The Journal of Open Source Software

MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets

Gabriele Lubatti[*]^[1,2,3], **Elmir Mahammadov**[†]^[1,2,3], **and Antonio Scialdone**[1,2,3¶]

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 ¶ Corresponding author

DOI: 10.21105/joss.04265

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @BatoolMM
- @juanvillada

Submitted: 21 February 2022
Published: 01 June 2022

Summary

Eukaryotic cells rely on mitochondria, organelles that are equipped with their own DNA (mtDNA) to produce the energy they need. Each cell includes multiple mtDNA copies that are not perfectly identical but have differences in their sequence; such sequence variability is called heteroplasmy. mtDNA heteroplasmy has been associated with diseases (Nissanka & Moraes, 2020), that can affect cellular fitness and have an impact on cellular competition (Lima et al., 2021). Several single-cell sequencing protocols provide the data to estimate mtDNA heteroplasmy, including single-cell DNA-seq, RNA-seq and ATAC-seq, in addition to dedicated protocols like MAESTER (Miller et al., 2022). Here, we provide MitoHEAR (Mitochondrial HEteroplasmy AnalyzeR), a user-friendly software written in R that allows the estimation as well as downstream statistical analysis of the mtDNA heteroplasmy calculated from single-cell datasets. MitoHEAR takes as input BAM files, computes the frequency of each allele and, starting from these, estimates the mtDNA heteroplasmy at each covered position for each cell. The analysis parameters (e.g., the filtering of the mtDNA positions based on read quality and coverage) are easily tuneable. Moreover, statistical tests are available to explore the dependency of the mtDNA heteroplasmy on continuous or discrete cell covariates (e.g., culture conditions, differentiation states, etc), as extensively shown in the detailed tutorials we include.

Statement of need

Although mtDNA heteroplasmy has important consequences on human health (Stewart & Chinnery, 2015) and embryonic development (Floros et al., 2019), there are still many open questions on how heteroplasmy affects cells' ability to function and how cells keep it under control. With the increasing availability of single-cell data, many questions can begin to be answered. Still, it is fundamental to have efficient and streamlined computational tools enabling researchers to estimate and analyse mtDNA heteroplasmy. Existing packages (Calabrese et al., 2014; Huang & Huang, 2021; Prashant et al., 2021) focus only on the first step of quantifying heteroplasmy from BAM files, and do not provide any specific tools for further statistical analyses or plotting. Instead, MitoHEAR covers all steps of the analysis in a unique user-friendly package, with highly customisable functions. Starting from BAM files, MitoHEAR estimates heteroplasmy and offers several options for downstream analyses. For example, statistical tests are provided to investigate the relationship of the mtDNA heteroplasmy with continuous or

[*] first author
[†] co-author

discrete cell covariates. Moreover, there are plotting functions to visualise heteroplasmy and allele frequencies and to perform hierarchical clustering of cells based on heteroplasmy values.

Key functions

The two main functions of `MitoHEAR` are:

1. `get_raw_counts_allele`: a parallelised function that relies on Rsamtools and generates the raw counts matrix starting from BAM files, with cells as rows and bases with the four possible alleles as columns.
2. `get_heteroplasmy`: Starting from the output 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 frequency values, for all the cells and bases that pass a filtering procedure.

Among the downstream analyses implemented in the package are:

- Several statistical tests (e.g., Wilcoxon rank-sum test) for the identification of the mtDNA positions with the most different levels of heteroplasmy between discrete groups of cells or along a trajectory of cells (i.e., cells sorted according to a diffusion pseudo-time) (**Figure 1** and **Figure 2**).
- Plotting functions for the visualisation of heteroplasmy and the corresponding allele frequency values among cells.
- Unsupervised hierarchical clustering of cells based on a distance matrix defined from the angular distance of allele frequencies that could be relevant for lineage tracing analysis (Ludwig et al., 2019) (**Figure 3**).

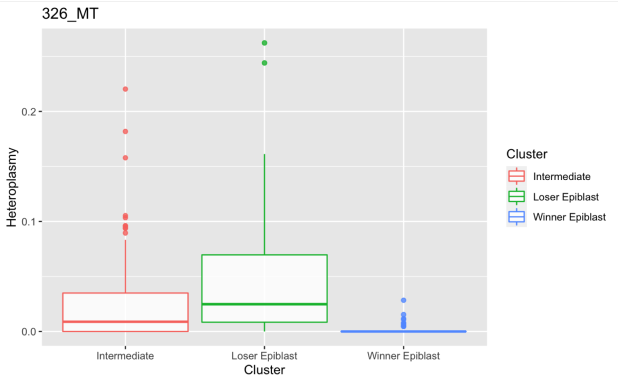

Figure 1: Example of an output plot generated by MitoHEAR showing heteroplasmy values at a given position estimated from single cells in three clusters indicated on the x-axis. Data from Lima et al. (2021).

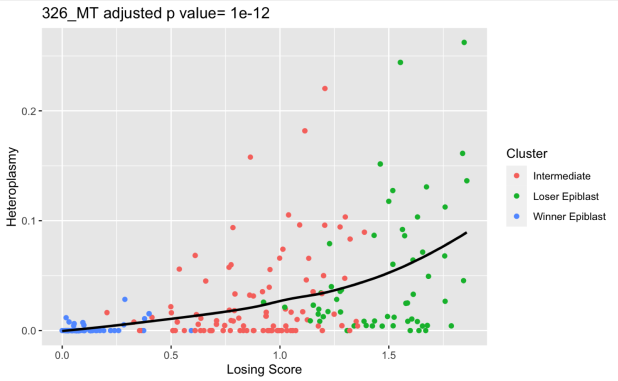

Figure 2: Example of an output figure generated by MitoHEAR where the heteroplasmy is plotted as a function of the pseudo-time coordinate of each cell. Cells are classified into three clusters. The heteroplasmy shows a statistically significant change along the pseudo-time, as indicated by the adjusted p-value reported at the top, which is computed by a generalised additive model fit. Data from Lima et al. (2021).

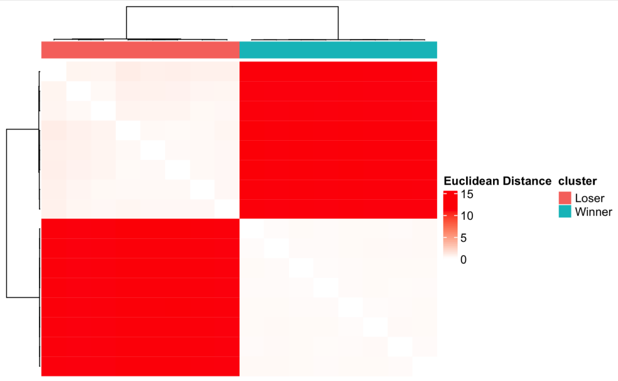

Figure 3: Unsupervised hierarchical clustering of cells based on a distance matrix defined from the angular distance of allele frequencies. The data shown is bulk RNA-seq mouse data from two mtDNA cell lines labelled *Loser* and *Winner*. Data from Lima et al. (2021).

The package has been used in a recently published paper (Lima et al., 2021), where we revealed that cells with higher levels of heteroplasmy are eliminated by cell competition in mouse embryos and are characterised by specific gene expression patterns.

References

Calabrese, C., Simone, D., Diroma, M. A., Santorsola, M., Guttà, C., Gasparre, G., Picardi, E., Pesole, G., & Attimonelli, M. (2014). MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, *30*(21), 3115–3117. https://doi.org/10.1093/bioinformatics/btu483

Floros, V., Pyle, A., Dietmann, S., Wei, W., Tang, W., Irie, N., Payne, B., Capalbo, A., Noli, L., Coxhead, J., Hudson, G., Crosier, M., Strahl, H., Khalaf, Y., Saitou, M., Ilic, D., Surani, M., & Chinnery, P. (2019). Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nature Cell Biology*. https://doi.org/10.1038/s41556-017-0017-8

Huang, X., & Huang, Y. (2021). Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab358

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., & van Galen, P. (2022). Mitochondrial variant enrichment from high-throughput single-cell RNA-seq resolves clonal populations. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01210-8

Nissanka, N., & Moraes, C. T. (2020). Mitochondrial DNA heteroplasmy in disease and targeted nuclease-based therapeutic approaches. *EMBO Reports*, *21*(3), e49612. https://doi.org/10.15252/embr.201949612

Prashant, N., Alomran, N., Chen, Y., Liu, H., Bousounis, P., Movassagh, M., Edwards, N., & Horvath, A. (2021). SCReadCounts: Estimation of cell-level SNVs expression from scRNA-seq data. *BMC Genomics*. https://doi.org/10.1186/s12864-021-07974-8

Stewart, J., & Chinnery, P. (2015). The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3966