Katdetectr: utilising unsupervised changepoint analysis for robust kataegis detection

Daan M. Hazelaar[1,2,†], Job van Riet[1-3,†], Youri Hoogstrate[1,4], Martijn P. Lolkema[2] and Harmen J. G. van de Werken[1,3,5]

[1]Cancer Computational Biology Centre, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands, [2]Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands, [3]Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands, [4]Department of Neurology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands, [5]Department of Immunology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands.

[†]Shared first-author

All Authors have seen and approved this manuscript.

Abstract

Motivation:
Kataegis refers to the occurrence of regional hypermutation in cancer genomes and is a phenomenon that has been observed in a wide range of malignancies. Robust detection of kataegis is necessary to advance research regarding the origins and clinical impact of kataegis. Multiple kataegis detection packages are publicly available; however, the performance of their respective approaches have not been evaluated extensively. Here, we introduce katdetectr, an R-based, open-source, computationally fast, and robust package for the detection, characterisation and visualisation of kataegis.

Results:
The performance of katdetectr and five publicly available packages for kataegis detection were evaluated using an in-house generated synthetic dataset and an *a priori* labelled pan-cancer dataset of whole genome sequenced malignancies. The performance evaluation revealed that katdetectr has the highest accuracy and normalized Matthews Correlation Coefficient for kataegis classification on both the synthetic and the *a priori* labelled dataset. Katdetectr is in particularly more robust for kataegis detection within samples with a high tumour mutational burden.

Availability and Implementation:
Katdetectr imports standardised variant calling formats (MAF and VCF) as well as standard Bioconductor classes (GRanges and VRanges). Katdetectr segments genomic variants utilising unsupervised changepoint detection and the Pruned Exact Linear Time search algorithm. Kataegis foci are flagged based on the historical definition, namely that a kataegis foci is a continuous segment harbouring ≥6 variants and has a mean intermutation distance ≤1000 bp. Additionally, the implementation of changepoint detection utilised by katdetectr results in fast computation. Furthermore, katdetectr is available on Bioconductor which ensures reliability, and operability on common operating systems (Windows, macOS and Linux). Katdetectr is available on Bioconductor at https://www.bioconductor.org/packages/devel/bioc/html/katdetectr.html and on GitHub at https://github.com/ErasmusMC-CCBC/katdetectr. All code used for the performance evaluation is available on GitHub at: https://github.com/ErasmusMC-CCBC/evaluation_katdetectr

Contact: h.vandewerken@erasmusmc.nl

Introduction

Next-generation sequencing of cancer genomes has revealed that mutations can cluster together, i.e., the acquired mutations are found in proximity to one another, much closer than would be expected if they had been dispersed uniformly throughout the genome purely by chance (Alexandrov et al., 2013a; Nik-Zainal et al., 2012a). This phenomenon was termed kataegis and its respective genomic location was termed a kataegis foci. Kataegis, which is Greek for thunderstorm or shower, was first observed and visualised in whole genome sequencing (WGS) data of 21 primary breast cancers (Nik-Zainal et al., 2012b). Alexandrov et al. subsequently detected 873 kataegis foci in a pan-cancer dataset containing 507 WGS samples from primary malignancies (Alexandrov et al., 2013b).

Extensive exploration of the aetiology of kataegis revealed a significant positive correlation between kataegis and two distinct mutational signatures both attributed to the APOBEC enzyme-family Alexandrov et al., 2020; Bergstrom, Luebeck, et al., 2022; Burns et al., 2013; Taylor et al., 2013b).

Subsequently, multiple studies confirmed the importance of the APOBEC enzymes in cancer, showing that APOBEC is a major cause of mutagenesis, both seen in clusters, dispersed throughout the cancer genome and in extrachromosomal DNA (Bergstrom et al., 2021; Bergstrom, Luebeck, et al., 2022; Langenbucher et al., 2021; Maciejowski et al., n.d.; Taylor et al., 2013a).

Previous studies have shown that kataegis occurs within known cancer genes including TP53, EGFR and BRAF which are associated with overall survival (Bergstrom, Luebeck, et al., 2022). Still, the clinical significance of kataegis remains to be validated and therefore obfuscates kataegis as a clinical biomarker for predicting prognosis. Nevertheless, any future clinical application requires accurate and robust detection of kataegis.

Here, we introduce katdetectr, an R-based and Bioconductor package that contains a complete suite for the detection, characterisation and visualisation of kataegis. Additionally, we have evaluated the performance of katdetectr and five publicly available kataegis detection packages (Bergstrom, Kundu, et al., 2022; Lin et al., 2021; Lora, 2016; Mayakonda et al., 2018; Yousif et al., 2020).

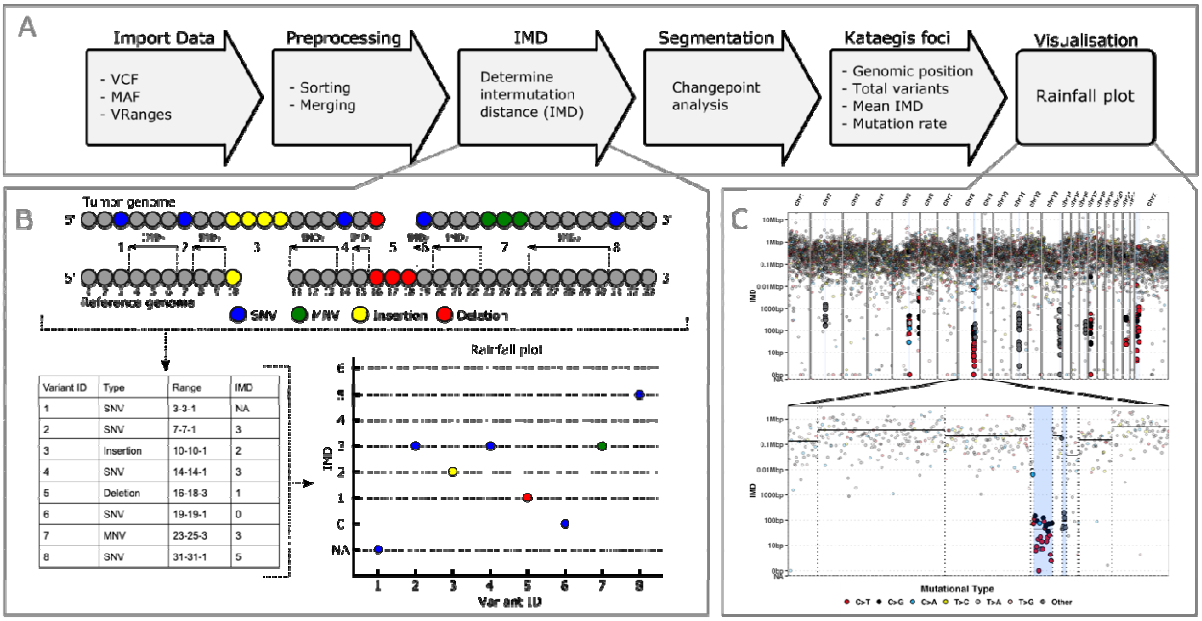

Figure 1, Overview of the katdetectr workflow, Intermutation distance and rainfall plots. A) General workflow of katdetectr represented by arrows. B) The intermutation distance (IMD) is determined for each two subsequent genomic variants per chromosome and rainfall plots are used to visualise these IMDs and corresponding detected changepoint segments. C) Rainfall plot of PD7049a (breast cancer) from the Alexandrov dataset as interrogated by katdetectr (Alexandrov et al., 2013a). Y-axis: IMD, x-axis: variant ID ordered on genomic appearance, light blue rectangles: kataegis foci with genomic variants within kataegis foci shown in bold. The mutational type is depicted by the colour. The determined segmentation (as mean IMD per segment) is shown by black horizontal solid lines whilst vertical lines represent detected changepoints. Note that the first variant of a kataegis foci has a high IMD due to the usage of the upstream-oriented IMD.

Approach

Katdetectr was programmed in the R statistical programming language (v4.1.2) (R Core Team, 2022). Briefly, katdetectr can import standardised formats denoting genomic variants including: Variant Calling Format (VCF), Mutation Annotation Format (MAF) and VRanges objects. Per sample, the genomic variants are pre-processed and subsequently the upstream-oriented intermutation distance (IMD) is calculated (Nik-Zainal et al., 2012a). The distribution of IMDs is then segmented based on unsupervised detection of changepoints using the changepoint package (v2.2.3) and the Pruned Exact Linear Time (PELT) search method (Haynes et al., 2017; Haynes & Killick, 2021; Killick et al., 2012; Killick & Eckley, 2014).

After segmentation, putative kataegis foci are called based on the following definition: 1) a continuous segment harbouring ≥6 variants and 2) the captured IMDs within the segment contain a mean IMD of ≤1000 bp (Alexandrov et al., 2013a). Moreover, katdetectr can

visualise the IMD, changepoints and their continuous segments and can highlight all putative kataegis foci within a sample using an intuitive rainfall plot (Figure 1).
The output of katdetectr consists of an S4 object containing the putative kataegis foci (GRanges), the annotated genomic variants (VRanges) and the annotated segments (GRanges).

See supplementary note 1 for an extended description of the design of katdetectr and parameters settings.

Method

The performance of katdetectr (v1.0.0) was compared to alternative packages by utilising an in-house generated synthetic dataset containing 1024 samples and a publicly available pan-cancer dataset containing 507 WGS samples with a priori labelled kataegis foci as curated by Alexandrov et al. (2013) (Alexandrov et al., 2013a; Bergstrom, Kundu, et al., 2022; Lin et al., 2021; Lora, 2016; Mayakonda et al., 2018; Yousif et al., 2020).

In order to quantify and compare performances, the task of kataegis detection was reduced to a binary classification problem. The task of the kataegis detection packages was to correctly label each variant for kataegis, i.e., whether or not a genomic variant lies within a kataegis foci.

In order to assess performance related to sample-specific Tumour Mutational Burden (TMB), we binned samples based on TMB. The synthetic dataset contained eight TMB classes (0.1, 0.5, 1, 5, 10, 50, 100, 500) whilst the Alexandrov dataset was binned into three TMB classes (low: TMB < 0.1, middle: 0.1 ≥ TMB < 10, high: TMB ≥ 10).
Due to large class imbalance, we used the normalised Matthews Correlation Coefficient (nMCC) as the main performance metric for performance evaluation (Chicco & Jurman, 2020).

See supplementary note 1 for an extended description of the datasets, synthetic data generation and confusion matrices.

Performance kataegis classification

				Synthetic dataset					Dataset labelled by Alexandrov et al.				
	Package	Reference	Language	Accuracy	nMCC	F1	TPR	TNR	Accuracy	nMCC	F1	TPR	TNR
1	katdetectr	Hazelaar, van Riet et al., 2022	R	0.99	0.98	0.97	0.94	0.99	0.99	0.92	0.83	0.91	0.99
2	SeqKat	Taylor et al., 2013	R	0.84	0.54	0.02	0.93	0.84	0.99	0.85	0.69	0.59	0.99
3	MafTools	Mayakonda et al., 2018	R	0.74	0.53	0.01	0.96	0.74	0.99	0.85	0.66	0.93	0.99
4	SigProfilerClusters	Bergstrom, Kundu, et al., 2022	Python	0.65	0.52	0.01	0.88	0.65	0.99	0.84	0.68	0.66	0.99
5	ClusteredMutations	Lora, 2016	R	0.70	0.53	0.01	0.99	0.74	0.99	0.83	0.61	0.99	0.99
6	kataegis	Lin et al., 2021	R	0.99	0.80	0.52	0.36	0.99	0.99	0.56	0.03	0.02	0.99

Table 1, performance metrics of evaluated kataegis detection packages. Accuracy, normalized Matthews Correlation Coefficient (nMCC), F1 score, True Positive Rate (TPR) and True Negative Rate (TNR) of the kataegis detection packages on 1024 synthetic samples and 507 a priori labelled WGS samples (Alexandrov et al., 2013a). Rows were sorted in descending order based on nMCC score on the Alexandrov dataset (grey transparent background). For each performance metric, the highest score is underlined.

Results

Out of all evaluated packages, katdetectr revealed the highest overall accuracy and nMCC in correctly labelling kataegis foci within both the synthetic and Alexandrov et al. dataset (Table 1). The performance of all packages was found to be associated with the sample-respective TMB (**Supplementary Figure 1**). Performance evaluation per TMB-binned category revealed that katdetectr is on par with alternative packages for samples with TMB ≤50. However, in contrast to alternative packages, the nMCC of katdetectr remains high for samples with high TMB (ranging between 50-500; **Supplementary Figures 2-3**). Furthermore, katdetectr demonstrated the fastest computational runtimes of all evaluated packages (**Supplementary Figures 4**).

Conclusion

Here, we described katdetectr; an R-based Bioconductor package capable of the detection, characterization and visualization of putative kataegis foci within genomic variants. Performance evaluation revealed that katdetectr robustly detects kataegis in a wide range of malignancies, irrespectively of low or high TMB. Additionally, katdetectr is user-friendly and computationally inexpensive with fast runtimes. In conclusion, the robust and reproducible methodologies of katdetectr can help facilitate further research into the clinical significance and underlying biological mechanism of kataegis.

Acknowledgements

We would like to thank John Martens, Marcel Smid and Guido Jenster for their discussions, input and support. Additionally, we would like to thank Coen Berns and Yi Ping for their initial efforts on detecting kataegis.

Funding

This research received funding from the Daniel den Hoed Fonds - Cancer Computational Biology Center (DDHF-CCBC) grant.
Conflict of Interest: none declared.

Data availability

All data used in the performance evaluation can be found on Zenodo at:
https://zenodo.org/record/6623289#.YqBxHi8Rr0o

References

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., ... & Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. Nature, 578(7793), 94-101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., ... & Stratton, M. R. (2013). Signatures of mutational processes in human cancer. Nature, 500(7463), 415-421. https://doi.org/10.1038/nature12477

Bergstrom, E. N., Kundu, M., Tbeileh, N., & Alexandrov, L. B. (2022). Examining clustered somatic mutations with SigProfilerClusters. bioRxiv. https://doi.org/10.1093/BIOINFORMATICS/BTAC335

Bergstrom, E. N., Luebeck, J., Petljak, M., Khandekar, A., Barnes, M., Zhang, T., ... & Alexandrov, L. B. (2022). Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. Nature, 602(7897), 510-517. https://doi.org/10.1038/s41586-022-04398-6

Bergstrom, E. N., Luebeck, J., Petljak, M., Bafna, V., Mischel, P. S., Harris, R. S., & Alexandrov, L. B. (2021). Comprehensive analysis of clustered mutations in cancer reveals recurrent APOBEC3 mutagenesis of ecDNA. bioRxiv. https://doi.org/10.1101/2021.05.27.445689

Burns, M. B., Lackey, L., Carpenter, M. A., Rathore, A., Land, A. M., Leonard, B., ... & Harris, R. S. (2013). APOBEC3B is an enzymatic source of mutation in breast cancer. Nature, 494(7437), 366-370. https://doi.org/10.1038/NATURE11881

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7

Haynes, K., Fearnhead, P., & Eckley, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. Statistics and computing, 27(5), 1293-1305. https://doi.org/10.1007/s11222-016-9687-5

Haynes, K., & Killick, R. (2021). changepoint.np: Methods for Nonparametric Changepoint Detection. https://CRAN.R-project.org/package=changepoint.np

Killick, R., & Eckley, I. (2014). changepoint: An R package for changepoint analysis. Journal of statistical software, 58(3), 1-19.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association, 107(500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745

Langenbucher, A., Bowen, D., Sakhtemani, R., Bournique, E., Wise, J. F., Zou, L., ... & Lawrence, M. S. (2021). An extended APOBEC3A mutation signature in cancer. Nature communications, 12(1), 1-11. https://doi.org/10.1038/s41467-021-21891-0

Lin, X., Hua, Y., Gu, S., Lv, L., Li, X., Chen, P., Dai, P., Hu, Y., Liu, A., & Li, J. (2021). kataegis: an R package for identification and visualization of the genomic localized hypermutation regions using high-throughput sequencing. BMC Genomics, 22(1), 1–6. https://doi.org/10.1186/s12864-021-07696-x

Lora, D. (2016). ClusteredMutations: Location and Visualization of Clustered Somatic Mutations. https://CRAN.R-project.org/package=ClusteredMutations

Maciejowski, J., Chatzipli, A., Dananberg, A., Chu, K., Toufektchan, E., Klimczak, L. J., ... & de Lange, T. (2020). APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. Nature genetics, 52(9), 884-890. https://doi.org/10.1038/s41588-020-0667-5

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome research, 28(11), 1747-1756. https://doi.org/10.1101/gr.239244.118

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., ... & Breast Cancer Working Group of the International Cancer Genome Consortium. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell, 149(5), 979-993. https://doi.org/10.1016/j.cell.2012.04.024

R Core Team. (2022). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Taylor, B. J., Nik-Zainal, S., Wu, Y. L., Stebbings, L. A., Raine, K., Campbell, P. J., ... & Neuberger, M. S. (2013). DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. elife, 2. doi: 10.7554/eLife.00534

Yousif, F., Lin, X., Fan, F., Lalansingh, C., & Macdonald, J. (2020). SeqKat: Detection of Kataegis. https://CRAN.R-project.org/package=SeqKat