GeneBreak: detection of recurrent DNA copy number aberration-associated chromosomal breakpoints within genes
(version 2; referees: 2 approved)

Evert van den Broek¹,⁵, Stef van Lieshout¹, Christian Rausch¹,⁵, Bauke Ylstra¹, Mark A. van de Wiel²,³, Gerrit A. Meijer¹,⁵, Remond J.A. Fijneman¹,⁵, Sanne Abeln ⁴

¹Department of Pathology, VU University Medical Center, Amsterdam, 1081 HZ, Netherlands
²Department of Epidemiology & Biostatistics, VU University Medical Center, Amsterdam, 1081 HZ, Netherlands
³Department of Mathematics, VU University Medical Center, Amsterdam, Amsterdam, 1081 HV, Netherlands
⁴Department of Computer Science, VU University Medical Center, Amsterdam, 1081 HV, Netherlands
⁵Department of Pathology, Netherlands Cancer Institute, Amsterdam, 1066CX, Netherlands

Abstract
Development of cancer is driven by somatic alterations, including numerical and structural chromosomal aberrations. Currently, several computational methods are available and are widely applied to detect numerical copy number aberrations (CNAs) of chromosomal segments in tumor genomes. However, there is lack of computational methods that systematically detect structural chromosomal aberrations by virtue of the genomic location of CNA-associated chromosomal breaks and identify genes that appear non-randomly affected by chromosomal breakpoints across (large) series of tumor samples. ‘GeneBreak’ is developed to systematically identify genes recurrently affected by the genomic location of chromosomal CNA-associated breaks by a genome-wide approach, which can be applied to DNA copy number data obtained by array-Comparative Genomic Hybridization (CGH) or by (low-pass) whole genome sequencing (WGS). First, ‘GeneBreak’ collects the genomic locations of chromosomal CNA-associated breaks by a genome-wide approach, which can be applied to DNA copy number data obtained by array-Comparative Genomic Hybridization (CGH) or by (low-pass) whole genome sequencing (WGS). First, ‘GeneBreak’ collects the genomic locations of chromosomal CNA-associated breaks that were previously pinpointed by the segmentation algorithm that was applied to obtain CNA profiles. Next, a tailored annotation approach for breakpoint-to-gene mapping is implemented. Finally, dedicated cohort-based statistics is incorporated with correction for covariates that influence the probability to be a breakpoint gene. In addition, multiple testing correction is integrated to reveal recurrent breakpoint events. This easy-to-use algorithm, ‘GeneBreak’, is implemented in R (www.cran.r-project.org) and is available from Bioconductor (www.bioconductor.org/packages/release/bioc/html/GeneBreak.html).

This article is included in the RPackage gateway.
This article is included in the Bioconductor gateway.

**Corresponding authors:** Remond J.A. Fijneman (r.fijneman@nki.nl), Sanne Abeln (s.abeln@vu.nl)

**Author roles:** van den Broek E: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; van Lieshout S: Methodology, Software, Writing – Review & Editing; Rausch C: Resources, Software, Validation, Writing – Review & Editing; Ylstra B: Resources, Writing – Review & Editing; van de Wiel MA: Conceptualization, Formal Analysis, Methodology, Software, Validation, Writing – Review & Editing; Meijer GA: Conceptualization, Funding Acquisition, Resources, Supervision, Writing – Review & Editing; Fijneman RJA: Conceptualization, Funding Acquisition, Methodology, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; Abeln S: Conceptualization, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** van den Broek E, van Lieshout S, Rausch C et al. GeneBreak: detection of recurrent DNA copy number aberration-associated chromosomal breakpoints within genes [version 2; referees: 2 approved] F1000Research 2017, 5:2340 (doi: 10.12688/f1000research.9259.2)

**Copyright:** © 2017 van den Broek E et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Grant information:** This work was supported by the VUmc-Cancer Center Amsterdam [to E.vd.B.]; performed within the framework of the Center for Translational Molecular Medicine, DeCoDe project [03O-101]; and CTMM-TraIT [05T-401 to EvdB, SvL, BY, GM, RF and SA]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**First published:** 19 Sep 2016, 5:2340 (doi: 10.12688/f1000research.9259.1)
Introduction
Tumor development is driven by irreversible somatic genomic aberrations such as single nucleotide variants (SNVs) and chromosomal aberrations including numerical as well as structural changes. Genome-wide somatic DNA copy number aberrations (CNA) profiling is a widely established approach to characterize chromosomal aberrations in cancer genomes. At present, application of computational methods has mainly been focused on the analysis of numerical aberrations of chromosomal segments. Evidence is emerging that genes affected by structural chromosomal aberrations, i.e. genes affected by chromosomal breaks, represent a biologically and clinically relevant class of mutations in many cancer types including solid tumors. Importantly, the actual locations of chromosomal CNA-associated breakpoints, which are the points of copy number level shift in somatic CNA profiles, indicate underlying chromosomal breaks and thereby genomic locations affected by somatic structural aberrations. Hence, the wide availability of large series of high-resolution DNA copy number data by for instance array-Comparative Genomic Hybridization (CGH) or by (low-pass) whole genome sequencing (WGS) approaches enables to systematically search for regions and genes that are affected by CNA-associated structural chromosomal changes. Computational methods determining numerical CNAs, consequently, also yield CNA-associated breakpoint locations. However, it is not trivial to identify genes that are recurrently affected by CNA-associated chromosomal breakpoints across (large) series of cancer samples since this methodology also requires dedicated computational methods including comprehensive statistical evaluation.

We here provide a computational method, ‘GeneBreak’, that identifies chromosomal breakpoint locations using DNA copy number profiles. A tailored annotation approach maps breakpoint locations to genes for each individual profile. Moreover, dedicated comprehensive cohort-based statistical analysis including correction for covariates that influence the probability to be a breakpoint gene and multiple testing pinpoints genes that are non-randomly and recurrently affected by chromosomal breaks across multiple tumor samples. ‘GeneBreak’ is implemented in R (www.cran.r-project.org) and is available from Bioconductor (www.bioconductor.org/packages/release/bioc/html/GeneBreak.html). The Bioconductor vignette describes a detailed example workflow of CNA data obtained by analysis of 200 array-CGH samples. A schematic overview of computational methods is depicted in Figure 1.

Methods
DNA copy number profiles
The breakpoint detection method we provide is amenable for data from any DNA copy number discovery platform, e.g. array-CGH and (low-pass) WGS. For optimal results, ‘GeneBreak’ takes DNA copy number data that are pre-processed by the R-package ‘CGHcall’ or ‘QDNAseq’, both based on the Circular Binary Segmentation algorithm, as input. Alternatively, segmented values (log2-ratios) from a different copy number detection algorithm can be used. In addition, it is recommended to provide discrete DNA copy number states (e.g. loss, neutral, gain) that can be used for breakpoint selection. Bioconductor vignette and manual describe commands and workflows in detail (See Supplementary material).

Breakpoint detection and filter options
Breakpoints are defined by the chromosomal locations that separate the contiguous DNA copy number segments pinpointed by a segmentation algorithm. ‘GeneBreak’ identifies chromosomal breakpoint locations for each individual DNA copy number profile. Instead of taking all detected breakpoints, users may want to define more precisely what breakpoints to take into account, based on the two flanking DNA copy number segment characteristics. One of the following three selection options can be applied. A) Copy number deviation: this selects breakpoints where the shift in log2-ratio between two consecutive DNA copy number segments exceeds the user-defined threshold; B) CNA-associated breakpoints: this selects all breakpoints between consecutive DNA copy number segments, except for breakpoints flanked by two copy number neutral segments; C) CNA-breakpoints: this selects only those breakpoints flanked by segments with dissimilar discrete DNA copy number states.

Breakpoints are defined by the genomic start position of the copy number segments. DNA copy number profiling data is typically granular due to the distance between microarray probes or bin size of WGS copy number data. This means that the genomic location of a breakpoint is not detected at nucleotide resolution but represents a chromosomal interval with a size that is determined by microarray probe density or WGS bin size.

Breakpoint gene identification
For identification of genes affected by chromosomal breakpoints the built-in gene annotations can be used. Alternatively, a user-defined gene annotation file can be provided (see Bioconductor vignette and manual for further details). The implemented mapping approach identifies genes that are associated with one or multiple chromosomal breakpoint intervals.

Cohort-based breakpoint statistics: breakpoint and gene level
Identification of statistically recurrent breakpoint events across all samples can be performed on both chromosomal location- and gene-level. As features, i.e. microarray probes or bins of WGS copy number data, are (nearly) equally distributed over the genome, we assume that the null-probability for breakpoint occurrence is equal for all individual candidate breakpoints (features). It differs per sample, though, and equals \( p_s = N_s/N \), where \( N \) is the number of probes, and \( N_s \) the total number of breakpoint for samples. The test statistic is \( T_p \), the total number of breakpoints for probe \( p \) across all samples. Then, under the null-hypothesis, \( T_p \) is simply a sum of independent Bernoulli \( p_s \) random variables, the null-distribution of which is the same for all probes. It is quickly computed by using probability generating functions, giving also the p-values for any observed value of \( T_p \).
The probe-based statistical analysis uses Benjamini-Hochberg false discovery rate (FDR) correction for multiple testing. For the intended use at gene level, a more advanced statistical null-model is required. For the gene level, the null-probability for a breakpoint to occur within an individual gene, depends on 1) the length of the gene, 2) the number of gene-associated features and 3) the number of breakpoints in the entire tumor profile for the specific sample. Therefore, at gene-level, we apply a linear regression-based correction for covariates. These regression-estimates are then used as gene- and sample-specific breakpoint null-probabilities ($p_{null}$). The test statistic remains the same, and so does the null-distribution computation, although it has to be repeated for each gene now. Finally, the Gilbert FDR correction that accounts for discreteness in the null-distribution is applied in this analysis to determine significance of recurrent breakpoint genes. Commands and example workflow can be found in Bioconductor vignette and manual.

Use case

Identification of recurrent breakpoint genes in advanced colorectal cancers

We applied our method to 352 high-resolution array-CGH samples from a series of advanced colorectal cancers following CNA detection using ‘CGHcall’. Array-CGH data are available in the Gene Expression Omnibus database under accession number GSE63216 (www.ncbi.nlm.nih.gov/projects/geo). We selected for the CNA-associated breakpoints (setting: ‘CNA-associated’), used gene annotations from ensembl (human genome NCBI build36/hg18, release 54) and applied the dedicated Benjamini-Hochberg-type FDR correction (setting: ‘Gilbert’), for recurrent breakpoint gene identification. A total of 748 genes appeared to be recurrently affected by chromosomal breaks (FDR<0.1). Breakpoint frequencies of chromosome 20p are visualized with the built-in plot function (Figure 1; see Bioconductor vignette and manual).
manual for further details about this function). Interestingly, patient stratification based on recurrent gene breakpoints and well-known point mutations by propagation to the predefined STRING human protein interaction network revealed one CRC subtype with very poor prognosis, which supported clinical relevance of this class of somatic aberrations in advanced colorectal cancers.\(^5\)

**Conclusion**

Genome instability including numerical and structural somatic chromosomal aberrations is a hallmark of cancer. Several tools are available that focus on detection of numerical aberrations of large chromosome segments. The R-package ‘GeneBreak’ extracts additional information from CNA data. ‘GeneBreak’ provides an easy-to-use algorithm, which handles identification of genomic breakpoint locations, mapping of breakpoints to genes and includes a comprehensive statistical approach to reveal recurrent breakpoint genes from series of tumor samples. Therefore, ‘GeneBreak’ can be applied to detect CNA-associated chromosomal breaks in individual tumor samples and facilitates detection of recurrent breakpoint genes across multiple tumor samples.

**Data and software availability**

Publicly available copy number data used for the use case is deposited at Gene Expression Omnibus database under accession number GSE63216 (https://protect-eu.mimecast.com/s/6LQhBmNGvCG).


**Supplementary material**

GeneBreak vignette.
Click here to access the data.

Click here to access the data.

**References**


7. Edwards PA: Fusion genes and chromosome translocations in the common


Data Source
Open Peer Review

Current Referee Status: ✔️ ✔️

Version 2

Referee Report 06 July 2017
doi:10.5256/f1000research.12600.r24055

Angel Rubio
Group of Bioinformatics, TECNUN, University of Navarra, San Sebastian, Spain

The authors provided an brief explanation of the statistics involved in the detection of recurrent copy number changes. I would have liked a slightly deeper description but, for most users the explanation is sufficient and give an idea of the methodology.

I think that this paper is ready.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 13 February 2017
doi:10.5256/f1000research.9967.r18598

Angel Rubio
Group of Bioinformatics, TECNUN, University of Navarra, San Sebastian, Spain

The paper shows an inspiring vision of the copy number changes in the genome focusing on the "changes" more than on the levels of change. The underlying reasoning is that a copy number change, if occurs within the loci occupied by a gene, implies an alteration in the coding sequence of the gene.

In addition, it is shown that these changes occur recurrently, i.e. the loci where the copy number changes tend to be similar in different samples with the same type of cancer.

The methodology has been uploaded to Bioconductor. The stringent quality checks of Bioconductor guarantees the availability for different platforms and, in fact, the vignette is easy to follow and use.

My main concern with this paper is the (lack of) description of the statistical method to state the
recurrence of the copy number changes. Within the methods section is only stated that there are two methods (genome location and gene-level) but the differences between them or the underlying statistical model is missing.

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

**Tobias Marschall**  
Center for Bioinformatics, Max-Planck Institute for Infomatics, Saarbrücken, Germany

GeneBreak is an R package to help identifying recurrent breakpoints of copy number variants (CNVs). While the offered analyses are straightforward from a methodological point of view, this package can be valuable in practice, providing an easy and reproducible way to conduct such analyses. I appreciate that it is available from bioconductor (and hence easily installable), openly developed on github, and archived on zenodo.

I merely have some minor suggestions for improvements:

- When trying to use the package, I couldn't open the example data (got a "data set 'copynumber.data.chr20' not found" error). Could you verify that it's available?

- First sentence: SNV commonly means "single nucleotide variation" (not "small").

- P1, L9: "Recently, ..." Of course what you consider "recent" is a matter of taste, but here you are citing a review paper from 2007. I wouldn't call this recent.

- P1, Methods, L3: "... and copy number detection algorithm" Either explain what exactly you mean here, or remove.

- P1, paragraph "Due to the typical granularity [...], in fact represent a chromosomal interval." I can guess what you mean here, but writing this more clearly would be good.

- P1, "This method assumes the same permutation null- distribution for all candidate breakpoint events for the analysis of breakpoints at the level of genomic location." Could you describe in more detail how the null distribution is obtained?

- P1, "In addition, a more comprehensive and powerful dedicated Benjamini-Hochberg FDR correction that accounts for discreteness in the null-distribution is supplied." The Benjamini-Hochberg procedure is a well defined statistical method. I would rephrase the respective sentence(s) to explicitly say that you are talking about Gilbert's method.

**Competing Interests:** No competing interests were disclosed.
I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reader Comment 29 May 2017

Evert van den Broek, Netherlands Cancer Institute / UMCG, Netherlands

We thank the reviewer for careful evaluation of our work and providing helpful recommendations. As suggested by the reviewer, we rephrased some sentences and provided a more detailed description of the used statistics. With respect to the error by loading the example data of GeneBreak, we verified availability of the example data on different computers with different operating systems (MacOS and Linux) on which we installed the GeneBreak package from Bioconductor and CGHcall with all dependencies. The data ('copynumber.data.chr20.rda') was also available in the ‘data’ directory that was retrieved from Bioconductor. The exact code we used is provided by https://www.bioconductor.org/packages/release/bioc/vignettes/GeneBreak/inst/doc/GeneBreak.R.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com