














# mutyper: assigning and summarizing mutation types for analyzing germline mutation spectra

William S. DeWitt <sup>1</sup><sup>2</sup>, Luke Zhu <sup>2</sup>, Mitchell R. Vollger <sup>3</sup>, Michael E. Goldberg <sup>3,4</sup>, Andrea Talenti <sup>5</sup>, Annabel C. Beichman <sup>3</sup>, and Kelley Harris <sup>3</sup>

**1** Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, CA, United States of America **2** Department of Bioengineering, University of Washington, Seattle, WA, United States of America **3** Department of Genome Sciences, University of Washington, Seattle, WA, United States of America **4** Departments of Human Genetics and of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States of America **5** The Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian, United Kingdom   
Corresponding author

DOI: [10.21105/joss.05227](https://doi.org/10.21105/joss.05227)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Lorena Pantano](#)  

## Reviewers:

- [@izabelcavassim](#)
- [@vladsavelyev](#)

Submitted: 24 January 2023

Published: 10 May 2023

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

The germline mutation process drives genetic variation and provides the raw material for adaptive evolution. Germline mutations arise from spontaneous DNA damage caused by environmental mutagens, or errors in DNA replication. Populations and species may experience distinct mutational histories due to variation in environmental exposure, life history, and heritable variation in the machinery controlling DNA replication fidelity.

Mutational mechanisms often have *mutation signatures* in terms of the nucleotide sequence contexts where they act. Population genomics has given increasing attention to nucleotide sequence context in the study of the germline mutation process (reviewed in Carlson et al. (2020)). Single-nucleotide polymorphisms (SNPs) can be assigned to *mutation types* by the ancestral and derived nucleotide states and a window of local nucleotide context in the ancestral background. The *mutation spectrum* of an individual or population is the relative distribution of these mutation types.

Inter- and intra-specific germline mutation spectrum variation has revealed a dynamic and evolving germline mutation process shaping modern genomic diversity. Parsing mutation spectra temporally (via allele frequency) and spatially (via genomic annotations) has revealed the history and present of mutational processes, and applying such analysis to *de novo* mutation data may be clinically informative for rare or undiagnosed genetic diseases.

Here we describe mutyper, a command-line utility and Python package that assigns ancestrally polarized mutation types to SNP data, computes mutation spectra for individuals and populations, and computes sample frequency spectra stratified by mutation type for population genetic inference. Documentation is provided at <https://harrispopgen.github.io/mutyper>; source code is available at <https://github.com/harrispopgen/mutyper>.

## Statement of need

Despite many exciting findings in this growing area, there is a lack of software for germline mutation type annotation and spectrum generation from population-scale genomic data. We developed mutyper, an open-source command-line utility and Python package, to address the field's need for efficient and well-tested software for both larger bioinformatics pipelines and exploratory analysis.

The literature on cancer somatic mutation signatures includes several software tools for clustering and dimensionality reduction that are either not scalable or not flexible enough for general population-scale germline variation data (Gehring et al., 2015; Goncarenco et al., 2017; Lee et al., 2018; S. Li et al., 2020; Manders et al., 2022; Rosales et al., 2017; Rosenthal et al., 2016), but the package `helmsman` (Carlson et al., 2018) enables partial interoperability with some of these tools. Complementing this work, `mutyper` is a flexible, efficient, and extensible software package for low-level bioinformatic workflows in germline mutation spectrum studies.

## Implementation

### CLI

The core functionality of the `mutyper` command-line interface (CLI) is to augment SNP data (input or piped in VCF/BCF format) with ancestral mutation type annotations and stream to stdout. Fast and memory-efficient processing of VCF input (Danecek et al., 2011) is achieved with `cyvcf2` (Pedersen & Quinlan, 2017), and mutation types are assigned via the INFO field for each variant via a key-value pair such as `mutation_type=GAG>GTG`. Reference and alternative alleles are polarized to the ancestral and derived states, respectively, and genotype counts are updated accordingly. The `mutyper` CLI is fully compatible with standard CLIs (i.e. `bcftools` (H. Li, 2011)) for filtering SNPs or samples, masking regions, and merging/concatenating VCFs.

To polarize ancestral and derived allelic states, and define ancestral  $k$ -mer backgrounds, an ancestral genome in FASTA format is required. `Mutyper` uses the package `pyfaidx` (Shirley et al., 2015) for fast random access to ancestral genomic content, with minimal memory requirements. Ancestral genomes can be specified by various means. The ancestral FASTA sequence provided by the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) was estimated from a multi-species alignment using `orthus` (Paten et al., 2008). In such a case, the ancestral FASTA can be passed to `mutyper` directly. Alternatively, `mutyper` can estimate ancestral states by polarizing SNPs using an outgroup genome aligned to the reference (e.g. the chimp genome leftover to the human reference genome).

The user may specify the  $k$ -mer context size desired (e.g.  $k = 3$  for triplet mutation types). As in previous work, mutation type annotations are, by default, collapsed by reverse complementation such that the ancestral state is either A or C. Alternatively, a BED file can be supplied to define the strand orientation for nucleotide context at each site (e.g. according to direction of replication or transcription).

In addition to this core functionality, the CLI includes several other subcommands that summarize mutation-type-annotated SNP data piped from the core command described above. Individual- and population-level mutation spectra and sample frequency spectra are streamed to stdout in tab-separated form, and can be used to characterize modern mutation spectrum variation, and infer its evolutionary history.

### Python API

The `mutyper` Python API exposes the functions above in an interactive notebook session to implement custom analyses of mutation type data by interfacing with the strong ecosystem of scientific computing packages available in Python. For example, dimensionality reduction (such as principal components analysis or non-negative matrix factorization) is often used to summarize mutation spectra, and the `scikit-learn` package (Pedregosa et al., 2011) can be used in conjunction with the `mutyper` API for this purpose. The `mutyper` API produces mutation spectra or sample frequency spectrum matrices as `pandas` data frames (McKinney, 2010), which can be easily manipulated, visualized, and analyzed with standard python scientific computing packages.

## Applications

mutyper was first used by DeWitt et al. (2021) alongside the Python package `mushi` to infer mutation rate histories from mutation spectra using coalescent theory. Sasani et al. (2022) used mutyper in work reporting the discovery of a mutator allele in a unique mouse model system. Vollger et al. (2022) used mutyper to analyze long-read sequencing data from humans, finding elevated mutation rates and distinct mutation spectra in segmentally duplicated regions. As of this writing, mutyper is being used in several ongoing studies in multiple labs.

## Acknowledgements

The authors thank reviewers Izabel Cavassim and Vlad Savelyev for comments and corrections. Jedidiah Carlson and Sarah Hilton provided comments on an early draft. WSD was supported by the National Institute Of Allergy And Infectious Diseases (F31AI150163), and a Fellowship in Understanding Dynamic and Multi-scale Systems from the James S. McDonnell Foundation. AT has been supported by the Institute Strategic Programme Grant BBS/E/D/10002070 from the Biotechnology and Biological Sciences Research Council (BBSRC). ACB was supported by the Biological Mechanisms of Healthy Aging Training Program, NIH T32AG066574. KH was supported by the National Institute of General Medical Sciences (1R35GM133428-01), a Burroughs Wellcome Career Award at the Scientific Interface, a Pew Biomedical Scholarship, a Searle Scholarship, and a Sloan Research Fellowship.

## References

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. <https://doi.org/10.1038/nature15393>
- Carlson, J., DeWitt, W. S., & Harris, K. (2020). Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. *Current Opinion in Genetics & Development*, *62*, 50–57. <https://doi.org/10.1016/j.gde.2020.05.024>
- Carlson, J., Li, J. Z., & Zöllner, S. (2018). Helmsman: Fast and efficient mutation signature analysis for massive sequencing datasets. *BMC Genomics*, *19*(1), 845. <https://doi.org/10.1186/s12864-018-5264-y>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- DeWitt, W. S., Harris, K. D., Ragsdale, A. P., & Harris, K. (2021). Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, *118*(21), e2013798118. <https://doi.org/10.1073/pnas.2013798118>
- Gehring, J. S., Fischer, B., Lawrence, M., & Huber, W. (2015). SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, *31*(22), 3673–3675. <https://doi.org/10.1093/bioinformatics/btv408>
- Goncarenco, A., Rager, S. L., Li, M., Sang, Q.-X., Rogozin, I. B., & Panchenko, A. R. (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Research*, *45*(W1), W514–W522. <https://doi.org/10.1093/nar/gkx367>
- Lee, J., Lee, A. J., Lee, J.-K., Park, J., Kwon, Y., Park, S., Chun, H., Ju, Y. S., & Hong, D. (2018). Mutalisk: A web-based somatic MUTation AnaLYSis toolKit for genomic,

- transcriptional and epigenomic signatures. *Nucleic Acids Research*, 46(W1), W102–W108. <https://doi.org/10.1093/nar/gky406>
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
- Li, S., Crawford, F. W., & Gerstein, M. B. (2020). Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. *Nature Communications*, 11(1), 3575. <https://doi.org/10.1038/s41467-020-17388-x>
- Manders, F., Brandsma, A. M., Kanter, J. de, Verheul, M., Oka, R., Roosmalen, M. J. van, Roest, B. van der, Hoeck, A. van, Cuppen, E., & Boxtel, R. van. (2022). MutationalPatterns: The one stop shop for the analysis of mutational processes. *BMC Genomics*, 23(1), 134. <https://doi.org/10.1186/s12864-022-08357-3>
- McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., & Birney, E. (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, 18(11), 1829–1843. <https://doi.org/10.1101/gr.076521.108>
- Pedersen, B. S., & Quinlan, A. R. (2017). cyvcf2: Fast, flexible variant analysis with python. *Bioinformatics*, 33(12), 1867–1869. <https://doi.org/10.1093/bioinformatics/btx057>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://doi.org/10.48550/arXiv.1201.0490>
- Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E., & Da Silva, I. T. (2017). signeR: An empirical bayesian approach to mutational signature discovery. *Bioinformatics*, 33(1), 8–16. <https://doi.org/10.1093/bioinformatics/btw572>
- Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). DeconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1), 1–11. <https://doi.org/10.1186/s13059-016-0893-4>
- Sasani, T. A., Ashbrook, D. G., Beichman, A. C., Lu, L., Palmer, A. A., Williams, R. W., Pritchard, J. K., & Harris, K. (2022). A natural mutator allele shapes mutation spectrum variation in mice. *Nature*, 605(7910), 497–502. <https://doi.org/10.1101/2021.03.12.435196>
- Shirley, M. D., Ma, Z., Pedersen, B. S., & Wheelan, S. J. (2015). *Efficient “pythonic” access to FASTA files using pyfaidx* (No. e1196). PeerJ PrePrints; PeerJ Inc. <https://doi.org/10.7287/peerj.preprints.970v1>
- Vollger, M. R., DeWitt, W. S., Dishuck, P. C., Harvey, W. T., Guitart, X., Goldberg, M. E., Rozanski, A. N., Lucas, J., Asri, M., Munson, K. M., Lewis, A. P., Hoekzema, K., Logsdon, G. A., Porubsky, D., Paten, B., Harris, K., Hsieh, P., & Eichler, E. E. (2022). Increased mutation rate and interlocus gene conversion within human segmental duplications. *bioRxiv*. <https://doi.org/10.1101/2022.07.06.498021>