Review of: "Novel method for multiplexed full-length single-molecule sequencing of the human mitochondrial genome"

Daniele Ghezzi[1], Andrea Legati[1], Chiara Frascarelli[1]

1 Fondazione I.R.C.C.S. Istituto Neurologico Carlo Besta

Potential competing interests: The author(s) declared that no potential competing interests exist.

Definitions

Heteroplasmy
Defined by Daniele Ghezzi

The manuscript by Keraite et al.[1] describes an interesting approach for sequencing the mitochondrial DNA (mtDNA) by using long reads. The authors exploited RNA-guided DNA endonuclease Cas9, which has been previously used for the PCR-free enrichment of specific nuclear regions [2] [3], and developed a method for mtDNA. Another valuable point is the use of different RNA guides (gRNA), and thus different break sites, that make possible to analyse multiple samples pooled on one without the need of barcoding.
The advantages of long-read sequencing are diverse and, for what concerns mtDNA, they mainly include the possibility to determine phase of different heteroplasmic variants and to define accurately complex deletion patterns.

We would like to report some comments about this paper and to highlight caveats linked to the proposed approach.
The heteroplasmy threshold was set at 1%, and all called variants with ≥1% heteroplasmy were concordant with the previous laboratory results. This statement suggests that 1% is the detection limit but it is misleading. Indeed, a series of variants of unknown significance were identified in diverse samples, with heteroplasmy up to 10% (or between 90% and 100%). While the authors compared the percentages of known pathogenic variants with short-read sequencing method (see Table 1 in Keraite et al. [1]), they did not performed the same analysis for these additional variants. Furthermore, no sequencing replicates were performed. Because of this missing information, it is not possible to exclude that these variants are errors in variant calling, thus suggesting a precision lower than expected for the proposed method.
In the paper, the “ont_align_view” program was mentioned for filtering SNV as well as for investigating their phasing. However, we were not able to find this tool on line and no reference citing it can be found in the literature. For SNV detection there are many tools that have been described to perform well and are now widely used, such as Nanopolish (https://github.com/jts/nanopolish) and Medaka (https://github.com/nanoporetech/medaka). Moreover, samples presenting low coverage could be processed using a more recent tool such Clair3 (https://github.com/HKU-BAL/Clair3), which is particularly suitable for low coverage regions.
The authors stated that their protocol allows the analysis of low integrity genomic DNA (gDNA). However, since the first step is the dephosphorylation of all free 5'-ends, in order to avoid ligation of sequencing adaptors to all non-targeted DNA

fragments, the presence of highly fragmented mtDNA would results in loss or decreased depth of coverage for the region far from the Cas9 cutting sites. Samples from oral mucosa and urine were reported to have high DNA degradation level. This is in contrast with our experience; morever, since urinary epithelial cells have been proposed as a good (and more accessible) alternative to muscle for detection of heteroplasmic point mutations and mtDNA deletions, it would be convenient to use this specimen.

The authors suggested dedicated selection strategies to produce higher coverage, especially for highly degraded DNA. However, we expect that these approaches also strongly increased the risk of sequencing 'nuclear mitochondrial sequences' (NUMTs).

An additional issue regards the identification of multiple mtDNA deletions. As we also reported in our studies[4], the amplification by PCR leads to significant bias due to preferential amplification of short molecules and inaccurate estimation of the mtDNA species present in the sample. This is an important point to stress since most of the diagnostic procedures to analyse mtDNA still rely on PCR amplification.

Nevertheless, the paper reported data on only one sample with mtDNA structural variants, and it was characterized by the presence of two deleted mtDNA populations. This finding is not usual in subjects with multiple mtDNA deletions, who usually harbour a large set of deleted species, mainly located in the major arc of replication and having deletion size in the range of 1,000-10,000 bp [5]. Mutations in nuclear genes involved in mtDNA maintenance are the cause of multiple mtDNA deletions; the molecular diagnosis of the sample investigated in the paper by Keraite et al. is not reported.

Regarding deletions calls, in this study they were performed by examining the sequence coverage, looking for regions showing a "significant coverage reduction". This approach seems to be quite approximate and very likely to be affected by bias. An easier way to identify mtDNA deletions is to perform a Minimap2 alignment using the **- -splice** option, which enables long reads splicing and thus allows to identify mtDNA deletions (single or multiple) directly by visualizing coverage profile from Integrative Genome Viewer. For the identification and calling of SV many tools are available, such as Sniffle2 (https://github.com/fritzsedlazeck/Sniffles), a fast SV caller for long-read sequencing, or Nanovar (https://github.com/benoukraflab/NanoVar), a genomic SV caller that utilizes low-depth long-read sequencing. These tools are suitable for identifying mtDNA structural alterations without the need to introduce PCR amplifications, which can be detrimental for deletions heteroplasmy overestimation.

In the presence of a deleted species, the location of Cas9 induced break needs to be outside the mtDNA deletion otherwise only the full-length sequence is cut and then sequenced. The authors selected three gRNAs targeting at different positions in order to avoid problems in detecting deleted species. However, this was possible because they have three populations with known breakpoints; in the real cases with multiple deletions, a large number of different deleted regions (with possible overlaps) is present and it is not possible to select a priori the best set of gRNAs. In the specific case reported in the paper by Keraite et al., the cutting site of the mt3 Cas9 localize at position 3127, very close to the coverage decrease observed at 3257 (Figure 2C in Keraite et al.); it could be possible that the level of heteroplasmy calculated for the m.3257_16071del is inaccurate, probably underestimated.

Two major advantages of the method described by Keraite et al. are the enrichment of the mtDNA and the possibility to multiplexed sequencing. Nevertheless, for both there are some doubts due to missing information.

While the enrichment of mtDNA by Exonuclease V is reported in the methods, it is not well defined in the text the degree

of improvement in mtDNA over total DNA that is obtained with this approach (and when it was used), as well as the comparison with the enrichment obtained by the Cas9 method. According to the data reported in supplementary tables, only 2-3% of reads from blood DNA sequencing mapped to mtDNA while this parameter was >30% in HEK293 cells. Concerning multiplexing, the method was used for up to 4 samples pooled together but most of the experiments were performed using one flow cell for each sample. In addition, it is not clear which flow cells were used, since in the method section diverse versions (R9.4, R10.3 and R10.4) are reported.

In the discussion session, the authors stated that their studied revealed undeniable benefits of applying Cas9-mtDNA enrichment protocol. It should be noted that most of the results were produced on 15 samples for the analysis of single point mutations, while only one sample carrying SV was analysed. As the authors reported in Table 1, short-reads NGS obtained very similar heteroplasmic percentages respect to their long-read approach; for this reason the benefits of using Cas9-mtDNA enrichment and long reads for the detection of single point mutations appear limited. The most promising benefits of long-reads NGS should come from the analysis of mtDNA SV and complex-rearrangements; however, in the present study the authors tested their approach on only one sample affected by large deletions, using a deletion call method simply based on the examination of sequence coverage.

We think that further tests on additional subjects with mtDNA SV are needed to properly evaluate the utility of the proposed method.

References

1. [a, b] *Ieva Keraite, Philipp Becker, Davide Canevazzi, Maria C. Frias-López, et al. (2022).Novel method for multiplexed full-length single-molecule sequencing of the human mitochondrial genome. doi:10.1101/2022.02.08.479581.*
2. [^] *Pay Giesselmann, Björn Brändl, Etienne Raimondeau, Rebecca Bowen, et al. (2019).Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nat Biotechnol, vol. 37 (12), 1478-1481. doi:10.1038/s41587-019-0293-x.*
3. [^] *Amelia D. Wallace, Thomas A. Sasani, Jordan Swanier, Brooke L. Gates, et al. (2021).CaBagE: A Cas9-based Background Elimination strategy for targeted, long-read DNA sequencing. PLoS ONE, vol. 16 (4), e0241253. doi:10.1371/journal.pone.0241253.*
4. [^] *Andrea Legati, Nadia Zanetti, Alessia Nasca, Camille Peron, et al. (2021).Current and New Next-Generation Sequencing Approaches to Study Mitochondrial DNA. The Journal of Molecular Diagnostics, vol. 23 (6), 732-741. doi:10.1016/j.jmoldx.2021.03.002.*
5. [^] *David C. Samuels, Eric A. Schon, Patrick F. Chinnery. (2004).Two direct repeats cause most human mtDNA deletions. Trends in Genetics, vol. 20 (9), 393-398. doi:10.1016/j.tig.2004.07.003.*