amsterdam (VUmc-CCA) and performed within the framework of CTMM, the Center for Translational Molecular Medicine. DeCoDe project (grant 03O-101). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: b.ylstra@vumc.nl

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

The increase in the resolving power of DNA copy number profiling techniques has led to the simultaneous discovery of the extend of (1) copy number variations (CNVs) of germ-line origin in the general population[1] as well as (2) focal copy number aberrations (CNAs) of somatic origin in cancer specimens.[2] The limited size of focal CNAs offers an excellent opportunity to pinpoint potential driver genes in cancer.[3–6] CNV detection usually is an obstacle in the identification of cancer driver genes. Unfortunately, with copy number assessment in tumors, a mix of focal CNAs and CNVs is detected, of which most have the same appearance (Fig. 1). A procedure that partly circumvents the interference of CNVs in tumor samples is the simultaneous analysis of matched patient normal DNA. However, if the diploid balance in a tumor is disturbed, ie, a single copy gain, a heterozygous CNV will still give rise to a superimposed focal signal. To recognize the CNAs, a negative selective procedure can be applied by identifying CNVs detected in the healthy population through the analysis of a series of healthy normal copy number profiles, preferably patient group matched, or otherwise an external database of genomic variants (ie, DGV).[7] Alternatively, an effective positive selection is through the identification of focal homozygous deletions and high-level amplifications that differ in amplitude from CNVs.[5] This approach however neglects many heterozygous focal CNAs.

Despite the great opportunities focal CNAs offer for cancer gene discovery, only few software tools are available that appreciate them, eg, GISTIC, WIFA, and control-FREEC.[8–10]

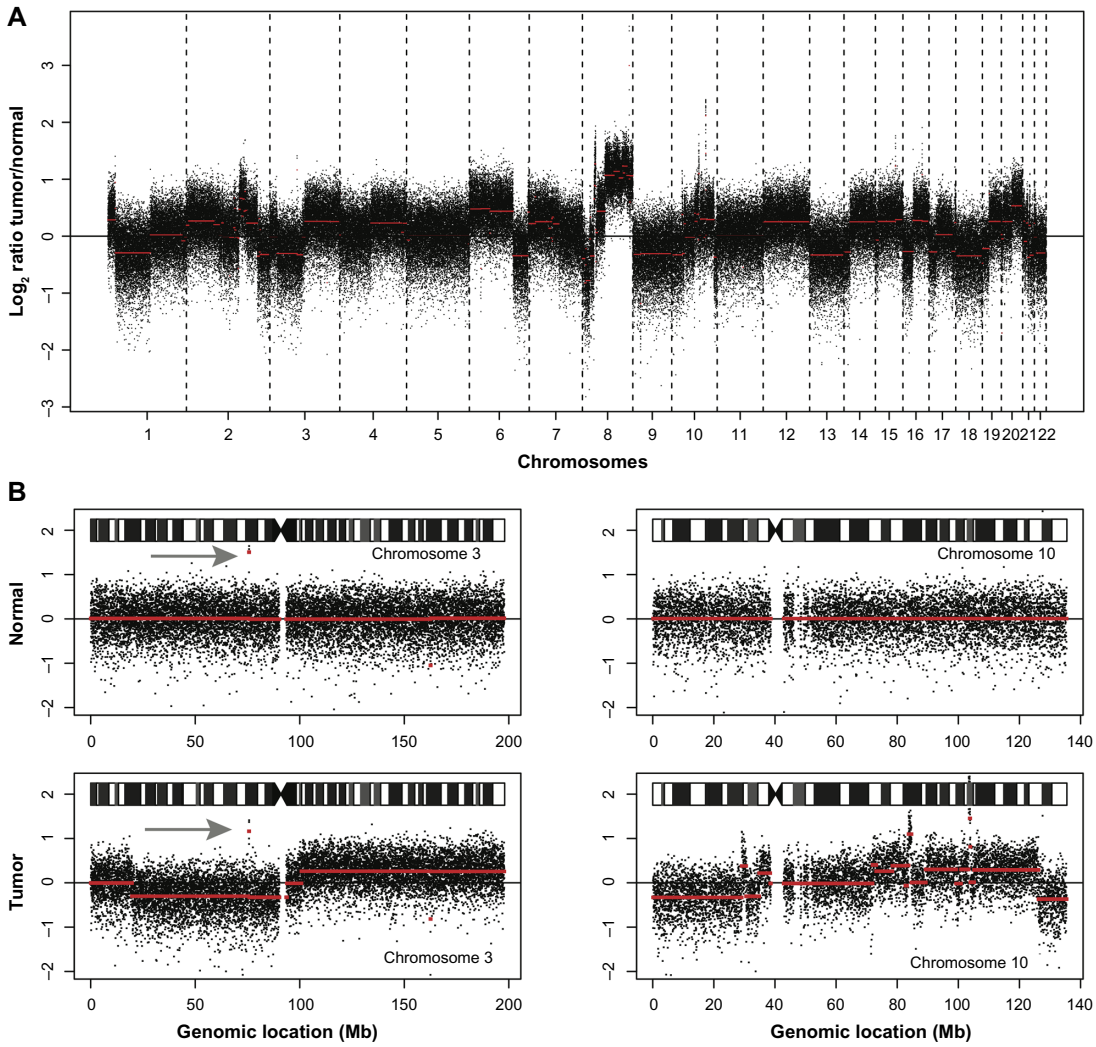

Figure 1. Copy number profiles of a lung cancer sequencing sample and matched patient normal signal.[12] Panel (**A**) shows all aberrations in the tumor sample. X-axis represents the bins ordered according to their chromosomal location. Y-axis represents the $\log_2$ ratio (right side). The red line indicates the segmented values as obtained using circular binary segmentation in CGHcall.[11] Panel (**B**) shows chromosomes 3 (left) and 10 (right) both for patient normal and tumor sample. The gray arrow in the left panels indicates a focal CNV present in both tumor and matched patient normal sample. Somatic focal CNAs on chromosome 10 are only present in the tumor and not in the matched patient normal sample. Focal CNAs and CNVs were detected using *focalCall()*.

Both GISTIC and WIFA were developed for array data and can detect focal CNAs in series of samples, but not in individual tumor profiles. GISTIC has a dedicated option to discriminate focal CNAs from CNVs based on an external database. Control-FREEC was developed to calculate genome-wide copy number information from whole genome sequencing data and can distinguish CNAs from CNVs, provided a matched patient normal signal is available.

Here, we present FocalCall, which elaborates on commonly used segmentation and calling algorithms.[11] A user-defined size cutoff allows for the identification of focal CNAs in individual samples as well as series of samples and can distinguish them from CNVs. FocalCall accepts copy number data from both high-resolution genome-wide array comparative genome hybrizations (aCGH) and single nucleotide polymorphism (SNP) arrays as well as data from sequencing data experiments,[12] with or without a matched patient normal signal.

Methods

Patient materials and settings. FocalCall was evaluated with four publicly available data sets: (1) shallow whole genome sequencing data (~0.2 × genome coverage) from tumor and normal DNA of a lung cancer patient[12]; (2) SNP array (250K) data of 371 lung cancer patients without matched patient normal samples[2]; (3) aCGH data (244K) of 74 glioblastoma multiforme (GBM) patients hybridized against its matched normal[13]; and (4) aCGH data (105K) of 60 high-grade cervical cancer pre-curser lesions hybridized against a pool of 100 healthy individuals.[4] Dataset 4 is available from the Gene Expression Omnibus (GSE34575) and used as an example dataset in the R-package.

Detection of recurrent aberrations. Standard data output as produced by CGHcall[11] was used as input for the main function *focalCall()*. Aberrations below the user-defined size threshold for focal CNAs (default = 3 Mb) were identified

in each cancer sample and categorized as "gain", "loss", "amplification", or "homozygous deletion". For each region, the smallest region of overlap (SRO) was calculated over the complete sample set. Complex regions may contain multiple SROs (Supplementary Fig. 1 and 2). To determine whether focal CNAs were enriched for cancer driver genes, enrichment analysis was performed.[3] In brief, enrichment analysis was implemented whereby 10,000 sets of simulated focal CNAs were randomly generated throughout the genome, with the same amount and length as the observed focal CNAs in the dataset. Overlap was determined of the simulated focal CNAs with the published list of cancer sensus genes and the significance of enrichment expressed as a *P* value.

Distinction between focal CNAs and CNVs. For each SRO (Supplementary Fig. 1), the percentage of overlap of focal CNAs with a normal reference or known CNVs is returned. If matched patient reference data are available, this can be provided in *focalCall()* as a separate CGHcall object. If no matched patient reference is available, focal CNAs are compared to a list of genomic locations of known CNVs, which can be provided in *focalCall()* as a flat text or bed file.

Reporting of focal CNA. The function *igvFiles()* generates tracks compatible with the Integrative Genome Viewer (IGV, www.broadinstitute.org/igv/home) for CNA frequency, focal CNA frequency, and segmentation values per sample (Supplementary Fig. 3). This allows the user to visually inspect the results generated by FocalCall. The functions *freqPlot()* and *freqPlotFocal()* generate .png file for CNA frequency and focal CNA frequency, respectively (Fig. 2).

Computational time. Computational times for the detection of focal CNAs in the GBM dataset ($n = 74$, 244K probes) with default parameters are approximately 7 minutes on a standard desktop computer with a 1.7 GHz CPU and 4 Gb of RAM.

Results

Detection of focal CNAs in single patient and series of tumors. The lung cancer sequencing data yielded a total of 38 focal gains and losses: 7 were identified as CNVs and 31 as focal CNAs, of which 6 were high-level amplifications (including FGFR1) and 4 were homozygous deletions (including CDKN2A, Fig. 1 and Supplementary Table).

The lung cancer SNP array dataset yielded a total of 503 focal CNAs with a frequency >5%. A total of 43 of the focal gains and losses overlapped with the CNV regions as archived in the DGV database.[7] All genes in focal CNAs detected by GISTIC in the original paper were also detected by FocalCall.2 The remaining 460 detected focal CNAs were enriched for known cancer driver genes ($n = 6$, $P < 0.05$) and included GNAS and KDM5A.

The GBM aCGH dataset yielded a total of 434 somatic focal CNAs and 90 CNVs. The focal CNAs encompassed known cancer driver genes like EGFR, PTEN, and CDKN2A. All 20 focal CNAs previously reported by GISTIC[13] were recognized by FocalCall. Additionally detected focal CNAs showed a highly significant enrichment for known cancer driver genes ($n = 38$, $P < 0.008$).

The cervical precursor lesion aCGH dataset yielded a total of 94 focal CNAs with FocalCall. Two of the identified

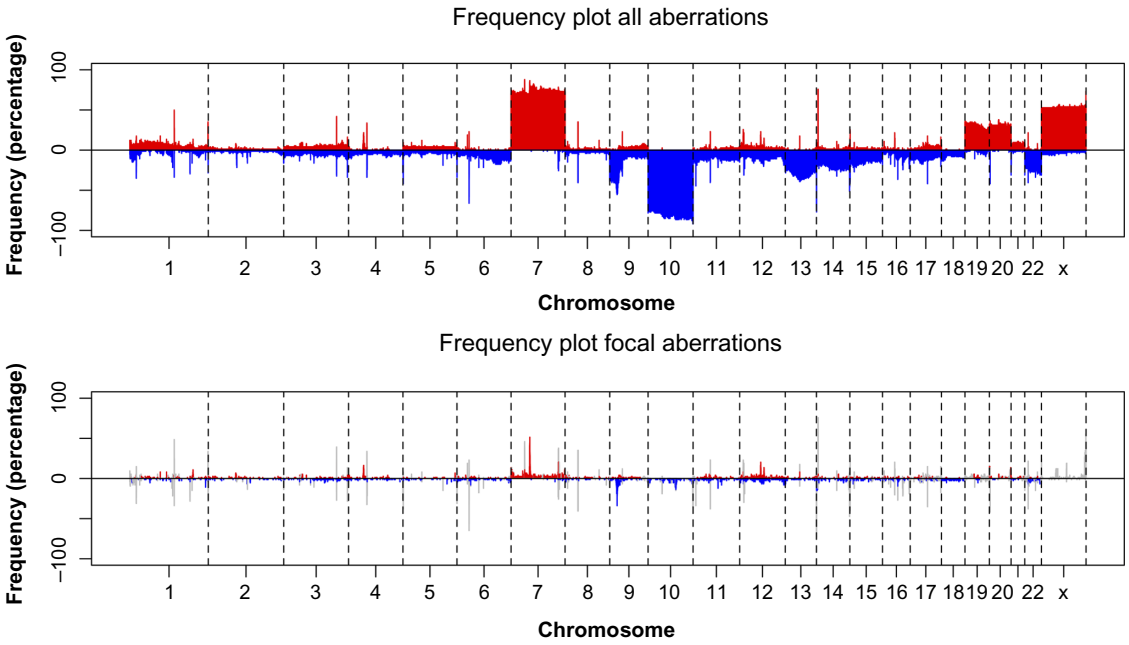

Figure 2. Frequency plots of the GBM dataset of all aberrations (top) and focal aberrations and CNVs (bottom) as generated by FocalCall functions *freqPlot()* and *FreqPlotFocal()*. Red indicates a gain and blue indicates a loss. In the frequency plot of focal aberrations (bottom), the somatic focal aberrations are indicated in red for gains and blue for losses. CNVs are indicated in gray, both for gains and losses.

genes, hsa-mir-375 and EYA2, were functionally tested and validated as a new oncogene and tumor suppressor gene.[4] The data and example scripts for this dataset are available in the R-package.

Conclusion

Focal CNAs provide an excellent opportunity to detect potential cancer driver genes.[6] Through advances in techniques, the resolution of DNA copy number detection has increased enormously and the changes we can identify have become smaller. Accurate detection and distinction of somatic aberrations from germ-line CNVs are thereby mandatory. FocalCall offers researchers a user-friendly tool to detect focal CNAs in high-resolution DNA copy number data and provides multiple methods to distinguish these from CNVs. FocalCall elaborates on a widely used DNA copy number tool CGHcall[11] and comprehensive genome analysis packages in the R/Bioconductor environment. In addition, FocalCall output in the IGV data format allows for easy browsing through the data and provides a direct link with the genes affected.

In conclusion, we provide an alternative and sensitive procedure for the detection of focal CNAs applicable to both individual and series of samples analyzed by either array or next-generation sequencing.

Acknowledgments

We would like to thank Vanessa St. Aubyn for critically reading our manuscript and for useful comments.

Author Contributions

Conceived and designed the experiments: OK, BY. Analyzed the data: OK, CB. Wrote the first draft of the manuscript: OK, BY. Contributed to the writing of the manuscript: OK, BY, MvdW, GAM. Agree with manuscript results and conclusions: OK, CD, GAM, MvdW, BY. Jointly developed the structure and arguments for the paper: OK, BY. Made critical revisions and approved final version: OK, CB, MvdW, GAM, BY. All authors reviewed and approved of the final manuscript.

Supplementary Materials

Supplementary Figure 1. Graphical explanation how the smallest region of overlap is calculated.

Supplementary Figure 2. Flowchart for FocalCall procedures from input to output data.

Supplementary Figure 3. IGV example with the segment values of the GBM dataset.

Supplementary Table. FocalCall output for the single sample lung patient data.

Supplementary Vignette. Explanation, R-code and output of the executable example data provided with the R-package.

REFERENCES

1. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*. 2006;7:85–97.
2. Weir BA, Woo MS, Getz G, et al. Characterizing the cancer genome in lung adenocarcinoma. *Nature*. 2007;450:893–98.
3. Brosens RP, Haan JC, Carvalho B, et al. Candidate driver genes in focal chromosomal aberrations of stage II colon cancer. *J Pathol*. 2010;221(4):411–24.
4. Bierkens M, Krijgsman O, Wilting SM, et al. Focal aberrations indicate EYA2and hsa-miR-375as oncogene and tumor suppressor in cervical carcinogenesis. *Genes Chromosomes Cancer*. 2012;52:56–68.
5. Leary RJ, Lin JC, Cummins J, et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci USA*. 2008;105:16224–9.
6. Krijgsman O, Carvalho B, Meijer GA, Steenbergen RDM, Ylstra B. Focal chromosomal copy number aberrations in cancer-needles in a genome haystack. *Biochim Biophys Acta*. 2014;1843:2698–704.
7. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42:D986–92.
8. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28:423–5.
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12:R41.
10. Hur Y, Lee H. Wavelet-based identification of DNA focal genomic aberrations from single nucleotide polymorphism arrays. *BMC Bioinformatics*. 2011;12:146.
11. van de Wiel MA, Picard F, van Wieringen WN, Ylstra B. Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform*. 2011;12:10–21.
12. Gusnanto A, Taylor CC, Nafisah I, Wood HM, Rabbitts P, Berri S. Estimating optimal window size for analysis of low-coverage next-generation sequence data. *Bioinformatics*. 2014;30(13):1823–9.
13. Brennan CW, Verhaak RG, McKenna A, et al. The somatic genomic landscape of glioblastoma. *Cell*. 2013;155:462–77.