

# RNAheteroplasmy

Gabriele Lubatti<sup>\*1, 2, 3</sup> and Antonio Scialdone<sup>†1,2, 3</sup>

**1** Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Munich, Germany **2** Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany **3** Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

DOI: [DOIunavailable](#)

## Software

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

Editor: [Pending Editor](#) ↗

## Reviewers:

- [@Pending Reviewers](#)

Submitted: N/A

Published: N/A

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## In partnership with



This article and software are linked with research article DOI [10.3847/xxxxx](#), published in the Astrophysical Journal <- The name of the AAS journal..

## Summary

A given base on the DNA can be occupied by 4 different alleles (A, C, T and G). The DNA heteroplasmy measures the frequency of different DNA variants in a given base of the genome. For the mitochondrial DNA (mtDNA), mutation rates are estimated to be 10 to 100 fold higher than for nuclear DNA and therefore mtDNA variants can achieve an higher level of heteroplasmy (Khrapko et al., 1997). Heteroplasmy from mtDNA data helps to distinguish cells that are in different states or fate commitment as well as for lineage tracing analysis (Ludwig et al., 2019), (Miller et al., 2021). It has been recently shown that single cell RNA seq (scRNA-seq) can be used to reliably identify mtDNA variants, although with a lower statistical power compared to more direct approaches, like mtDNA sequencing (Ludwig et al., 2019). Here we provide RNAheteroplasmy, an highly user friendly R library that allow the quantification of heteroplasmy from RNA seq data starting directly from the raw fastq files.

## Statement of need

Although some tools (based on C or python (Huang & Huang, 2021), (Prashant et al., 2021)) have been developed recently to asses heteroplasmy from scRNA-seq data, a unified approach directly from raw fastq files all implemented in R that include both the quantification of heteroplamsy with some more down stream analysis (i.e statistical tests, plotting and clustering based on heteroplasmy values) is not yet available. RNAheteroplasmy is an R package developed in order to fill this gap by providing a user friendly set of functions for the detection of allele frequencies, heteroplasmy and for additional down-stream analysis: The two main functions of RNAheteroplasmy are:

1. `get_raw_counts_allele`: an highly parallelized function that rely on Rsamtools, used to generate the raw counts matrix starting from fastq files, with cells as rows and bases with the four possible allele as columns.
2. `get_heteroplasmy`: Starting from the ouput of `get_raw_counts_allele`, it computes the matrix with heteroplasmy values (defined as 1 minus the frequency of the most common allele) and the matrix with allele frequencies values,for all the cells and bases that pass a filtering step procedure.

Among the down stream analysis implemented in the package there are: 1. Several statistical tests for identification of most different bases according to heteroplasmy, between group of cells or along a trajectory of cells (i.e cells sorted according to a diffusion pseudo time). 2. Plotting functions for the visualization of heteroplamsy and corresponding allele frequencies values among cells 3. Hierarchical clustering on cells based on a distance matrix defined from angular distance of allele frequencies, that could be relevant for lineage tracing analysis

The main concepts implemented in the current package were proven to be extremely powerful in our recent paper (Lima et al., 2021), where we revealed that cells eliminated by apoptosis

<sup>\*</sup>first author

<sup>†</sup>corresponding author

in the mouse embryo at the beginning of gastrulation showed an higher level of mt-DNA heteroplasmy.

## References

- Huang, X., & Huang, Y. (2021). Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btab358>
- Khrapko, K., Coller, H. A., André, P. C., Li, X.-C., Hanekamp, J. S., & Thilly, W. G. (1997). Mitochondrial mutational spectra in human cells and tissues. *Proceedings of the National Academy of Sciences*, 94(25), 13798–13803. <https://doi.org/10.1073/pnas.94.25.13798>
- Lima, A., Lubatti, G., Burgstaller, J., Hu, D., Green, A., Gregorio, A. D., Zawadzki, T., Pernaute, B., Mahammadov, E., Dore, M., Sanchez, J. M., Bowling, S., Sancho, M., Karimi, M., Carling, D., Jones, N., Srinivas, S., Scialdone, A., & Rodriguez, T. A. (2021). Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. *Nature Metabolism*. <https://doi.org/10.1038/s42255-021-00422-7>
- Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A., & Sankaran, V. G. (2019). Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, 176(6), 1325–1339.e22. <https://doi.org/10.1016/j.cell.2019.01.022>
- Miller, T. E., Lareau, C. A., Verga, J. A., Ssozi, D., Ludwig, L. S., Farran, C. E., Griffin, G. K., Lane, A. A., Bernstein, B. E., Sankaran, V. G., & Galen, P. van. (2021). Mitochondrial variant enrichment from high-throughput single-cell RNA-seq resolves clonal populations. *bioRxiv*. <https://doi.org/10.1101/2021.03.08.434450>
- Prashant, N., Alomran, N., Chen, Y., Liu, H., Bousounis, P., Movassagh, M., Edwards, N., & Horvath, A. (2021). SCReadCounts: Estimation of cell-level SNVs from scRNA-seq data. *bioRxiv*. <https://doi.org/10.1101/2020.11.23.394569>