HAL
open science

A new method for long-read sequencing of animal mitochondrial genomes: application to the identification of equine mitochondrial DNA variants

Sophie Dhorne-Pollet, Eric Barrey, Nicolas Pollet

To cite this version:

Sophie Dhorne-Pollet, Eric Barrey, Nicolas Pollet. A new method for long-read sequencing of animal mitochondrial genomes: application to the identification of equine mitochondrial DNA variants. BMC Genomics, 2020, 21 (1), 10.1186/s12864-020-07183-9 . hal-03104281

HAL Id: hal-03104281

https://hal.science/hal-03104281

Submitted on 11 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dhorne-Pollet *et al. BMC Genomics* (2020) 21:785
https://doi.org/10.1186/s12864-020-07183-9

BMC Genomics

METHODOLOGY ARTICLE Open Access

A new method for long-read sequencing of animal mitochondrial genomes: application to the identification of equine mitochondrial DNA variants

Sophie Dhorne-Pollet[1], Eric Barrey[1†] and Nicolas Pollet[2*†]

Abstract

Background: Mitochondrial DNA is remarkably polymorphic. This is why animal geneticists survey mitochondrial genomes variations for fundamental and applied purposes. We present here an approach to sequence whole mitochondrial genomes using nanopore long-read sequencing. Our method relies on the selective elimination of nuclear DNA using an exonuclease treatment and on the amplification of circular mitochondrial DNA using a multiple displacement amplification step.

Results: We optimized each preparative step to obtain a 100 million-fold enrichment of horse mitochondrial DNA relative to nuclear DNA. We sequenced these amplified mitochondrial DNA using nanopore sequencing technology and obtained mitochondrial DNA reads that represented up to half of the sequencing output. The sequence reads were 2.3 kb of mean length and provided an even coverage of the mitochondrial genome. Long-reads spanning half or more of the whole mtDNA provided a coverage that varied between 118X and 488X. We evaluated SNPs identified using these long-reads by Sanger sequencing as ground truth and found a precision of 100.0%; a recall of 93.1% and a F1-score of 0.964 using the Twilight horse mtDNA reference. The choice of the mtDNA reference impacted variant calling efficiency with F1-scores varying between 0.947 and 0.964.

Conclusions: Our method to amplify mtDNA and to sequence it using the nanopore technology is usable for mitochondrial DNA variant analysis. With minor modifications, this approach could easily be applied to other large circular DNA molecules.

Keywords: Mitochondria, mtDNA, Horse, Nanopore sequencing, MinION device, Single nucleotide polymorphism

* Correspondence: Nicolas.Pollet@egce.cnrs-gif.fr
[†]Eric Barrey and Nicolas Pollet should be considered senior co-authors.
[2]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198 Gif-sur-Yvette, France
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Mitochondria have become crucial organelles of eukaryote cells since an ancestral bacterial endosymbiosis event gave rise to the eukaryote branch of the tree of life, about 1.5 billion years ago [1, 2]. The discovery that mitochondria contains DNA in the early 1960s attracted the attention of many scientists and more recently the word of mitogenomics has been coined to unite this field of study [3]. Today, thousands of mitochondrial genomes (mtDNA) from many species have been sequenced, and ongoing studies require robust, rapid and cheap method to decipher novel genomes and identify genetic variants [4]. In human alone, more than 50 thousand full length mtDNA sequences are known, but much less sequences and variants are available for other mammals [5]. Numerous and diverse applications take advantage of mtDNA sequencing, such as diagnostic of mitochondrial diseases in human and animals, breeding selection, forensic studies, biodiversity surveys, conservation genetics and evolutionary analysis.

Animal mitochondrial genomes are usually encoded by a single circular DNA of about 10 to 20 kbp, but weird cases are known such as linear or very large molecules [6]. The mitochondrial genomes are characterized by a maternal transmission, a lack of recombination and higher frequency of mutations than in the nuclear genome [7]. In vertebrates, the ratio of mtDNA over nuclear DNA mutation rate is usually above 20 [8]. This means that mtDNA polymorphisms may occur relatively often in animal populations with the potential for associated disorders, and this is why sequencing these organelle genomes is an important component of contemporaneous animal genetics. Knowing the whole mtDNA sequence enables the identification of all possible genetic variants: s*ingle-nucleotide polymorphisms (*SNP), insertions, deletions, duplications, inversions, rearrangements. Although mitochondrial variants are typically maternally inherited, they can be sporadic. Such de novo variants are difficult to assess due to the co-existence of several mtDNA population in a cell (heteroplasmy) that may be below an assay's detection level, or different between tissues of the same organism [9]. The presence of DNA segments transferred from the mitochondria to the nuclear DNA, the so-called Nuclear Mitochondrial DNA Sequences (NUMTs), can seriously complicate mtDNA analysis in a variety of animals from bees to humans [10–12] . Such NUMTs can span several kilobases, can be nearly identical to mtDNA sequence and can exhibit variation between individuals of the same species [13].

The current methods for mtDNA sequencing rely on standard sequencing methods such as Sanger sequencing or massively parallel sequencing [4]. The choice of the starting DNA material is of importance since the relative mass amount of mtDNA is very low in comparison to nuclear DNA. Typical DNA preparations comprise approximately 0.1% mtDNA, even though there are hundreds to tens of thousands copies of mtDNA per cell [14]. One strategy relies on the amplification of mtDNA using regular or long-range PCR. This process is associated with an unpredictable risk of PCR errors, it can introduce a sequence bias, and it can be tricky because it depends on primers, on PCR efficiency, and finally the whole procedure can be time consuming and costly. In addition, the reliance on PCR primers derived from a reference sequence can drive the failure of some amplifications because they may suffer from a lack of specificity to mtDNA, and they may lead to the production of artefactual amplicons, including some derived from NUMTs [15–17]. The other strategy rely on whole-genome shotgun using massively parallel sequencing approaches, an efficient approach for mitogenomics because mtDNA reads are more prevalent than nuclear DNA reads, but it is not bullet-proof and short read mapping can also be confounded by NUMTs [18, 19]. These approaches involve significant resources for bench work and efforts in sequence handling, they are cost-effective only when high number of samples are pooled and therefore they do not scale down easily [18]. The principal caveat of the current mtDNA sequencing methods is that raw sequences enable to read only a tiny portion of the DNA at a time, preventing definitive evidence for long-range phasing or structural variant identification.

Long-read sequencing using Oxford Nanopore or PacBio technologies provide the possibility to capture in a single read the majority or the whole sequence of mtDNA molecules. Therefore, it can provide definitive evidence that a given mitogenome exists in a single cell or in a sample. So far, a few studies harnessed the Nanopore technology for sequencing mtDNA. Istace and colleagues assembled 21 yeast mitogenomes in a study where they evaluated the performances of nanopore-only assemblers. They used only 2D reads for mitogenome assembly, a chemistry that provided better sequence quality but that has been abandoned [20]. Ranjard and collaborators assembled the green-lipped mussel mitogenome (*Perna canaliculus* Gmelin 1791) using a whole-genome shotgun approach with long-reads to obtain a draft assembly that was polished using short-reads from a RNAseq experiment; but they do not give any details on the long-read sequencing output [21]. In another study, the european lobster (*Homarus gammarus*) mitogenome was assembled separately with both Illumina short-reads and Nanopore long-reads [22]. The report by Gan and collaborators is instructive since they could not obtain a large mtDNA contig using de novo or single gene bait-based assemblies of six million Illumina paired-end reads totalling 1.5 Gbp. However, the 20 kbp lobster mitogenome could be assembled from 40,687

long-reads totalling 100 Mbp. The same authors reported another mitogenome assembled similarly, with only nine long-reads [23]. These reports testify the usefulness of the Nanopore technology to sequence mitochondrial genomes, but the corresponding protocols are not tailored to identify mtDNA variants on a large-scale.

In horses, many mitochondrial functions have been shown to play important roles for muscular exercise and disorders [24–27]. Yet, the study of mtDNA variation is lagging behind even though it has been instrumental to study ancient horse's DNA and unravel the domestication events of the wild horse [28–30]. As in other mammals, the abundance and polymorphism of NUMTs is a characteristic of the horse genome and this has negative impact on the reliability of mtDNA variant identification [12, 31, 32].

The aim of our study was to trial a new method for mtDNA resequencing using the long-read technology provided by Oxford Nanopore Technologies (ONT) with the goal of identifying mtDNA variants. We developed an optimised method for mtDNA sequencing starting from total genomic DNA obtained using routine extraction methods. We implemented a protocol based on a nuclear DNA removal by nuclease treatment followed by mtDNA enrichment using multiple displacement amplification (MDA) and sequencing using the 1D chemistry on the MinION platform of ONT. We anticipate that this method will ease the identification of mtDNA variation.

Results

Enrichment of circular mtDNA by exonuclease V treatment

We started our strategy by enriching the circular forms of animal mtDNA in genomic DNA extracts. Since traditional biochemical purification of mitochondria does not scale up easily and requires a large amount of fresh tissue, we selected an alternative method. We chose an exonuclease treatment to deplete linear DNA fragments and thereby improve the proportion of mtDNA versus nuclear DNA, as previously described [33]. A frequently used enzyme to degrade linear DNA molecules is the exonuclease V that has different nuclease activities, comprising an ATP-dependent double-stranded and bi-directional exonuclease activity. Since the amount of nuDNA vastly exceeds that of mtDNA, we tried four conditions of nuclear DNA removal and evaluated their efficiency using qPCR (Fig. 1a). The effect of exonuclease V treatments impacted dramatically the quantity of nuDNA, since it diminished by several thousand fold relative to mtDNA. mtDNA quantity diminished slightly after the various exonuclease treatment to only about half of its amount in the starting DNA extract. We observed that increasing the incubation time had an effect while the amount of exonuclease V seemed not to be a limiting factor. We found that the best condition was 2 h of incubation with 10 Units of Exonuclease V, since it resulted in a ratio of mtDNA over nuDNA of 3755+/− 0.7 (Fig. 1a).

Such an enrichment was of interest, but it came at the cost of a limited amount of DNA which is inadequate for sequencing without an amplification step. We decided to amplify the mtDNA using a Whole Mitochondrial Genome Amplification (WMGA) method based on multiple displacement amplification [34]. We optimized primer design and concentration in the MDA reaction to avoid template-independent amplification, a technical bias usually observed with MDA. To evaluate the impact of the Exonuclease V treatment and the amplification efficiency of WMGA, we estimated the relative amounts of mtDNA versus nuDNA with or without an exonuclease V treatment, and using two primer concentrations. We present our results on the quantification of the amplification efficiency in Fig. 1b. We observed a nearly 1000 fold better amplification using WMGA on DNA treated with exonuclease V versus the starting undigested DNA (Fig. 1b), with an overall amplification of more than 100 million fold in the best condition using 0.2 μM of each primer. We obtained on the order of

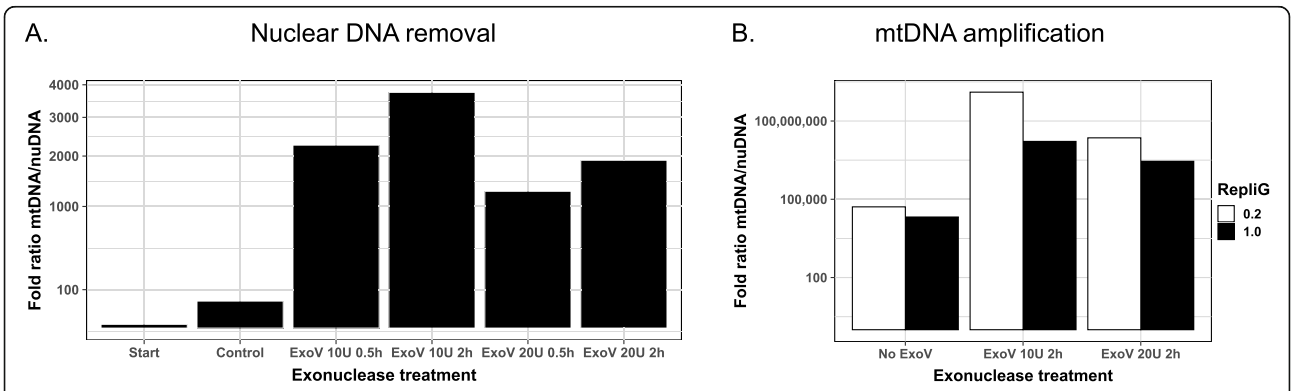

Fig. 1 Quantification of mtDNA enrichment by qPCR. **a** Quantification after nuclear DNA removal using exonuclease V (ExoV) as indicated. **b** Quantification after mtDNA amplification using MDA. The results are given for DNA untreated by exoV (No ExoV), or after two conditions of ExoV digestion as indicated. REPLI-g was performed using primers at a concentration of 0.2 μM or 1.0 μM

ten micrograms of DNA after WMGA, with an average fragment size of about 3 kbp that we deemed usable for long-read sequencing (SI_Figure 1).

Long-read sequencing

We performed a first sequencing run on nine DNA samples obtained from three horses (H25, H26 and H27) after WMGA on DNA treated or not by exonuclease V and amplified using the two primer concentration conditions mentioned previously. We will refer to DNA samples untreated by exonuclease and amplified using the lowest primer concentration as "exo –"; DNA samples treated by exonuclease and amplified using the lowest primer concentration as "exoA" and DNA samples treated by exonuclease and amplified using the highest primer concentration as "exoB" (Fig. 2, Table 1). We obtained a total of 634,062 reads with a mean length of 2370 nt and a mean quality of 10.2. After demultiplexing, we obtained 562,981 reads (88.8%) evenly distributed among the nine barcodes; with between 45,641 reads and up to 67,448 reads (Table 1 and SI Table 1). We did not observe gross differences of read size or quality between the nine barcoded libraries (Fig. 2, Table 1). We mapped these reads to the reference horse genome and counted the reads mapping to the mitochondrial genome and those mapping to the nuclear genome (Fig. 2b). The lowest proportion of mtDNA reads, about 1%, was from the exoB DNA samples (Fig. 2b). We observed around 5% of mtDNA reads in the exo- DNA samples (Fig. 2). The proportion of mtDNA reads was up to ten times higher in the exoA DNA samples (Fig. 2), with values ranging from 17,569 to 40,888 reads, i.e. 30 to 59% of the total sequences from a given sample. Upon inspection of the location of reads mapping to the nuclear genome, we found that about 20% of them were mapped to NUMTs. The effect of the exonuclease treatment and the amplification on the mapping of reads to mtDNA versus nuDNA was significant (Chi2 = 12, 552.68, $p < 0.001$). The distribution of read size and quality was similar whether they mapped to the mitochondrial or to the nuclear genome (SI Figure 2). In conclusion, the protocol we developed to amplify mtDNA led to a significant increase in the proportion of long sequences derived from the horse mitochondrial genome.

Overall coverage of the mitochondrial genome

We then looked at the coverage of the mitochondrial genome at the base level (Fig. 3). Whatever the treatment, we found that the coverage of the mitochondrial genome was complete and this pattern was reproducible between the samples. We could observe the effect of some primers on the coverage plot that appeared as bulges, for example between positions 13,946 and 14,748. For samples untreated by exonuclease in the S1 sequencing run, the minimum coverage was 66 reads at a given mitochondrial base (SI Table 1). In the same run and for the samples treated with exoA the minimum coverage was 526 reads. The variation between minimal and maximal coverage was between 2.67 (sample H26 exo-) and 5.33 (sample H27 exoB). In a second sequencing run, the coverage was more homogeneous between DNA samples with variations of 3.26–3.97 between minimal and maximal coverage, a minimum coverage of 936 and a maximum of 11,707 (SI Figure 3). In conclusion, the read coverage of the mitochondrial genome was complete and unbiased in our tested conditions.

Long-read distribution and coverage of the mitochondrial genome

Since we obtained a good coverage of the mitochondrial genome, we wondered about the proportion of long-reads covering a significant portion of the mitochondrial DNA. In H25, H26 and H27 exoA samples, we found respectively 52, 144 and 116 reads longer than half the mitochondrial genome while we only found between one and seven such reads in the exo- and exoB samples (Fig. 4, SI table 1). In about 10% of these long-reads, we observed that the longest alignment between the read and the mitochondrial genome was remarkably smaller than the read length even though we selected reads for which the alignment coverage on mtDNA exceeded 80% (Fig. 4a). Upon inspection of these cases, we identified two categories of such reads. The first category corresponded to reads spanning more than one full revolution of the circular mtDNA. For example, we obtained five reads longer than the mitochondrial genome, one of which measuring 17,501 nt and spanning more than one full turn with about 1 kbp of additional coverage (Fig. 4). The second category of reads were usually compound and made of two inverted repeated sequences. These could be reads originating from the sequencing of both strands of the same molecule, one after the other like in the 2D or 1D2 approaches developed by ONT, with usually the second strand being incomplete. Such reads could correspond to non-chimeric or chimeric DNA molecules produced during WMGA. Altogether, these long-reads covered the three horse's mitochondrial genome at 30.8X, 87.8X and 70.1X (Fig. 4b). In conclusion, we obtained very long-reads from the exonuclease treated DNA spanning large portions of the mitochondrial genome up to its entirety.

These results of the first sequencing run prompted us to reproduce in a second set of experiments the analysis of the same set of horse DNAs. We prepared sequencing libraries using the LSK109 chemistry from 12 new samples treated using the exoA conditions. We changed the primer pool composition, excluding three primers

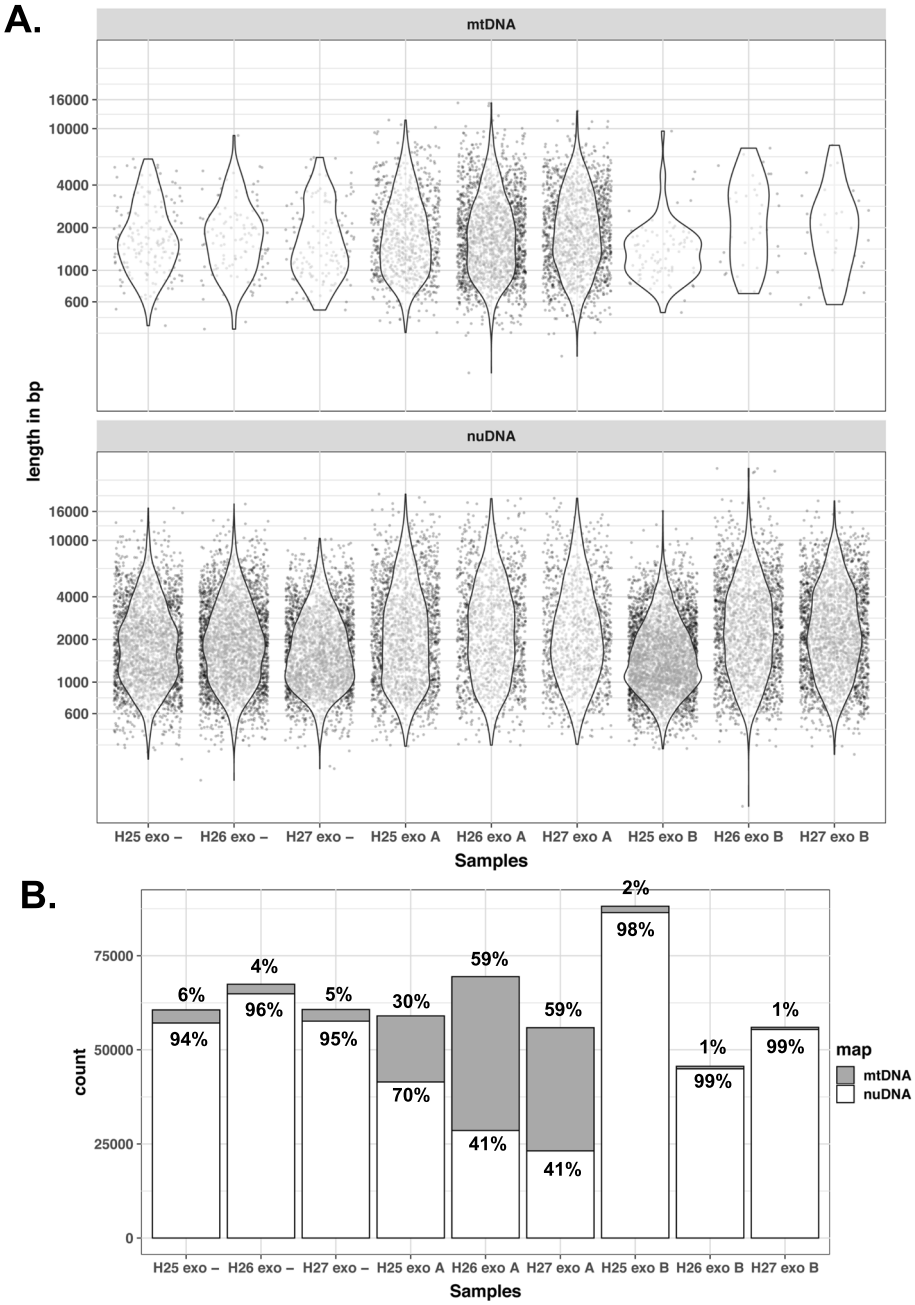

Fig. 2 Read length distribution and proportions of sequences mapping to the mitochondrial or nuclear genomes. **a** Distribution of reads length mapping to mtDNA (top) or nuDNA (bottom). For clarity, values are shown for a randomly selected subset representing 0.005% of reads. **b** Numbers and proportions of sequences mapping to mtDNA or nuDNA. The number of mapped sequences is shown on the y axis. exo-: no Exonuclease V treatment; exoA: Exonuclease V 10 U 2 h + REPLI-g 0.2 μM primers; exoB: Exonuclease V 10 U 2 h + REPLI-g 1.0 μM primers

matching to numerous NUMTs loci in the nuclear genome, performed the exonuclease treatment in duplicate and made a technical replicate of each experiment to have a better readout of the factors that could affect the efficiency and the reproductibility of our protocol. We multiplexed the 12 resulting samples for sequencing on the same flow cell. We obtained more reads from this second sequencing run and therefore a much higher coverage of the mitochondrial genome (SI table 1). The fraction of mitochondrial reads ranged from 16 to 20%, with the exception of two technical replicates reaching 50%. Like previously, the mitochondrial coverage was unbiased (SI Figure 3). The mitochondrial genome coverage by long-reads (> 8000 nt) varied between 118X

Table 1 Sequencing statistics. H25, H26 and H26 are the three different horse's DNA analysed. Barcode01 to Barcode09 are the names of the barcodes used in the multiplexing with the kit EXP-NBD-104. The category Unclassified refers to reads that were filtered out at the demultiplexing steps because the barcode sequence was not attributed. exo -: no Exonuclease V treatment; exo A: Exonuclease V 10 U 2 h + REPLI-g 0.2 μM primers; exo B: Exonuclease V 10 U 2 h + REPLI-g 1.0 μM primers

Sample	**H25 exo -**	**H26 exo -**	**H27 exo -**	**H25 exo A**	**H26 exo A**	**H27 exo A**	**H25 exo B**	**H26 exo B**	**H27 exo B**		
Barcode	**Barcode01**	**Barcode02**	**Barcode03**	**Barcode04**	**Barcode05**	**Barcode06**	**Barcode07**	**Barcode08**	**Barcode09**	**Unclassified**	**Total**
Mean read length (nt)	2209	2248	2004	2733	2526	2553	1821	3241	2869	2145	2379
Median read length (nt)	1706	1763	1579	1918	1815	1865	1440	2336	2172	1545	1741
Read length N50 (nt)	2786	2807	2462	3867	3409	3414	2172	4635	3873	2909	3137
Number of reads	60,619	67,448	60,729	59,028	69,453	55,892	88,175	45,641	55,996	71,073	634,054
% of Total reads	10%	11%	10%	9%	11%	9%	14%	7%	9%	11%	100%
Total bases (nt)	133,906,674	151,599,799	121,723,567	161,303,032	175,444,343	142,699,042	160,526,846	147,900,066	160,661,882	152,461,373	1,508,226,624
% of Total bases	9%	10%	8%	11%	12%	9%	11%	10%	11%	10%	100%
Mean read quality	11.5	11.5	11.5	10.9	10.9	10.9	11.3	11.4	11.5	6.8	10.8
Median read quality	11.7	11.8	11.7	11.1	11.2	11.1	11.6	11.7	11.8	5.8	11.3

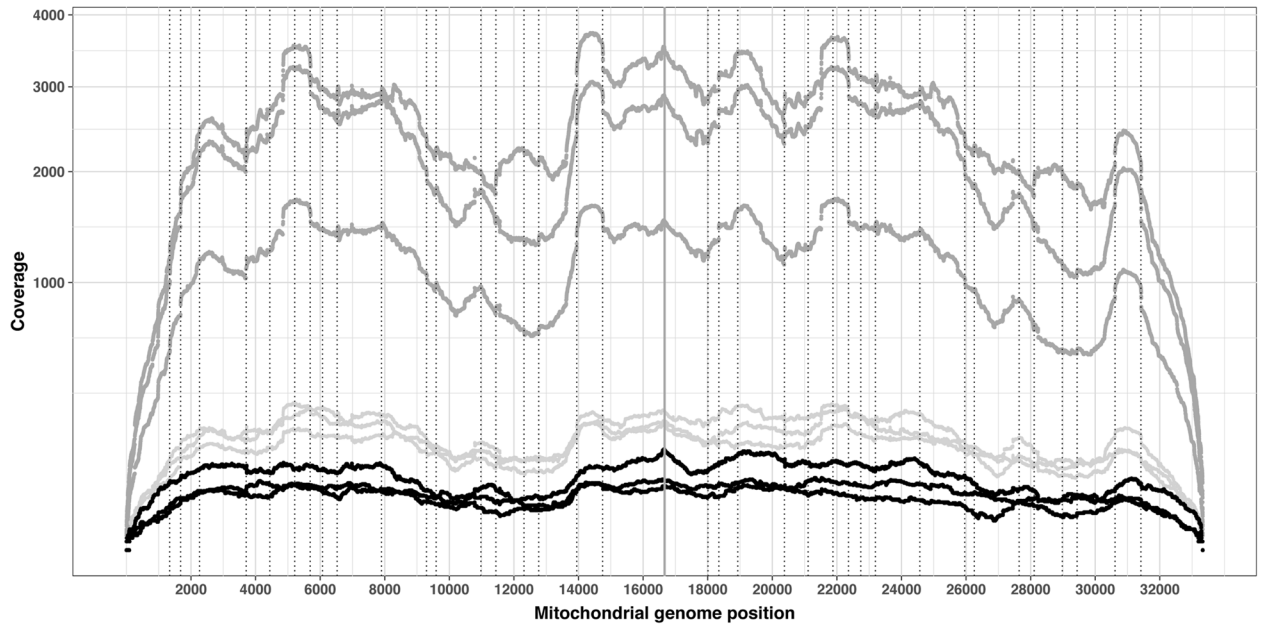

Fig. 3 Sequence coverage of the mitochondrial genome. The base position coverage is shown on the y axis. Sequences derived from the exo minus samples are shown in light grey, exo plus A are in dark grey and exo plus B are in black. Coverage was computed from mapping sequences to a duplicated mitochondrial genome. The grey vertical line at position 16,666 represents the end of the reference mtDNA sequence. Dotted vertical lines indicate the starting positions of the primers used for MDA

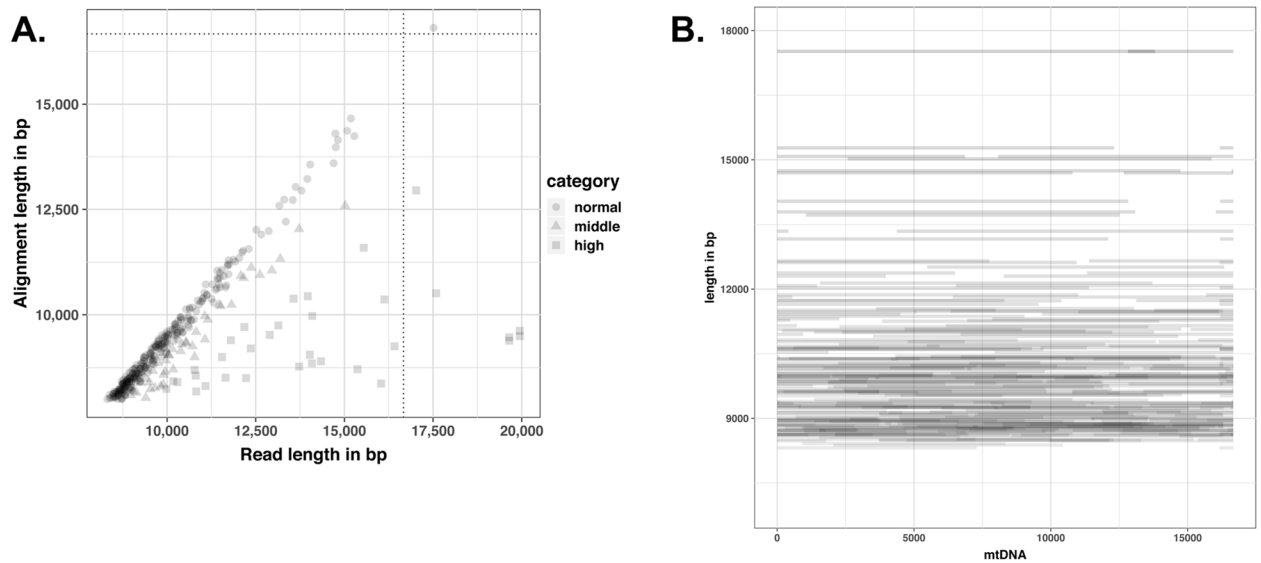

Fig. 4 Mitochondrial long-reads. **a** Scatter plot of alignment and read lengths. Dotted lines represent the size of the mitochondrial genome. Each square, triangle or dot represents one read satisfying the following criteria: the read aligns to more than 8000 bp of the mtDNA reference sequence and more than 80% of the read aligns to the reference mtDNA sequence. Circles, triangles and squares label different read categories correspond to the ratio of read length over alignment length: « normal » for values less than 1.1, « middle » for values between 1.1 and 1.2, « high » for values over 1.2. Values higher than 1 mean that there are mutiple alignments between the read and the mitochondrial genome; possibly because the read spans more than one full turn of the mtDNA when the read is longer than the mtDNA or when it is « 2D like ». Most of these reads align as expected over their whole length to a single portion of the mtDNA reference sequence (circles that can be seen along the diagonal). A small portion of these reads (triangles and squares) show more complicated alignment figures with the mtDNA reference sequence. We called some reads « 2D like » because their sequence encompass the same mtDNA segment on the same or opposite orientation. **b** Long-read coverage display on the horse mitochondrial genome. Each horizontal grey bar corresponds to one read satisfying the criteria defined in **a**, and is displayed according to its length on the y-axis. The read matching mtDNA region is indicated by the span of the bar on the x-axis (mtDNA bp coordinates). Remark the longest read (17.5 kbp) that spans more than the length of the mtDNA with the region covered twice highlighted by the dark grey color

and 488X, and because the sequencing yield by sample was higher, we also obtained significant coverage by reads longer than 16 kb, from 2 to 40X (SI Table 1, SI Figure 4). In conclusion, we obtained reproducible results in terms of mitochondrial reads enrichment, mitochondrial genome coverage and long-reads.

Mitochondrial DNA variants calling

We then performed variants calling on all sequencing data sets obtained from the three different horses DNA, i.e. from the seven sequencing libraries (three from the first sequencing experiment and four from the second sequencing experiment). For this purpose, we used sequence reads of 2000 nt or longer and satisfying a minimal quality threshold of 10, since we obtained a deep coverage and we wanted to avoid reads potentially derived from NUMTs. We used the medaka variant calling method and identified a total of 88 variable positions based on the same Arabian breed mtDNA reference as before: 77 for H25 samples, 10 for H26 and 83 for H27 (Fig. 5a). All variants were transitions, and all were found in at least one known horse mtDNA sequence available in GenBank. The majority of these variable positions were identified in the seven sequencing libraries prepared from the same DNA: 73/77 for H25 (94.8%), 10/10 for H26 (100.0%) and 77/83 for H27 (92.8%). There were 52 coding and 36 noncoding SNPs, 26 of which from the control region. The remaining positions were usually of low quality and were counted as false negatives, because they were missed in at least one sequence sample (see below).

We reasoned that a reference allele bias effect could impact the quality of variant calling, especially in the highly polymorphic and repetitive control region [35]. Therefore, we performed the same variant calling procedure using two alternative references: a Thoroughbred breed more closely related to H25 and H27, and the Twilight horse (a Thoroughbred) used for the equine reference genome. The results are presented in Fig. 5 and details are provided in SI Table 2. We identified 82 variants when we used the Thoroughbred reference mtDNA sequence: 4 for H25, 74 for H26 and 9 for H27

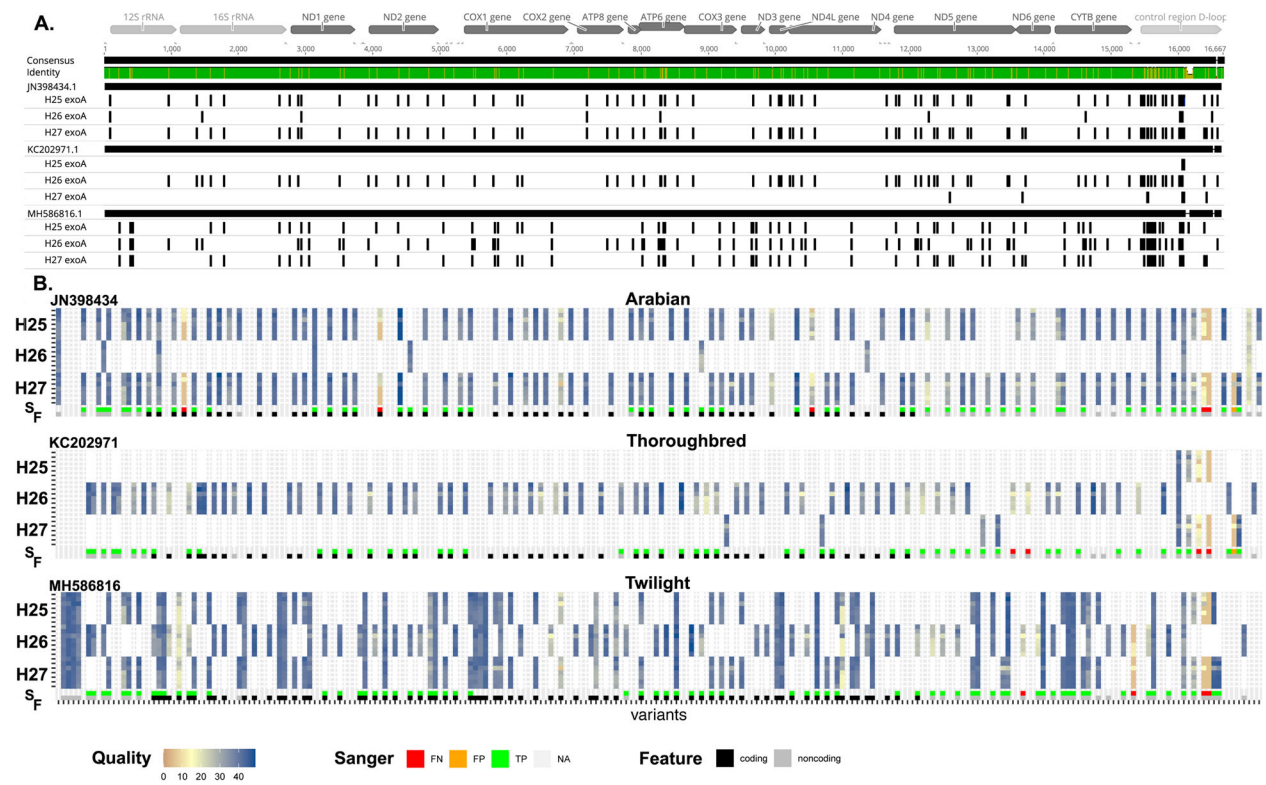

Fig. 5 Variant calling from long mtDNA reads. **a** The top panel represents the mitochondrial genome with the scale in base pairs and its rRNA, tRNA, gene and control region features. The three mtDNA reference sequences were aligned to visualise their differences. Remark the significant differences in the control region. The position of the identified polymorphisms in the H25, H26 and H27 exoA samples are indicated just below as small black vertical lines. **b** The lower panel is an heatmap showing all variants identified along the mtDNA (from the left to the right) for the different samples indicated and the three different references (from top to bottom). Variants are coloured according to their quality score or are shown in white when they were not found in a given horse DNA. SNPs categories defined by comparison with Sanger sequencing are shown below the heatmap: false negatives (FN, in red), false positives (FP, in orange) and true positives (TP, in green). Coding SNPs are shown in black in the last row, while non-coding SNPs are shown in grey

(Fig. 5b). Similarly, most variants were found in all seven sequencing libraries from the same horse: 2/4 for H25 (50.0%), 72/74 for H26 (97.3%) and 7/9 for H27 (77.8%). We found 50 coding and 32 noncoding SNPs, including 23 in the control region (Fig. 5b). We identified 82 variants when we used the Twilight reference mtDNA sequence: 66 for H25, 85 for H26 and 70 for H27 (Fig. 5b). Again, most variants were found in all seven sequencing libraries from the same horse: 63/66 for H25 (95.4%), 82/85 for H26 (96.5%) and 66/70 for H27 (94.3%). We found 50 coding and 44 noncoding SNPs, including 31 in the control region (Fig. 5b). As expected, the distribution of variants was globally the opposite between the one found using an Arabian breed mtDNA reference and the one found using the Thoroughbred, but the latter was characterized by higher variant qualities (compare qualities on Fig. 5b). The distribution of variants found using Twilight as a reference was more even between the three horses. We found the same relative abundances of polymorphisms in coding and noncoding sequences, and as expected the density of polymorphisms in the control region was the highest.

We evaluated the validity of these polymorphisms by sequencing nine PCR fragments using Sanger sequencing as ground truth. We could thus compare 7493 bp read using Sanger and Nanopore sequencing. The overall sequence identity between the Sanger and the Nanopore sequences was 99.89%. We made a concordance analysis based on variants encompassed by both Sanger and Nanopore sequencing (Table 2 and SI Table 2). The values were affected mainly by the number of variants and by the reference being used, and the results obtained using Twilight are globally more meaningful. Overall, we can underscore that the precision was excellent with a single case of false positive observed on H27 with the Arabian and the Thoroughbred reference. We looked at each sample individually to further analyse the between sample variation of true positive and false negative variant calling (SI Table 2). Overall, the mean true positive rate were 93.42, 93.33 and 94.75% with a false negative rate of 5.71, 5.00 and 5.02% using Arabian, Thoroughbred and Twilight references, respectively. These values underscored the critical impact of the reference sequence on variant calling. We observed that all cases classified as false negatives because they were missed in at least one sample were detected in at least another sample from the same horse DNA and according to the same quality criteria (Fig. 5). There were only two variations that were always missed using Nanopore reads. The first, detected using the Arabian reference, was a short deletion in the control region at position 16,402 in an array of cytosines. The second, detected using the Thoroughbred reference, was a dinucleotide polymorphism in a minisatellite repeat of the control region at position 16,125-16,126. This meant that there is no fundamental technical limitation to detect these variants using currently available nanopore long read sequencing, it is just a matter of thresholds and reproducibility in sequence quality in these regions.

We studied in more details the polymorphisms of the control region. While seven variants turned as false negatives or false positives using the Arabian reference, there was only one difficult position identified using the Thoroughbred reference (Fig. 5). This difficulty lies at the 5′ end of a minisatellite segment made of 29 GTGC ACCT repeats. Upon close inspection, both Sanger and Nanopore sequencing confirmed that the first motif of the minisatellite varied (GTGCACCT- > GCACACCT). Yet there were many long reads that aligned with deletions in the minisatellite region in a reference-dependent manner, leading to lowest prevalence of the correct sequence and to an absence or a low quality variant call (SI Figure 5).

In conclusion, the concordance for SNP calling was very good and depended mostly on the reference used and on variations in the mtDNA control region, a notoriously difficult region to sequence.

Discussion

We combined two previously described approaches to enrich the fraction of sequence reads derived from mtDNA, namely nuclear DNA removal using an exonuclease treatment and amplification using MDA [33, 34]. We found that these two techniques require optimisation steps to obtain a significant mtDNA selective amplification factor. For nuclear DNA removal, the exonuclease V amount and incubation time are likely to be dependent on the source and the quality of the input DNA. Using this step, we obtained a 3000 fold enrichment of the mtDNA/nuDNA ratio, so that theoretically the relative mass

Table 2 Concordance analysis

	Arabian (JN398434)			Thoroughbred (KC202971)			Twilight (MH586816)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
H25	1.000	0.904	0.949	1.000	0.500	0.667	1.000	0.795	0.886
H25	1.000	1.000	1.000	1.000	0.957	0.978	1.000	0.955	0.977
H27	0.980	0.909	0.943	0.857	0.750	0.800	1.000	0.925	0.961
All	0.982	0.915	0.947	0.980	0.926	0.952	1.000	0.931	0.964

amount of mtDNA in the whole cellular DNA increased from 0.63% (considering 1000 mitochondrial genome per cell) to 99.99%. This value is in line with the report of Jayaprakash and colleagues [33].

We obtained on the order of a 100,000 fold enrichment after MDA on DNA untreated by exonuclease V, i.e. about ten times more than the value of 100 to 10,000 reported by Marquis and colleagues [34]. We used the same reagents as Marquis and colleagues, with the exceptions of primers and DNA, and the two primer concentrations we used in REPLI-g lead to similar enrichment. This difference of a factor of ten may be explained by a difference between human and horse mtDNA/nuDNA ratio, by the qPCR quantification method and by the DNA isolation procedure, the latter being known to impact the quantification of mtDNA copy number [36]. We observed that there were less mtDNA reads in samples treated using the exoB condition (1 μM of primers) versus the exoA condition (0.2 μM of primers). A possible explanation is that higher primer concentration could lead to non-specific amplification, including template independent amplification, a well-known issue in MDA [37].

A concern related to MDA is its propensity to produce chimeric molecules, an artefact found in 3 to 6% of reads derived from Illumina sequencing [38, 39]. We think that it is unlikely that chimeric reads affect variant calling since chimerism frequency is more than ten times lower than the prevalence of variants. Moreover, such chimeric reads were found to be mostly formed during Illumina sequencing library constructions [38].

We obtained a ratio of mtDNA to nuDNA reads of up to 59%, but this varied for the same starting DNA between different MDA. Still, we obtained a minimum of 16% of reads derived from mtDNA, and this provided an even coverage of the mitochondrial genome. These results are similar to what is reported using some commercial solutions but less than those obtained using PCR on mtDNA extracted using a plasmid preparation kit [40]. The principal difference here lies in the horse NUMTs, because most reads mapped to nuDNA originated from NUMTs.

We used a set of primer to amplify horse mtDNA that were uniformly distributed along the mitochondrial genome, and we did not observe a significant effect on the coverage. The absence of a strong coverage bias lead by mtDNA amplification using MDA was also found by Marquis and collaborators [34]. In the future, it could be desirable to reduce the number of primers because this could enable the production of larger DNA molecules obtained from MDA. Alternatively, random primers could be used for de novo mtDNA sequencing like in the work of Simison and colleagues [41].

We obtained a sequencing depth of at least 1000X that should allow a much higher level of multiplexing. At the time of initial writing, multiplexing by ligation was limited to 24 but the possibility to multiplex 96 samples using the same protocol is now a reality. According to Quail and colleagues, heterozygous variant calling with a 15x coverage using Ion Torrent or various Illumina datasets was comparable to a coverage of 190x for PacBio datasets [42]. We can deduce conservatively that a similar coverage using Nanopore datasets would be enough to call mitochondrial variants in an efficient manner. This means that we could multiplex four or five times more samples, i.e. a hundred, in a single run. With such a capacity, this method could become very competitive in terms of rapidity, scalability and cost-effectiveness.

We estimated at 99.9% the precision of the obtained sequence, which is satisfactory given that the read median identity was estimated at ~ 89% for 1D sequencing using R9.0 chemistry. The process of MDA is known to be highly accurate and can not be considered as the limiting factor when it comes to long-read sequencing. But at the same time we observed a variant false negative rate of ~ 5% that stems from nanopore sequencing imprecisions around homopolymers of small sizes or around micro or minisatellite segments. The precision is likely to increase through improvement of the sequencing chemistry, like the recent introduction of the R10 flow-cells that combines different pores, and by the improvement of the algorithms used for base-calling. We detected only SNPs using our workflow, and the current level of sequencing precision on R9.4 flowcells is likely to preclude the identification of small indels. Further development of variant calling pipelines may alleviate this limitation, and specific tools to detect structural variants could leverage on the added value of long-reads. Another limitation is that the medaka pipeline is tailored for diploid genomes, but suitable models for variant calling on haploid such as FreeBayes could be used.

The overall quality of variant calling was improved when we used an alternative reference sequence to the point where we did not observe false negatives or false positives in coding regions. All the variants we found were transitions (A- > G or C- > T) and it is known that transitions are dominant among single nucleotide polymorphisms. This effect is more pronounced in mitochondrial genomes but also prevalent in nuclear genomes, both in coding and non-coding regions [43, 44]. The initial explanations advanced to explain this observation were that transitions are more conservative changes, but it seems to be a matter of discussion [45]. Nevertheless, the transition to transversion ratio is a useful tool for quality control in variant calling [46].

It turned out that the single variants that could be missed were in the control region, in a particularly difficult region containing a minisatellite. We found that the main factor involving variant calling accuracy was the

choice of the reference used for mapping the reads. We can foresee that the optimisation of mapping parameters in conjunction with the filtering of alignments could dramatically improve the robustness of variant calling [47, 48]. Actually, some level of heteroplasmy involving this minisatellite sequence like previously described in rabbit and other species could explain the lower level of quality of variant sequences in this region [49].

Previous studies have reported the use of Nanopore sequencing to sequence mitochondrial genomes, usually by sequencing long amplicons [50, 51]. In a recent study, Lindberg and colleagues evaluated variant allele frequencies from single or mixed DNA sources using both Illumina and Nanopore sequencing of two overlapping amplicons. In comparison to our findings, they reported a higher precision and recall for single DNA samples [52]. Similarly, the added value of long-read sequencing of amplicons for sorting out complex heteroplasmic conditions was leveraged by White and colleagues [50]. Only few reports mention in some details the prevalence of mtDNA reads in whole genome shotgun nanopore sequencing. In their report on Nanopore DNA sequencing aboard the International Space Station, Castro-Wallace and colleagues found that 29 out of 83,332 reads (0.03%) from mouse DNA were derived from the mouse mtDNA [53]. In another study, Zascavage and colleagues relied on Nanopore long-read sequencing for mtDNA analysis [54]. They reported an average coverage of mtDNA reads that varied between 15 and 118 fold from a whole genome sequencing analysis. In their study, the lower coverage limit was inadequate for accurate variant calling, and the coverage range was too large to warrant a routine use.

Conclusions

Overall, we provided a detailed and benchmarked method that should allow mtDNA genome variant identifications not only for horse geneticists but for other animal geneticists.

The identification of causal mtDNA polymorphisms require accurate variant discovery methods. The capacity to phase mtDNA variants, to identify structural variations and nucleotide modifications is a major asset of Nanopore long-read sequencing. The possibility to do so rapidly and at a relatively modest cost is of added interest, and may be of interest for field studies since the MinION sequencer is portable and able to perform real-time sequencing and base-calling [55]. At present, the technology is still hindered by the limited level of sequence accuracy but this is rapidly changing as sequencing chemistry and base-calling algorithms are evolving [56, 57]. The variety of flow cells and equipment available enables to better tailor the sequencing tools according to the number of samples. We believe that this methodology is easily amenable to sequence circular DNA molecules of similar sizes and even beyond, such as plasts, viral circular DNA or endogenous circular DNA [58].

Methods

Nuclear DNA depletion

We used DNA samples obtained in the course of a previous project and kindly provided by Dr. Céline Robert [59]. The DNA were initially extracted from blood samples of H25 (female, Selle Français), H26 (female, Thoroughbred) and H27 (male, Selle Français). We tested nuclear DNA removal by starting with 1 μg of total DNA using either 10 or 20 Units of Exonuclease V (New England Biolabs) and an incubation of 30 min or 2 h at 37 °C. We performed the enzymatic digestions in 30 μL reactions containing 1 mM ATP in NEB Buffer 4. We stopped the reaction by adding 0.8 μl of 0.5 M EDTA followed by heat inactivation at 70 °C for 30 min. DNA was recovered by precipitation with 4 μl of 3 M NaCl, 2 μl of 5 mg/ml glycogen and 150 μl of 100% ethanol. DNA was finally resuspended in 40 μl of nuclease-free water and stored at − 20 °C until used.

Whole mitochondrial genome amplification (WMGA)

We used the REPLI-g mitochondrial DNA kit (Qiagen) to perform WMGA according to the manufacturer's instruction with the following modifications. We designed 18 primers using CLC Genomic Workbench v6 (CLC bio) based on a horse mitochondrial genome (GenBank KX669268.1; see Table 1). We adopted the following recommendations for primer design: a length of 10–14 bases, the same number of primers hybridize to each strand; primer hybridization sites are evenly spaced and primers include phosphorothioate modifications between the last three bases of their 3′ end. All primers were purchased from Eurofins Genomics. We tested two primer concentrations: 1 μM and 0.2 μM of each oligonucleotide. The reaction set-up was as follows: 4 μl of DNA treated with Exonuclease V was loaded into a 0.2 ml PCR tube containing 16 μl of RNase DNase free water. 27 μl of REPLI-g mt reaction buffer and 10 μl of horse mitochondrial primer mix (1 μM or 0.2 μM each primer) was added to the DNA and the mixture was incubated at 75 °C for 5 min. We added 1 μl of REPLI-g Midi polymerase to the DNA and incubated at 33 °C for 8 h. Inactivation was done at 65 °C for 3 min. The concentration of amplified DNA was quantitated with Quant-iT™ PicoGreen dsDNA reagent (Molecular Probes) on an Infinite M200 Pro (Tecan). We followed published recommendations to avoid possible contaminations [60].

qPCR experiments

We used qPCR to estimate the ratio of mtDNA over nuclear DNA (nuDNA). The amount of mtDNA was evaluated using three amplicons from the *Cox1*, *Cox3* and *Nd2* genes. We also used three nuclear gene loci (*Epo*, *Myod1* and *Gapdh*) to quantify nuDNA amounts. All primers are given in SI Table 3. As recommended in the REPLI-g mitochondrial DNA kit handbook, the amplified DNA was diluted 1:1000 prior to qPCR. We made two technical replicates for each biological sample, and the same qPCR run contained all samples for a given gene to minimize inter-experimental variations. The 20 μL qPCR reactions were assembled manually. Reactions included 5 μL diluted DNA, 10 μL 2X powered Sybrgreen PCR master mix (Applied Biosystems) and 5 μL primers mix (300 nM each primer). Reactions were incubated in a 96 well optical plate (Applied Biosystems) at 95 °C for 10 min, followed by 40 cycles at 95 °C for 15 s and 60 °C for 1 min in a QuantStudio 12 K Flex instrument (Applied Biosystems).

The qPCR data were analysed using the QuantStudio 12 K Flex software (Applied Biosystems). Cq values were obtained using the auto baseline option and the threshold was manually set to the same value of $\Delta Rn = 0.1$ for all amplicons. The Cq values were considered valid if the duplicate measures differed by less than 0.5 cycles, undetermined values were set equivalent to a Ct value of 40. We computed the geometric mean of the Cq values of the three mtDNA and nuDNA amplicons, and then we computed the corresponding fold changes. Formation of primer-dimer and amplification specificity were assessed by melting curve analysis.

Long-read sequencing library preparation

Prior to the library construction, we purified the input DNA samples using a 0.5 volume ratio of Agencourt® AMPure® XP magnetic beads. We quantified DNA using a Qubit dsDNA BR assay and performed a debranching treatment on 8–10 μg of DNA in 100 μL reactions containing 50 Units of endonuclease T7E1 (New England Biolabs) in 1X NEB buffer 2, with an incubation of 2 h at 37 °C. We purified the debranched DNA using one volume of Agencourt® AMPure® XP magnetic beads and checked DNA concentration using a Qubit dsDNA BR assay and DNA integrity of using electrophoresis on a 12 K DNA chip (Experion, Bio-Rad).

These DNA samples were then prepared for sequencing according to native barcoding expansion (kit EXP-NBD-103, ONT) and the 1D Native barcoding genomic DNA protocol (kit SQK-LSK108 for the S1 run and SQK-LSK109 for the S2 run, ONT) with some modifications. We treated 2 μg debranched DNA using the UltraII end repair/dA-tailing enzyme mix (New England Biolabs) followed by a purification using a 0.6 volume ratio of AMPure beads. Native barcodes were added by ligation using the Blunt/TA ligase master mix (New England Biolabs) and the ligation reaction was purified using 1 volume ratio of AMPure beads. Equimolar amounts of each barcoded sample were pooled before adapter ligation using the NEBNext Quick ligation Module (New England Biolabs) and a final purification using a 0.4 volume ratio of AMPure beads with the S Fragment buffer.

For the S1 sequencing run, we used DNA amplified with the 18 primers and multiplexed nine samples using barcodes one to nine. We loaded 195 ng of sequencing library. In this S1 run, each of the three horse's DNA libraries was derived from three different conditions of exonuclease and amplification, so that each horse's DNA was independently sequenced three times. For the S2 sequencing run, we used DNA amplified with 15 primers, excluding three that were found to match NUMTs loci, and multiplexed 12 samples using barcodes (barcodes one to twelve). We loaded 439 ng of the sequencing library. In this S2 run, we performed each exonuclease treatment in duplicate and each amplification in duplicate, so that each horse's DNA was independently sequenced four times. Altogether, each horse's DNA was sequenced across seven libraries.

We performed the sequencing using R9.4 flowcells from two different batches on a MinION sequencer (ONT) operated using MinION Release v18.05.5 for the S1 run and 19.05.0 for the S2 run. The MinION sequencing device was connected to a computer running the linux operating system Ubuntu 16.04 LTS.

Data analysis

We used guppy for basecalling with the flip-flop model, trimming and homopolymer correction (guppy-cpu-v2.3.1). We then used guppy_barcoder for demultiplexing. For variant calling, we selected reads longer than 2000 bp with a mean quality higher than Q10 using NanoFilt [61]. The filtered sequences were then aligned using minimap2 on the *Equus caballus* genome reference assembly (GCF_002863925.1_EquCab3.0) [62]. We substituted the mitochondrial genome present in this assembly by a duplicated, pseudo-circularized, mitochondrial genome (GenBank JN398434) to increase the recovery of mapping to the circular mitochondrial genome. We used bioawk and samtools to extract mitochondrial reads and compute their quality scores [63]. These mtDNA reads were aligned with minimap2 on an Arabian breed reference horse mtDNA (JN398434), a Thoroughbred reference mtDNA (KC202971) or on the mtDNA from the reference Thoroughbred horse genome Twilight (MH586816). We selected medaka over nanopolish as the pipeline to call variant because our first tests showed that it required less computation time,

and finally filtered the SNPs using a quality cut-off of 15. All visualisation scripts were written using the ggplot2 package in R3.6.0 [64, 65].

Sanger sequencing for SNP validation

We designed nine primer pairs to amplify nine fragments evenly spaced along the horse mtDNA (see Table 1). PCRs were performed on 4 ng of the same DNA used for long-read sequencing and using the 2X HotStartTaq Master Mix (Qiagen). PCR products were checked by agarose gel electrophoresis and purified using Nucleospin Gel and PCR clean-up kit (Macherey-Nagel). Cycle sequencing reactions were performed using the BigDye 3.1 chemistry, purified by precipitation and sequenced using an ABI 3130 sequencer (Applied Biosystems). Base calling was performed using the Sequencing analysis software version 5. 3 (Applied Biosystems).

The assessment of true positives (tp), false positives (fp) and false negatives (fn) was performed by inspecting the results of Sanger sequencing (on both strands) and of variant calling using the Geneious 11.1.5 software (https://www.geneious.com). We used a strict rule for our concordance analysis: we declared a SNP as a false negative or false positive if a difference was observed in one or more of the samples sequenced for a given horse, and even if there was a concordance in one or another sample. For each sequence run and each horse's DNA we counted the occurences of tp, fp and fn. We then computed the values for the precision $p = tp/(tp + fp)$ and the recall $r = tp/(tp + fn)$. The F1 score was computed as the harmonic mean of precision and recall:

$$F1 = 2*\left[\frac{precision*recall}{precision + recall}\right]$$

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-020-07183-9.

Additional file 1: SI_table 1. Sequencing statistics, coverage, long-reads.

Additional file 2: SI_table 2. Sanger resequencing informations.

Additional file 3: SI_table 3. Primers.

Additional file 4: SI_Figure 1. DNA quantity and fragment size after REPLI-g.

Additional file 5: SI_Figure 2. Distribution of raw read length and quality by barcode.

Additional file 6: SI_Figure 3. Proportions of mtDNA and nuDNA sequences and sequence coverage in a second sequencing run.

Additional file 7: SI_Figure 4. Distribution of long reads in a second sequencing run.

Additional file 8: SI_Figure 5. Comparison of long-read and Sanger sequencing in the control region.

Acknowledgements

The authors would like to thank Prof. Céline Robert for the kind gift of horse's DNA.

Authors' contributions

S.D-P, E.B., N.P. designed the experiment. S.D-P. performed the DNA pre-treatments (Exonuclease treatment and MDA), qPCR analyses, PCR amplifications in S.D. and E.B. laboratory. S.D-P. and N.P. performed sequencing library preparations in N.P. laboratory. N.P. performed the Sanger sequencing experiments. N.P. carried out the bioinformatics analyses. N.P. drafted the main text and prepared figures with inputs from S.D-P and E.B. All authors reviewed the manuscript and approved the final version.

Funding

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are available at the EMBL European Nucleotide Archive PRJEB35937 and the European Variation Archive (EVA).
The other datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Horse's DNA samples were obtained in the course of a previous project [59].

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

Author details

[1]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France. [2]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198 Gif-sur-Yvette, France.

Received: 20 December 2019 Accepted: 26 October 2020
Published online: 11 November 2020

References

1. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The origin and diversification of mitochondria. Curr Biol CB. 2017;27:R1177–92.
2. Gray MW. Mitochondrial evolution. Cold Spring Harb Perspect Biol. 2012;4: a011403.
3. Nass MM, Nass S. Intramitochondrial fibers with DNA characteristics. I. Fixation and electron staining reactions. J Cell Biol. 1963;19:593–611.
4. Smith DR. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? Brief Funct Genomics. 2016;15:47–54.
5. Kogelnik AM, Lott MT, Brown MD, Navathe SB, Wallace DC. MITOMAP: a human mitochondrial genome database. Nucleic Acids Res. 1996;24:177–9.
6. Bernt M, Braband A, Schierwater B, Stadler PF. Genetic aspects of mitochondrial genome evolution. Mol Phylogenet Evol. 2013;69:328–38.
7. Ballard JWO, Whitlock MC. The incomplete natural history of mitochondria. Mol Ecol. 2004;13:729–44.
8. Allio R, Donega S, Galtier N, Nabholz B. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. Mol Biol Evol. 2017;34:2762–72.
9. Ramos A, Santos C, Mateiu L, del M Gonzalez M, Alvarez L, Azevedo L, et al. Frequency and pattern of heteroplasmy in the complete human mitochondrial genome. PLoS One. 2013;8:e74636.
10. Pamilo P, Viljakainen L, Vihavainen A. Exceptionally high density of NUMTs in the honeybee genome. Mol Biol Evol. 2007;24:1340–6.
11. Ricchetti M, Tekaia F, Dujon B. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol. 2004;2:E273.
12. Nergadze SG, Lupotto M, Pellanda P, Santagostino M, Vitelli V, Giulotto E. Mitochondrial DNA insertions in the nuclear horse genome. Anim Genet. 2010;41(Suppl 2):176–85.
13. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 2014;42:12640–9.
14. Robin ED, Wong R. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol. 1988;136:507–13.

15. Bensasson D, Zhang D-X, Hartl DL, Hewitt GM. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 2001;16:314–21.
16. Nguyen TTT, Murphy NP, Austin CM. Amplification of multiple copies of mitochondrial cytochrome b gene fragments in the Australian freshwater crayfish, Cherax destructor Clark (Parastacidae: Decapoda). Anim Genet. 2002;33:304–8.
17. Yao Y-G, Kong Q-P, Salas A, Bandelt H-J. Pseudomitochondrial genome haunts disease studies. J Med Genet. 2008;45:769–72.
18. Maddock ST, Briscoe AG, Wilkinson M, Waeschenbach A, Mauro DS, Day JJ, et al. Next-generation mitogenomics: a comparison of approaches applied to caecilian amphibian phylogeny. PLoS One. 2016;11:e0156757.
19. Briscoe AG, Hopkins KP, Waeschenbach A. High-throughput sequencing of complete mitochondrial genomes. Methods Mol Biol Clifton NJ. 2016;1452: 45–64.
20. Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, et al. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. GigaScience. 2017;6. https://doi.org/10.1093/gigascience/giw018.
21. Ranjard L, Wong TKF, Külheim C, Rodrigo AG, Ragg NLC, Patel S, et al. Complete mitochondrial genome of the green-lipped mussel, Perna canaliculus (Mollusca: Mytiloidea), from long nanopore sequencing reads. Mitochondrial DNA Part B. 2018;3:175–6.
22. Gan HM, Grandjean F, Jenkins TL, Austin CM. Absence of evidence is not evidence of absence: Nanopore sequencing and complete assembly of the European lobster (Homarus gammarus) mitogenome uncovers the missing nad2 and a new major gene cluster duplication. BMC Genomics. 2019;20:335.
23. Gan HM, Linton SM, Austin CM. Two reads to rule them all: Nanopore long read-guided assembly of the iconic Christmas Island red crab, Gecarcoidea natalis (Pocock, 1888), mitochondrial genome and the challenges of AT-rich mitogenomes. Mar Genomics. 2019;45:64–71.
24. Barrey E, Mucher E, Jeansoule N, Larcher T, Guigand L, Herszberg B, et al. Gene expression profiling in equine polysaccharide storage myopathy revealed inflammation, glycogenesis inhibition, hypoxia and mitochondrial dysfunctions. BMC Vet Res. 2009;5:29.
25. Barrey E, Jayr L, Mucher E, Gospodnetic S, Joly F, Benech P, et al. Transcriptome analysis of muscle in horses suffering from recurrent exertional rhabdomyolysis revealed energetic pathway alterations and disruption in the cytosolic calcium regulation. Anim Genet. 2012;43:271–81.
26. Mach N, Plancade S, Pacholewska A, Lecardonnel J, Rivière J, Moroldo M, et al. Integrated mRNA and miRNA expression profiling in blood reveals candidate biomarkers associated with endurance exercise in the horse. Sci Rep. 2016;6:22932.
27. Ricard A, Robert C, Blouin C, Baste F, Torquet G, Morgenthaler C, et al. Endurance exercise ability in the horse: a trait with complex polygenic determinism. Front Genet. 2017;8:89.
28. Keyser-Tracqui C, Blandin-Frappin P, Francfort H-P, Ricaut F-X, Lepetz S, Crubézy E, et al. Mitochondrial DNA analysis of horses recovered from a frozen tomb (Berel site, Kazakhstan, 3rd century BC). Anim Genet. 2005;36:203–9.
29. Achilli A, Olivieri A, Soares P, Lancioni H, Hooshiar Kashani B, Perego UA, et al. Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proc Natl Acad Sci U S A. 2012; 109:2449–54.
30. Fages A, Hanghøj K, Khan N, Gaunitz C, Seguin-Orlando A, Leonardi M, et al. Tracking five millennia of horse management with extensive ancient genome time series. Cell. 2019;177:1419–1435.e31.
31. Wolff JN, Shearman DCA, Brooks RC, Ballard JWO. Selective enrichment and sequencing of whole mitochondrial genomes in the presence of nuclear encoded mitochondrial pseudogenes (numts). PLoS One. 2012;7:e37142.
32. Santibanez-Koref M, Griffin H, Turnbull DM, Chinnery PF, Herbert M, Hudson G. Assessing mitochondrial heteroplasmy using next generation sequencing: a note of caution. Mitochondrion. 2019;46:302–6.
33. Jayaprakash AD, Benson EK, Gone S, Liang R, Shim J, Lambertini L, et al. Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. Nucleic Acids Res. 2015;43:2177–87.
34. Marquis J, Lefebvre G, Kourmpetis YAI, Kassam M, Ronga F, De Marchi U, et al. MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. BMC Genomics. 2017;18. https://doi.org/10.1186/s12864-017-3695-5.
35. Huszar TI, Wetton JH, Jobling MA. Mitigating the effects of reference sequence bias in single-multiplex massively parallel sequencing of the mitochondrial DNA control region. Forensic Sci Int Genet. 2019;40:9–17.
36. Fazzini F, Schöpf B, Blatzer M, Coassin S, Hicks AA, Kronenberg F, et al. Plasmid-normalized quantification of relative mitochondrial DNA copy number. Sci Rep. 2018;8:15347.
37. Hutchison CA, Smith HO, Pfannkoch C, Venter JC. Cell-free cloning using φ29 DNA polymerase. Proc Natl Acad Sci U S A. 2005;102:17332–6.
38. Ellegaard KM, Klasson L, Andersson SGE. Testing the reproducibility of multiple displacement amplification on genomes of clonal Endosymbiont populations. PLoS One. 2013;8:e82319.
39. Tu J, Guo J, Li J, Gao S, Yao B, Lu Z. Systematic characteristic exploration of the chimeras generated in multiple displacement amplification through next generation sequencing data reanalysis. PLoS One. 2015;10. https://doi.org/10.1371/journal.pone.0139857.
40. Quispe-Tintaya W, White RR, Popov VN, Vijg J, Maslov AY. Fast mitochondrial DNA isolation from mammalian cells for next-generation sequencing. BioTechniques. 2013;55:133–6.
41. Simison WB, Lindberg DR, Boore JL. Rolling circle amplification of metazoan mitochondrial genomes. Mol Phylogenet Evol. 2006;39:562–7.
42. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics. 2012;13:341.
43. Lanave C, Tommasi S, Preparata G, Saccone C. Transition and transversion rate in the evolution of animal mitochondrial DNA. Biosystems. 1986;19: 273–83.
44. Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, et al. The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet. 2009;84:628–40.
45. Stoltzfus A, Norris RW. On the causes of evolutionary transition: transversion bias. Mol Biol Evol. 2016;33:595–602.
46. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2014;15:879–89.
47. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun. 2019; 10:4660.
48. Luo R, Sedlazeck FJ, Lam T-W, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10:998.
49. Casane D, Dennebouy N, de Rochambeau H, Mounolou JC, Monnerot M. Nonneutral evolution of tandem repeats in the mitochondrial DNA control region of lagomorphs. Mol Biol Evol. 1997;14:779–89.
50. White EJ, Ross T, Lopez E, Nikiforov A, Gault C, Batorsky R, et al. Chasing a moving target: Detection of mitochondrial heteroplasmy for clinical diagnostics. bioRxiv. 2017:222109.
51. Tan AS, Baty JW, Dong L-F, Bezawork-Geleta A, Endaya B, Goodwin J, et al. Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. Cell Metab. 2015;21:81–94.
52. Lindberg MR, Schmedes SE, Hewitt FC, Haas JL, Ternus KL, Kadavy DR, et al. A comparison and integration of MiSeq and MinION platforms for sequencing single source and mixed mitochondrial genomes. PLoS One. 2016;11:e0167600.
53. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, et al. Nanopore DNA sequencing and genome assembly on the international Space Station. Sci Rep. 2017;7:1–12.
54. Zascavage RR, Thorson K, Planz JV. Nanopore sequencing: an enrichment-free alternative to mitochondrial DNA sequencing. Electrophoresis. 2019;40: 272–80.
55. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat Methods. 2016;13:751–4.
56. Noakes MT, Brinkerhoff H, Laszlo AH, Derrington IM, Langford KW, Mount JW, et al. Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. Nat Biotechnol. 2019;37:651–6.
57. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19:90.
58. McNaughton AL, Roberts HE, Bonsall D, de Cesare M, Mokaya J, Lumley SF, et al. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). Sci Rep. 2019;9:1–14.
59. Robert C, Valette J-P, Jacquet S, Lepeule J, Denoix J-M. Study design for the investigation of likely aetiological factors of juvenile osteochondral conditions (JOCC) in foals and yearlings. Vet J Lond Engl 1997. 2013; 197:36–43.

60. Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl E-M. An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PLoS One. 2010;5.
61. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018; 34:2666–9.
62. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinforma Oxf Engl. 2018;34:3094–100.
63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinforma Oxf Engl. 2009; 25:2078–9.
64. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2009. https://doi.org/10.1007/978-0-387-98141-3.
65. R Core Team. R: A language and environment for statistical computing. 2019. www.R-project.org.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC