


SOFTWARE

Open Access



kataegis: an R package for identification and visualization of the genomic localized hypermutation regions using high-throughput sequencing

Xue Lin^{1*†} , Yingying Hua^{2†}, Shuanglin Gu^{3†}, Li Lv³, Xingyu Li³, Pin Chen³, Peng Dai¹, Yunyun Hu³, Anna Liu³ and Jian Li^{3*} 

Abstract

Background: Genomic localized hypermutation regions were found in cancers, which were reported to be related to the prognosis of cancers. This genomic localized hypermutation is quite different from the usual somatic mutations in the frequency of occurrence and genomic density. It is like a mutations “violent storm”, which is just what the Greek word “kataegis” means.

Results: There are needs for a light-weighted and simple-to-use toolkit to identify and visualize the localized hypermutation regions in genome. Thus we developed the R package “kataegis” to meet these needs. The package used only three steps to identify the genomic hypermutation regions, i.e., i) read in the variation files in standard formats; ii) calculate the inter-mutational distances; iii) identify the hypermutation regions with appropriate parameters, and finally one step to visualize the nucleotide contents and spectra of both the foci and flanking regions, and the genomic landscape of these regions.

Conclusions: The kataegis package is available on Bioconductor/Github (<https://github.com/flosalbizziae/kataegis>), which provides a light-weighted and simple-to-use toolkit for quickly identifying and visualizing the genomic hypermutation regions.

Keywords: Kataegis, Visualization, High-throughput sequencing, R

Background

There are numerous somatic mutations in human genomes, especially in cancer genomes. Many exogenous and endogenous factors are known reasons for the occurrence of the somatic mutations, like the ultra-violet

lights, chemical mutagens, and DNA repair, etc. [1] And different mutational combinations are usually generated by different mutational processes, e.g., C > T and CC > TT transitions are common in ultra-violet light related skin cancers [2], and G > T in aflatoxin-B1 associated hepatocellular carcinomas [3]. The mutational combinations are called “signatures” of the mutational processes. These signatures were firstly analyzed in a small number of frequently mutated cancer genes like the TP53, however with the rapid development of the massively parallel sequencing technology, it has overcome the scale limitations, thus tens of thousand of variations can be

* Correspondence: xue.lin@njmu.edu.cn; jjanli2014@seu.edu.cn

Xue Lin, Yingying Hua and Shuanglin Gu contribute equally.

¹Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, 211166 Nanjing, People’s Republic of China

³Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, 210096 Nanjing, People’s Republic of China

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

identified in a cancer genome. Intriguingly, when screening the mutations and extracting the signatures, the researchers investigated a possibility of regional clustering of mutations by constructing “rainfall plots”, which present the inter-mutational distances between each mutation [4]. These regional clustering mutations are “hot-spots” of mutations in cancer genomes, i.e., regional hypermutation, which is just like a “violent storm” of the mutations in the cancer genomes, thus this phenomena is named with a Greek word “kataegis” which exactly means the same.

The kataegic foci were earlier investigated in a study of 21 breast cancers [4, 5]. Later with 7,042 primary cancers of 30 different classes analyzed, cancers of breast (67 of 119), pancreas (11 of 15), lung (20 of 24), liver (15 of 88), medulloblastomas (2 of 100), CLL (Chronic Lymphocytic Leukemia) (15 of 28), B-cell lymphomas (21 of 24) and acute lymphoblastic leukaemia (1 of 1) showed occasional (< 10), small (< 20 mutations) foci of kataegis [6]. The mechanism of the generation of the kataegic foci is not fully clear. But the kataegis foci were usually found co-locating with the genomic rearrangements. There was evidence in yeast and human that the clustered mutations can arise from damaged long single-strand DNA regions [7], the chromothripsis and kataegis were induced by telomere crisis [8], and the AID/APO-BEC editing deaminases were involved [9–11]. Further analysis of the kataegis expression signature found that, in breast cancer it is associated with late onset, better

prognosis and higher HER2 levels [12]. To decipher the role of the kataegis in the cancer genomes, there are needs for a light-weighted and simple-to-use toolkit to identify and visualize the localized hypermutation regions in genome. Thus we developed the R package “kataegis” to meet these needs.

Implementation

This package was coded in R language with RStudio version 1.2.5042 built on R version 4.0 on macOS Mojave version 10.14. It depends on the R core packages grDevices, graphics, stats, and utils, and is maintained and released through the Bioconductor project with an Artistic License.

The kataegis package provides a four-step workflow for localized hypermutation regions identification and visualization: data read in, inter-mutational distances calculation, kataegis identification, and visualization (Fig. 1).

The readVCF() and readMAF() functions can read in the standard Variant Call Format (VCF) (<http://github.com/samtools/hts-specs/blob/master/VCFv4.2.pdf>) and Mutation Annotation Format (MAF) (http://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/) files respectively. The VCF and MAF formats are both most commonly used file formats for storing variants information from high-throughput sequencing, e.g., whole genome sequencing. Both of these formats have standard specifications, and there are also mature tools to

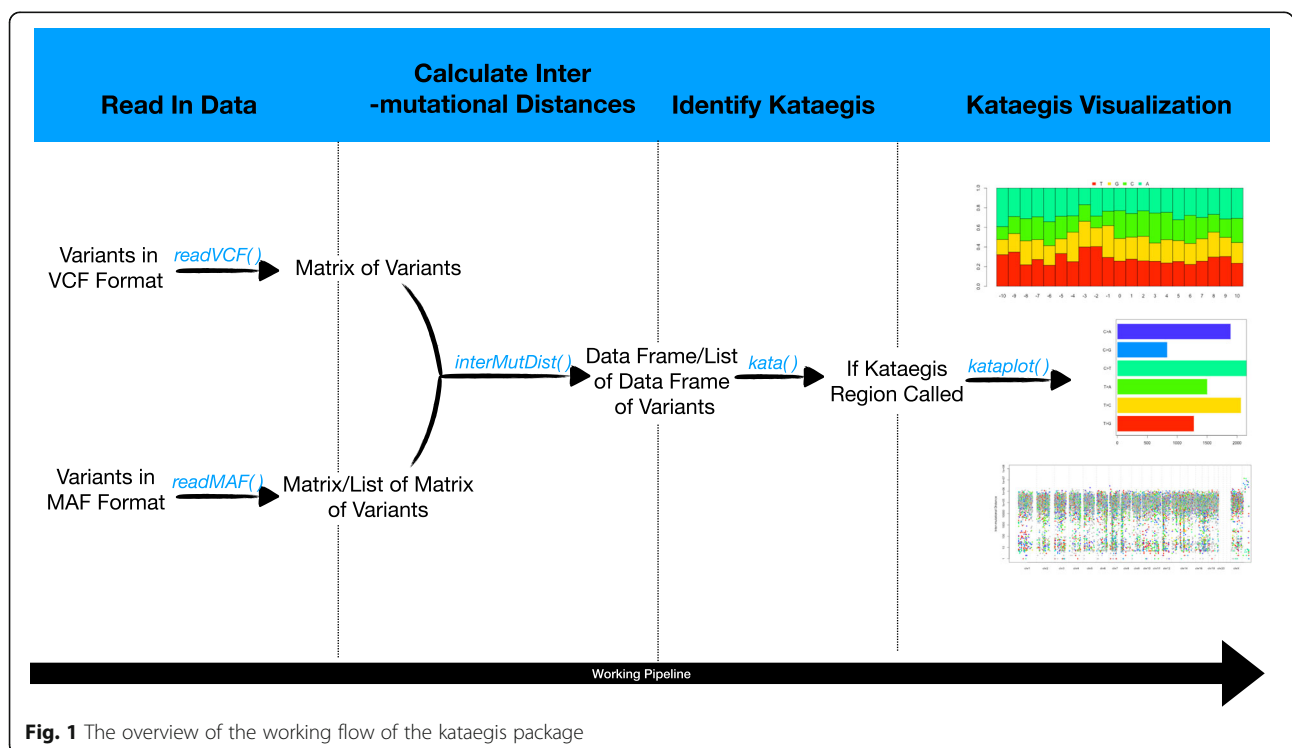


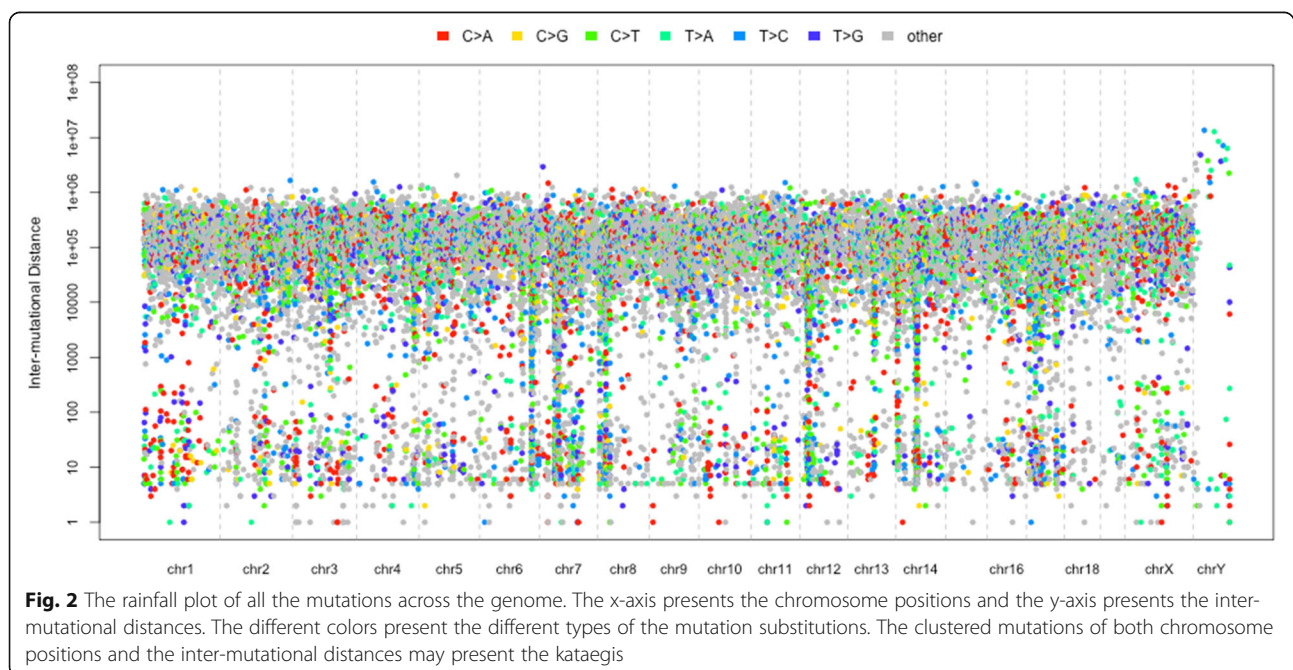
Fig. 1 The overview of the working flow of the kataegis package

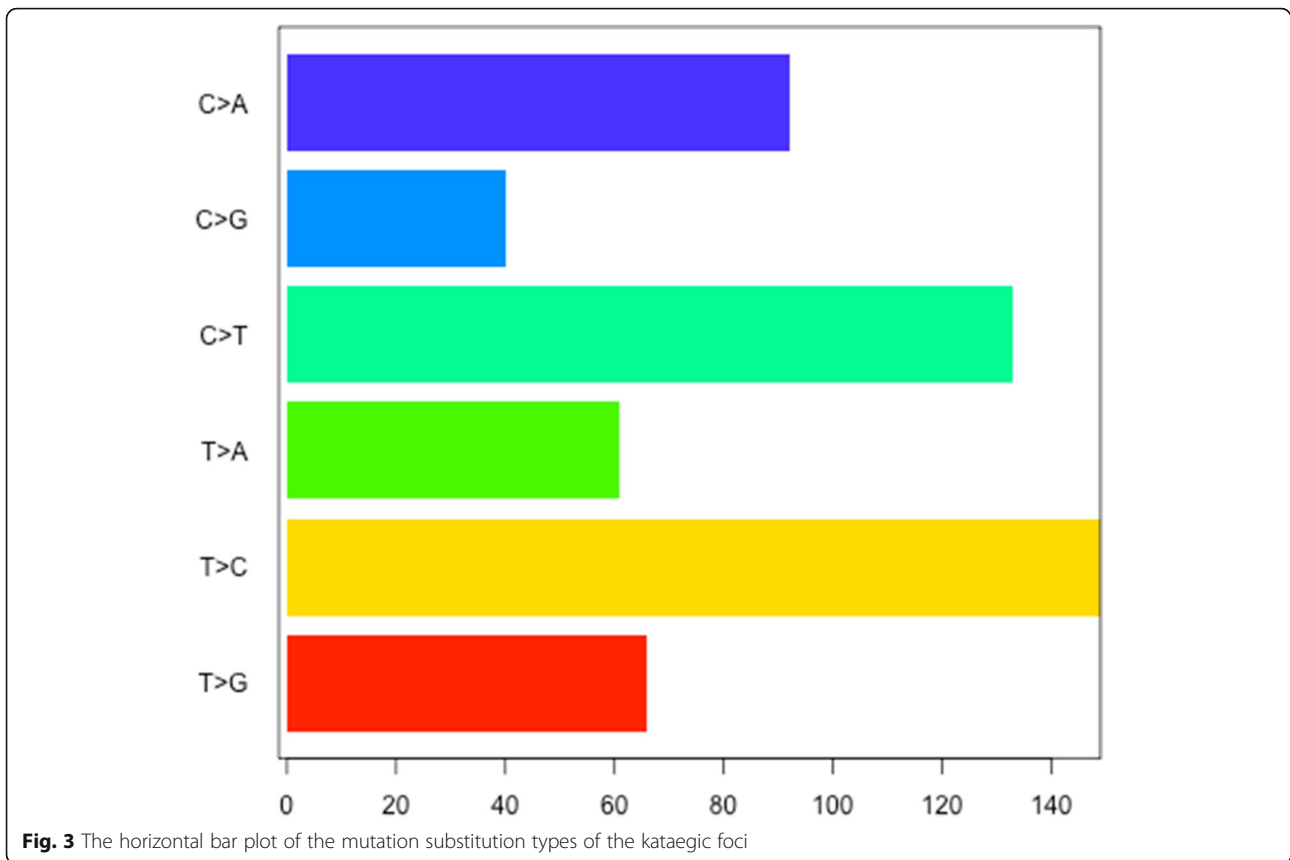
perform filtering and format conversion between them. The readVCF() function will read in the VCF format and suffixed files and do a crude filtering according to the VCF “FILTER” field. As the MAF file can hold the mutations’ annotation data of several samples, which is widely used by important bioinformatic databases like the The Cancer Genome Atlas (TCGA), etc., so we provide a function readMAF() to read in the MAF format and suffixed files with the samples merged or separated. If the user chooses to read in the MAF file with the samples separated, then the variants will be read in to a list of matrix, each matrix is named after the sample’s ID. The crude filtering also works for the readMAF() function. The package contains two simulated data sets, of which the VCF was generated from the mouse model of small cell lung cancer GSE149444_17686R (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149444>) and the MAF was generated from the human adenoid cystic carcinoma (http://www.cbiportal.org/study/summary?id=acyc_mskcc_2013) [13].

If the data is read in without any errors, the second step is to calculate the inter-mutational distances between the mutations. The variants will firstly be sorted according to the genomic coordinates for each chromosome, and then the distances will be calculated between the neighboring variants. For a list of separated samples, the samples with too few variants for calculating the inter-mutational distances will be abandoned. And a warning of this situation and the sample IDs will arise. This step will produce a data frame containing the information of the chromosomes, variants’ locations, and the inter-mutational distances.

With the previous two steps, the data is ready for calling the localized hypermutation regions. The localized hypermutation regions are mutations “hotspot” regions, which are defined as more than a certain number of mutations in a range of the genome. It was reported as more than five [9] or six mutations in a range of 1000 bp of the human genome [6]. With this concept, it is reasonable to segment the genomes with the mutations. We used a segmentation method based on the Piecewise Constant Fitting (PCF) algorithm [14]. The segmentations with the information of the number of mutations and its average inter-mutational distances are reported, and the genomic coordinates are also produced. Thus it’s simple for users to filter the segments with the threshold of the mutation number and the average inter-mutational distance. The kata() function will automatically perform the previous jobs.

After all the first three steps finished, and the localized hypermutation regions are identified successfully, it comes to the last step to visualize the regions. Researchers are usually interested in the global landscape of the distribution of the regions in the genome (Fig. 2), and also the nucleotides content (Fig. 3) and spectra (Fig. 4) of the foci and flanking regions of the localized hypermutation regions. Here we provide a kataplot() function, it will take in the data produced by the first three steps and produce these plots automatically. The users only have to control which type of plots will be produced, the size and format of the plots, and the name of the plots as well. The global landscape of the distribution of the regions can also become the landscape of one or several certain chromosomes other than the whole genome (Fig. 5).

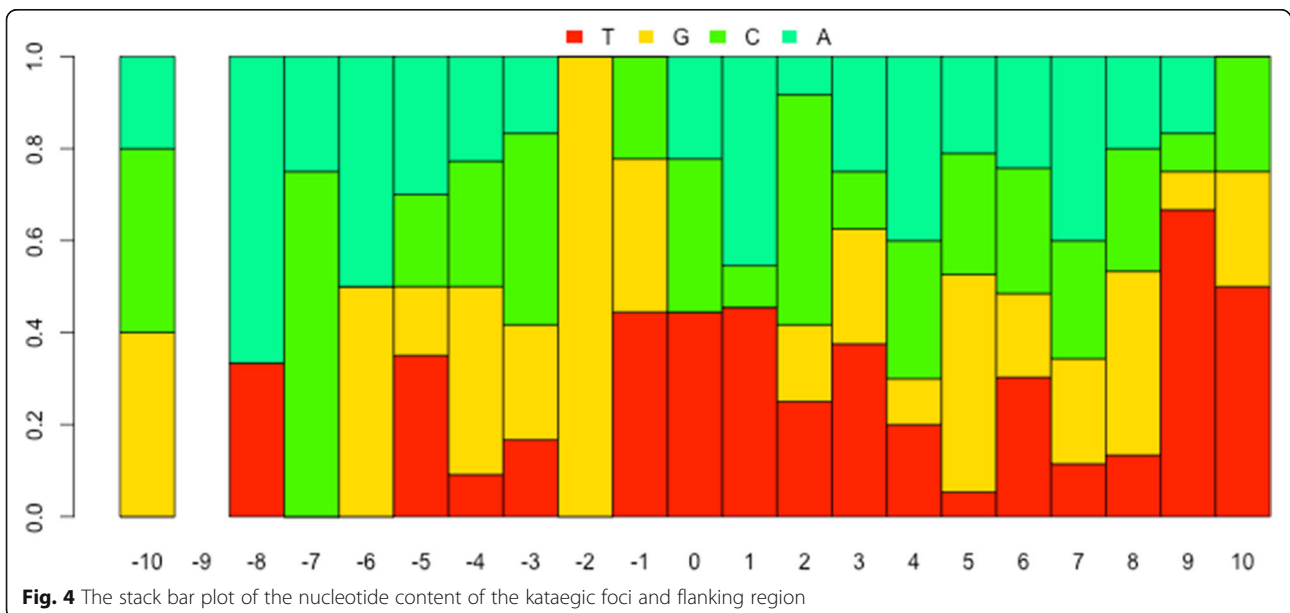


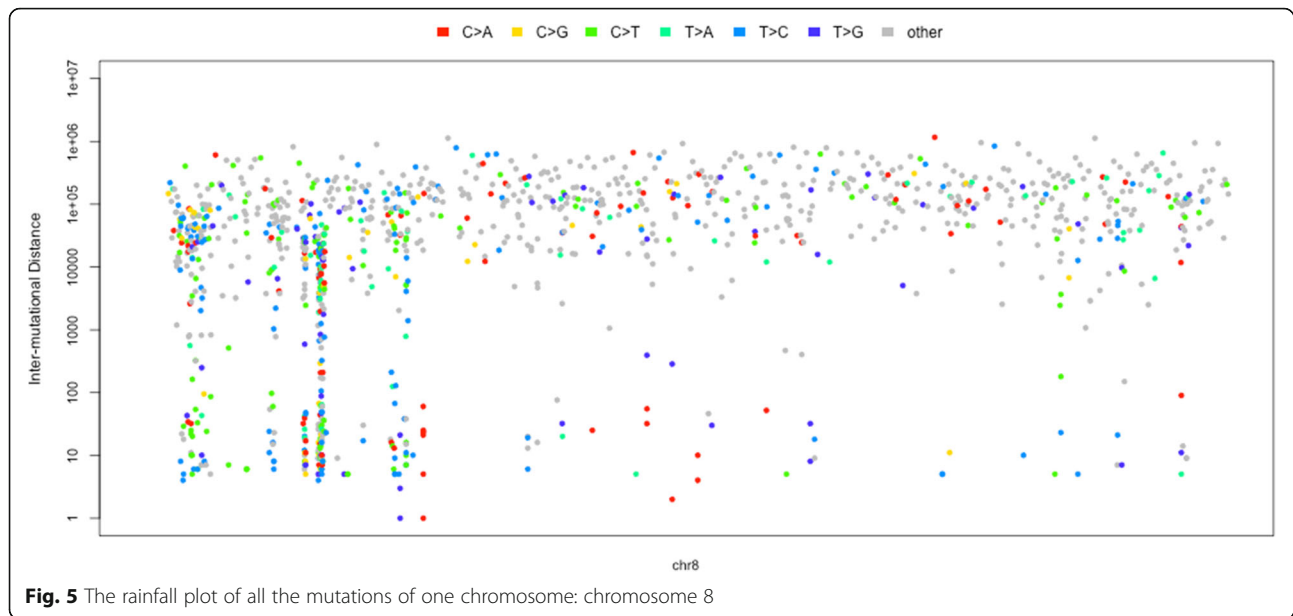


Results and discussion

We have prepared a set of simulated VCF files by using the script coded by ourselves, which is available on <https://github.com/flosalbizziae/BioinfoCollects/blob/main/katasim.py>. The script can mutate the DNA

sequence and provide the users an option `-k` to control whether or not to form kataegis during the mutating of the sequence. When the `-k` option is set to 1, the script will form kataegis in the sequence by mutating more than 6 mutations within 1 kb, which is just the same for





defining the kataegis, and the kataegis will be randomly distributed on the sequence. When the $-k$ is set to 0, the script will mutate the sequence completely randomly, and at the same time ensure that the mutations are not too close to each other to trigger the definition of kataegis. With this script, a hundred VCF files were generated, giving the same tumor mutational burden (TMB) on the same sequence, the chr1 of human genome version hg38 downloaded from UCSC Genome Browser web site. One half of the simulated VCFs are with kataegis, and the other half are without kataegis. Analyzing with our package, we obtained good sensitivity and specificity. We detected kataegis from all the fifty VCFs generated with kataegis, and on the other hand, we detected no kataegis from all the other fifty VCFs generated without kataegis. The sensitivity of the detection of kataegis largely depends on the quality of the definition of the mutations. As the package will segment the sequence according to the mutations and the inter-mutational distances, if the mutations were not detected in the variants calling analysis, then kataegis cannot be detected consequently. Thus our stringent criteria of the segmentation may under-represent the kataegis in the analysis.

Conclusions

In conclusion, we have provided a light-weighted, simple-to-use, and relatively flexible toolkit for identifying and visualizing the localized hypermutation regions using high-throughput sequencing data as an R package. This toolkit uses a straight-forward and statistical strategy to identify the kataegis regions, which provides the users a convenient usage experience and efficient researching tool for variants understanding and further study of integrating the mutations with the clinical outcomes in a new dimension.

Availability and requirements

Project name: kataegis.

Project page: <https://github.com/flosalbizziae/kataegis>.

Operating systems: platform independent.

Programming language: R.

Other requirements: R 4.0 or higher.

License: Artistic-2.0.

Any restriction to use by non-academics: Artistic-2.0 license needed.

Abbreviations

CLL: Chronic Lymphocytic Leukemia; VCF: Variant Call Format; MAF: Mutation Annotation Format; TCGA: The Cancer Genome Atlas; PCF: Piecewise Constant Fitting; TMB: Tumor Mutational Burden

Acknowledgements

Not applicable.

Authors' contributions

XL1 developed the package, and all the documentation of the package; YH1 contributed to the designing and debugging of the package, and provided English writing supports of the paper; SG, YH2 and AL assisted the development of software and performed the simulation for the package coding and testing; LL, XL2, PC, PD assisted the preparation of the simulation data for the package coding and testing and contributed to the discussion of the methodology of this work; JL conceived and guided this project and corresponding to this paper. The author(s) read and approved the final manuscript.

Funding

National Natural Science Foundation of China (31900473 and 31871322) Natural Science Foundation of the Jiangsu Higher Education Institutions of China (18KJB180015)

Availability of data and materials

The system data for building the package, including hg17, hg18, hg19, hg38, mm7, mm8, mm9, and mm10, was downloaded from the UCSC genome browser using its tool Table browser (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1097046771_ILSM8fQrnzE9fECLP2zBMeNfPZ6) with the options group setting to "Mapping and Sequencing" and track setting to "Chromosome Band (Ideogram)". The VCF example file was generated from

the GEO data set GSE149444 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149444>), and the MAF example file was generated from the human adenoid cystic carcinoma data of TCGA database, which was downloaded from http://www.cbioportal.org/study/summary?id=acyc_mskcc_2013. The VCF files for simulation was generated from the hg38 genome data, which was downloaded from the UCSC genome browser using its tool Downloads (<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/>).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, 211166 Nanjing, People's Republic of China. ²Traditional Chinese Medicine Department, Fuxing Hospital, Capital Medical University, 100038 Beijing, People's Republic of China. ³Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, 210096 Nanjing, People's Republic of China.

Received: 8 December 2020 Accepted: 10 May 2021

Published online: 12 June 2021

References

- Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;446:153–8.
- Pfeifer GP, You Y, Besaratinia A. Mutations induced by ultraviolet light. *Mutat Res*. 2005;571:19–31.
- Macé K, Aguilar F, Wang JS, Vautravers P, Gómez-Lechón M, Gonzalez FJ, et al. Aflatoxin B1-induced DNA adduct formation and p53 mutations in CYP450-expressing human liver cell lines. *Carcinogenesis*. 1997;18:1291–7.
- Nik-zainal S, Alexandrov LB, Wedge DC, Loo P, Van, Greenman CD, Raine K, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*. 2012;149:979–93.
- Nik-zainal S, Loo P, Van, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The Life History of 21 Breast Cancers. *Cell*. 2012;149:994–1007.
- Alexandrov LB, Nik-zainal S, Wedge DC, Aparicio SAJR. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415–21.
- Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, et al. Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Mol Cell*. 2012;46:424–35. doi:<https://doi.org/10.1016/j.molcel.2012.03.030>.
- Maciejowski J, Li Y, Bosco N, Campbell PJ, Lange T, De, Trust W, et al. Chromothripsis and kataegis induced by telomere crisis. *Cell*. 2015;163:1641–54.
- Lada AG, Dhar A, Boissy RJ, Hirano M, Rubel AA, Rogozin IB, et al. AID / APOBEC cytosine deaminase induces genome-wide kataegis. *Biol Direct*. 2012;7:1–7.
- Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. 2015;47:1067–72.
- Taylor BJM, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife*. 2013;2013:1–14.
- Antonio MD, Tamayo P, Jill P, Frazer KA, Antonio MD, Tamayo P, et al. Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Article Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels. *CellReports*. 2016;16:672–83. doi:<https://doi.org/10.1016/j.celrep.2016.06.026>.
- Ho AS, Kannan K, Roy DM, Morris LGT, Garly I, Ramaswami D, et al. The Mutational Landscape of Adenoid Cystic Carcinoma. *Nat Genet*. 2013;45:791–8.
- Nilsen G, Liestøl K, Loo P, Van, Kristian H, Vollan M, Eide MB, et al. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genom*. 2012;13:1–16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

