Bioinformatics, 38(7), 2022, 2036–2038
https://doi.org/10.1093/bioinformatics/btac037
Advance Access Publication Date: 31 January 2022
Applications Note

Genome analysis

GenomeTornadoPlot: a novel R package for CNV visualization and focality analysis

Chen Hong [1,2,3], **Robin Thiele** [1,2] **and Lars Feuerbach** [1,*]

[1]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany, [2]Faculty of Biosciences, Heidelberg University, Heidelberg 69120, Germany and [3]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany

*To whom correspondence should be addressed.

Associate Editor: Can Alkan

Received on July 30, 2021; revised on December 21, 2021; editorial decision on December 23, 2021

Abstract

Motivation: Analysis of focal copy number variations (CNVs) is highly relevant for cancer research, as they pinpoint driver genes. More specifically, due to selective pressure oncogenes and tumor suppressor genes are more often affected by these events than neighboring passengers. In cases where multiple candidates co-reside in a genomic locus, careful comparison is required to either identify multigenic minimally deleted regions of synergistic co-mutations, or the true single driver gene. The study of focal CNVs in large cancer genome cohorts requires specialized visualization and statistical analysis.

Results: We developed the GenomeTornadoPlot R-package which generates gene-centric visualizations of CNV types, locations and lengths from cohortwise NGS data. Furthermore, the software enables the pairwise comparison of proximate genes to identify co-mutation patterns or driver-passenger hierarchies. The visual examination provided by GenomeTornadoPlot is further supported by adaptable local and global focality scoring. Integrated into the GenomeTornadoPlot R-Package is the comprehensive PCAWG database of CNVs, comprising 2976 cancer genome entities from 46 cohorts of the Pan-cancer Analysis of Whole Genomes project. The GenomeTornadoPlot R-package can be used to perform exploratory or hypothesis-driven analyses on the basis of the PCAWG data or in combination with data provided by the user.

Availability and implementation: GenomeTornadoPlot is written in R script and released via github: <https://github.com/chenhong-dkfz/GenomeTornadoPlot/>. The package is under the license of GPL-3.0.

Contact: l.feuerbach@dkfz-heidelberg.de

1 Background

Copy number variations (CNVs) play a central role in the etiology of many cancer types (Zhang *et al.*, 2016). They can lead to amplifications, loss of heterozygosity or complete loss of genetic information, and thus, enact gain-of-expression in oncogenes, as well as, loss-of-function mutation in tumor suppressors.

In the analysis of CNVs, the notion of focality has become an important criterion to distinguish actual tumor driving lesions from functionally irrelevant alterations (Zhang *et al.*, 2016). The current working definition defines focal CNVs as events which are below 1 or 3 Mb in length (Bierkens *et al.*, 2013; Bignell *et al.*, 2010). These events have for instance been described for lung- (Bierkens *et al.*, 2013), colon- and breast cancer (Bierkens *et al.*, 2013; Garnis *et al.*, 2006), in which well-established cancer driver genes like *PTEN*, CDKN2A and RB1 are observed to be linked to recurrent focal CNVs (Garnis *et al.*, 2006; Leary *et al.*, 2008).

Cooperative cancer genome sequencing projects, such as TCGA and ICGC, provide massive cohortwise genomic datasets (Tate *et al.*, 2019; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). These datasets provide an unprecedented statistical power for basepair-resolution *in silico* CNV analysis. Here, we provide a user-friendly software package implemented in R for the visualization and quantitative analysis of focal CNVs in large pan-cancer genome cohorts.

2 Implementation

2.1 Data input

CNV information is provided in an extended BED format that includes chromosome identifier, start position, end position, a score that indicates the CNV ploidy as well as additional information on the cohorts-of-origin and patient ID. Data can be either provided exclusively for the gene or gene-pair of interest, or as a combination of

a genome-wide dataset and a gene model. In the second case, the gene model is used to generate a database of gene-specific CNVs.

We also provided preprocessed CNV data from 2976 whole-genome cancer genomes of the 46 cancer cohorts from the PCAWG project (Harrow *et al.*, 2012; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), which is divided in one RData file per chromosome for direct import to R. The data can be downloaded here: <https://github.com/chenhong-dkfz/GenomeTornadoPlot-files>.

2.2 Visualization

GenomeTornadoPlot provides multiple visualization modes, which can display the CNVs of any single gene or gene pair in one chromosome as a specialized ideogram. For any single gene of interest, the corresponding chromosome is displayed as an ideogram, and variation events will be aligned beside it in regard to the regions they span. By default, CNVs are ordered by length in ascending order, creating in case of focally affected genes a characteristic tornado shape. Users can also sort CNVs by variation type, ploidy or cohort by setting the appropriate parameters. For the genes whose functions are complex in different cancer types, GenomeTornadoPlot allows to display the loss-of-function and gain-of-function variations at each side of the chromosome ideogram. For the convenience of reading, our package allows users to zoom and rotate the plots through parameter settings.

In case of gene pairs, GenomeTornadoPlot will automatically generate a 'twin plot' for both genes. CNV events of each gene are sorted on each side of the chromosome ideogram. Meanwhile, a 'mixed plot' will be generated in which common CNVs are displayed differently, to discriminate co-mutation patterns from passenger effects. Moreover, the focality scores and distribution entropy are annotated beside the tornado plot. All plots can be exported as vector-graphics files or stored as R objects.

2.3 Focality score and entropy

Driver genes show distinct patterns for which either focal copy number gains or losses are enriched. Capturing this feature mathematically in the form of a focality score allows to screen and compare a large number of genes, and only generate GenomicTornadoPlots for the most promising candidates. The default focality score is defined as:

$$S = \sum_{i=1}^{m} (\log(L_{\max} - L_i)) \quad (1)$$

where m is the total number of focal variation events and the capping value $L_{\max}$ is the length of the longest CNV event that still is counted as focal and used to exclude large events such as chromosomal arm losses. Besides this default score, our package supports further scoring methods including user-defined schemas. A detailed discussion of all provided focality scores can be found at the repository page.

To discriminate pan-cancer driver genes from cancer-subtype specific events, the package computes the Shannon-entropy using the cohorts-of-origin as individual clusters. To improve the interpretability of the entropy score, users may consider grouping cohorts into meaningful metacohorts that, for instance, contain all cases belonging to a certain cancer-subtype.

2.4 ShinyApp

GenomeTornadoPlot package provides a shiny app, which can be launched from the R console. With this interface users can input their CNV files and modify the parameters. The tornado plots will be displayed in the window and can also be exported as image files.

3 Applications

The following four examples were generated with the PCAWG dataset and demonstrate the core functionality of the GenomeTornadoPlot package (Fig. 1).

In addition to using the comprehensive PCAWG dataset, it is possible to use self-produced data to create GenomicTornadoPlots. These smaller datasets can either be integrated with the PCAWG database or visualized independently.

4 Conclusion

We developed an R-package called GenomeTornadoPlot for visualizing and performing statistical analysis of focal CNVs for multiple large cancer cohorts. The integrated PCAWG database allows users to perform exploratory or hypothesis-driven analyses without additional data or to integrate self-produced small datasets to address biologically relevant questions *in silico*. For easy application, the GenomeTornadoPlot has an implemented ShinyApp. GenomeTornadoPlot is a helpful tool to identify cancer-related genes and explore the selection-driven accumulation of focal CNVs in cancer.

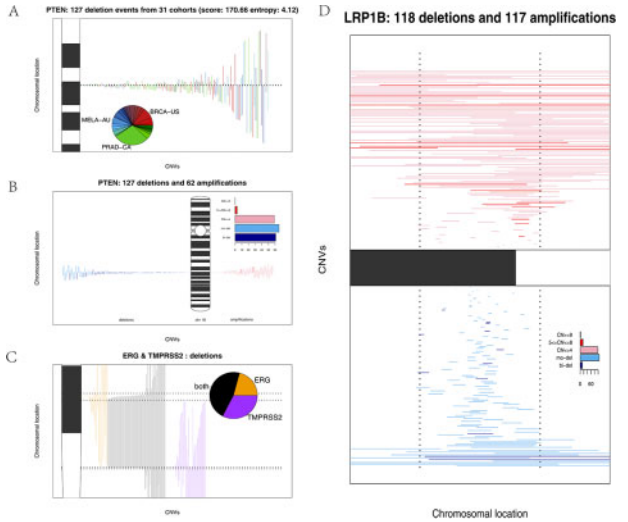

Fig. 1. Examples of broadly acknowledged cancer-related CNVs illustrated via TornadoPlots with the provided PCAWG dataset. (**A, B**) CNVs of *PTEN* throughout cohorts—*PTEN* plays important roles as a tumor suppressor gene in many cancer entities (Chu and Tarnawski, 2004). (A) The characteristic tornado shape illustrates the enrichment of focal deletions events of the *PTEN* locus. Coloration by cohorts-of-origin and the respective pie chart, support the pan-cancer nature of this event, but also illustrate an enrichment of focal events in prostate cancer. (B) Comparison of different levels of deletions (blue shades, left side) and amplifications (red shades, right side) verifies the abundance of focal, bi-allelic deletions as a typical tumor-suppressor signature. (**C**) Example of zoomed-in mixed plot—detection of minimally deleted co-mutation patterns in *ERG* and *TMPRSS2*: *ERG* is known as a proto-oncogene and its overexpression is related to cancer development. The promoter of *TMPRSS2* and the gene body of *ERG* are both located on chromosome 21 and frequently form a fusion gene in prostate cancer by genomic deletion (Liu *et al.*, 2001; Weischenfeldt *et al.*, 2013). This results in the oncogenic upregulation of *ERG*. In the mixed plot, co-mutation patterns dominate over single-locus events, implying a synergistic effect rather than two individual driver genes. CNV deletions occurring in *ERG* and *TMPRSS2* with nearly equal length further suggest positive selection pressure on a gain-of-function event. (**D**) CNVs of *LRP1B* throughout cohort—the zoomed-in and vertical arranged tornado shape plots shows the intragenic deletions and amplifications in *LRP1B*. *LRP1B* is a putative tumor suppressor. It is functionally related to clearance of extracellular ligands and signal transduction (Liu *et al.*, 2001). In the magnified tornado plot dashed lines illustrate the gene boundaries. Remarkably, many intragenic deletions can be observed which only affect particular regions in the gene (blue shades, bottom). Furthermore, also several amplifications and amplification breakpoints are intragenic, and thus, potentially truncate the gene

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [Research Consortium FOR 2674 to L.F.], by the German Cancer Consortium (DKTK).Furthermore, the project received funding from the European Union's Horizon 2020 research and innovation programme and the Canadian Institutes of Health Research under the grant agreement No 825325.

Conflict of Interest: none declared.

References

Bierkens,M. *et al.* (2013) Focal aberrations indicate EYA2 and hsa-miR-375 as oncogene and tumor suppressor in cervical carcinogenesis. *Genes Chromosomes Cancer*, **52**, 56–68.

Bignell,G.R. *et al.* (2010) Signatures of mutation and selection in the cancer genome. *Nature*, **463**, 893–898.

Chu,E.C. and Tarnawski,A.S. (2004) PTEN regulatory functions in tumor suppression and cell biology. *Med. Sci. Monitor Int. Med. J. Exp. Clin. Res.*, **10**, RA235–RA241.

Garnis,C. *et al.* (2006) High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *Int. J. Cancer*, **118**, 1556–1564.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

Leary,R.J. *et al.* (2008) Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl. Acad. Sci. USA*, **105**, 16224–16229.

Liu,C.X. *et al.* (2001) The putative tumor suppressor LRP1B, a novel member of the low density lipoprotein (LDL) receptor family, exhibits both overlapping and distinct properties with the LDL receptor-related protein. *J. Biol. Chem.*, **276**, 28889–28896.

Tate,J.G. *et al.* (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.*, **47**, D941–D947.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.

Weischenfeldt,J. *et al.* (2013) Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell*, **23**, 159–170.

Zhang,L. *et al.* (2016) Identification of recurrent focal copy number variations and their putative targeted driver genes in ovarian cancer. *BMC Bioinformatics*, **17**, 222.