Mitochondrial sequences or Numts – By-catch differs between sequencing methods

Running title: Numts in long-read data

Authors:

Hannes Becher[1,2] & Richard A Nichols[1]

[1] Queen Mary University of London, School of Biological and Chemical Sciences

[2] Present address: University of Edinburgh, Institute of Evolutionary Biology

Abstract

Nuclear inserts derived from mitochondrial DNA (Numts) encode valuable information. Being mostly non-functional, and accumulating mutations more slowly than mitochondrial sequence, they act like molecular fossils – they preserve information on the ancestral sequences of the mitochondrial DNA. In addition, changes to the Numt sequence since their insertion into the nuclear genome carry information about the nuclear phylogeny. These attributes cannot be reliably exploited if Numt sequence is confused with the mitochondrial genome (mtDNA). The analysis of mtDNA would be similarly compromised by any confusion, for example producing misleading results in DNA barcoding that used mtDNA sequence.

We propose a method to distinguish Numts from mtDNA, without the need for comprehensive assembly of the nuclear genome or the physical separation of organelles and nuclei. It exploits the different biases of long and short-read sequencing. We find that short-read data yield mainly mtDNA sequences, whereas long-read sequencing strongly enriches for Numt sequences. We demonstrate the method using genome-skimming (coverage < 1x) data obtained on Illumina short-read and PacBio long-read technology from DNA extracted from six grasshopper individuals. The mitochondrial genome sequences were assembled from the short-read data despite the presence of Numts. The PacBio data contained a much higher proportion of Numt reads (over 16-fold), making us caution against the use of long-read methods for studies using mitochondrial loci. We obtained two estimates of the genomic proportion of Numts. Finally, we introduce “tangle plots”, a way of visualising Numt structural rearrangements and comparing them between samples.

Keywords: mitochondrial genome, plastid, Numt content, Illumina, PacBio, tangle plot

Introduction

Sequences of mitochondrial DNA have proved indispensable markers for population genetics and phylogenetics for decades (Avise et al., 1987; Ballard & Rand, 2005). More recently, numerous ecological experiments have exploited the universal animal barcoding marker, COX1, which is a mitochondrial gene (Hebert, Ratnasingham, & deWaard, 2003). One valuable property of mitochondrial sequences is that, being more abundant, they tend to be

more effectively amplified by PCR than nuclear sequences, in particular in difficult or degraded samples. Mitochondrial markers are therefore widely used in research on museum specimens (Anmarkrud & Lifjeld, 2017), ancient DNA studies (Baca et al., 2018; Mohandesan et al., 2017) and analyses of faecal samples (van der Valk, Lona Durazo, Dalén, & Guschanski, 2017). A second advantage is that, thanks to their comparatively small size (approximately 16 kbp) and conserved structure (Boore, 1999) animal mitochondrial genomes are easy to assemble. Similar considerations make plastid genomes particularly valuable in genetic analysis of plants (Twyford & Ness, 2016).

The advantages of mtDNA analysis can be negated by the presence of Numts: nuclear inserts derived from mitochondrial DNA (and in plants, the plastid equivalent, Nupts). Evidence of such insertions was first found shortly after mitochondria were discovered to contain their own genetic material (Du Buy & Riley, 1967) and it has since become clear that Numts are present in many species (Bensasson, Zhang, Hartl, & Hewitt, 2001), often in multiple copies. The abundance of Numts strongly depends on whether a species has one or more mitochondria per cell, an observation which led Barbrook, Howe, & Purton (2006) to postulate a limited transfer window.

The confusion of Numts and mitochondrial sequence could lead to incorrect interpretations of molecular genetics studies (Blacket, Semeraro, & Malipatil, 2012; Hawlitschek et al., 2017; Jordal & Kambestad, 2014; Kim, Lee, & Ju, 2013; Thalmann, Hebler, Poinar, Pääbo, & Vigilant, 2004; Zhang & Hewitt, 1997). Any study targeting mitochondrial sequences will therefore benefit from knowledge of the genomic content of Numts.

While Numts are commonly seen as a nuisance, they are fascinating study objects in their own right. Accruing substitutions more slowly than the mitochondrial DNA lineage, they act as "molecular fossils" providing information about the ancestral mitochondrial sequence (Lopez, Yuhki, Masuda, Modi, & O'Brien, 1994; Thalmann, Serre, et al., 2004). Some can be distinguished, because substitutions which have accumulated since integration into the nucleus have a high incidence of non-synonymous changes relative to mtDNA (Bensasson, Zhang, & Hewitt, 2000). Such Numts can be used as genetic markers, for example providing evidence of past episodes of hybridisation between taxa (Brelsford, Mila, & Irwin, 2011; Miraldo, Hewitt, Dear, Paulo, & Emerson, 2012).

Before high-throughput sequencing data became readily available, Numts could be detected, albeit with some difficulty, by PCR-based methods (Bensasson et al., 2000) or cytologically, by *in situ* hybridisation of mitochondrial sequences to chromosomal preparations (Gellissen, Bradfield, White, & Wyatt, 1983). To physically separate mitochondrial and nuclear DNA, (ultra) centrifugation can be used (Garber & Yoder, 1983; Lansman, Shade, Shapira, & Avise, 1981), but these methods require considerable technical effort. Since the advent of high-throughput sequencing data, two further approaches have been applied to identifying Numts. (1) In well-assembled genomes sequenced at high coverage, Numts can be detected simply by screening the assembly for regions with similarity to mitochondrial DNA (Hazkani-Covo, Zeller, & Martin, 2010). Where genomic data has been assembled for multiple individuals, as in humans, insertion/deletion polymorphism for Numts can be readily studied (Dayama, Emery, Kidd, & Mills, 2014). (2) In absence of a well-assembled genome, for instance in genome skimming studies (Dodsworth, 2015; Straub et al., 2012), Numts may be

identified if reads (or read pairs) match the mitochondrial sequence for only part of their length. Otherwise, if the whole of the read maps to the mitochondrial genome, it may be possible to classify its Numt origin if the sequence has diverged from the mitochondria. Those that have not diverged sufficiently to be distinguished are termed "cryptic Numts" (Bertheau, Schuler, Krumböck, Arthofer, & Stauffer, 2011).

In this paper, we develop a further approach that investigates Numts by making use of the “by-catch” from High-throughput sequencing. We use this term to emphasise that much sequencing data is superfluous to the aims of a specific experiment, often the huge majority. The low price of sequencing data means these data could be discarded; yet this genomic by-catch can be mined for valuable incidental information. In particular it can be used to assemble genomes of organelles such as mitochondria. The choice of sequencing platform may influence the type of by-catch produced, particularly because of differences in fragmentation and size-selection protocols. We show that this difference can be exploited to investigate the Numt content of the genome.

In order to outline our approach, it is helpful to divide the data generated from a sequencing library into fractions (see Figure 1). The first major division is between A – reads without similarity to mitochondrial sequences and ML – mitochondrial-like sequences (i.e. those which align to the mitochondrial genome). This ML fraction may conceptually be subdivided further into D – those which can be identified as Numt sequences (i.e. sequences aligning for their full length but having a different sequence, or aligning for part of their length), C – cryptic Numts indistinguishable from actual mitochondrial sequences, and M – sequences derived from actual mitochondrial genomes. Fractions A, C, and D are derived from nuclear

105 DNA (fraction N, comprising A + C + D). In our analysis we assume that their ratios are
106 effectively constant among individuals from the same species. For instance, the proportion
107 of obvious Numts in the nuclear genome D/N should be constant. The size of fraction M,
108 which is contributed by mitochondria, may differ between samples because of differences in
109 the number of mitochondria per cell with tissue, sex, or developmental stage (Fernández-
110 Vizarra, Enríquez, Pérez-Martos, Montoya, & Fernández-Silva, 2011). This will cause
111 difference in M among samples, which will be observed in differences in the ratios of A and
112 ML in different individual samples.
113 Here we demonstrate two complementary approaches for estimating the proportion the
114 nuclear genome made up of Numts. One exploits the variation in M from sample to sample
115 in short-read data, which arises because of differences in the mitochondrial composition
116 with tissue, sex, or developmental stage. Secondly in some types of long-read data M is
117 minimal, so the ratio can be calculated directly. We demonstrate these approaches using
118 genome skimming data (coverage < 1/3x) generated by short-read (Illumina's NextSeq) and
119 long-read (PacBio's RSII) platforms from the grasshopper, *Podisma pedestris*. We assemble
120 the species's mitochondrial genome sequence and calculate these two estimates of the
121 proportion of Numts in the nuclear genome. We also introduce a method, which we call
122 "tangle plots", for the visualisation of Numts with structural re-arrangements
123
124

Materials and Methods

Samples and sequencing

Information about the samples can be found in Table 1.

Illumina NextSeq

Freshly removed hindlegs of *Podisma pedestris* were snap-frozen and stored at -79 °C. Before DNA extraction, the legs were dipped into boiling water to inactivate DNases. Subsequently, the denatured femur muscle was dissected out. DNA was then extracted using a Qiagen Blood and Tissue kit following the manufacturer's instructions. Using a Covaris ultra sonicator the DNA was sheared aiming to achieve a median size of 550 bp. Libraries for sequencing were prepared using an Illumina TruSeq DNA PCR-Free kit. Sequencing was carried out at QMUL's Genome Centre on Illumina's NextSeq Platform using v2 chemistry.

PacBio RSII

Freshly removed hindlegs were stored in pure ethanol. DNA was extracted from four samples using a Qiagen Gentra HMW kit resulting in molecules with a length of mainly > 48 kbp (TapeStation, Agilent Genomics). Further work was carried out by The University of Liverpool's Centre for Genomic Research. The aimed size for DNA fragmentation was 10 kbp. The libraries' median (non-redundant) insert sizes were 3125, 3167, and 2097 bp. Sequencing was carried out on a PacBio RSII machine using C6 chemistry. All PacBio data were obtained as circular consensus sequences (CSSs) in FASTQ format.

These are of a higher per-base quality than the raw reads, because they are generated from multiple reads generated from the same circular template.

Data cleaning

Two sets of clean NextSeq data were prepared. For the RepeatExplorer analysis (see below), the data were filtered using a custom python script keeping only read pairs where 90 % of the bases had a phred quality score > 20. Pairs matching the TruSeq adapters (detected by BLASTn num_aligments 1) were discarded to remove adapter dimers.

A second cleaned set of NextSeq data was generated for mapping and variant calling. Here, we aimed to remove as many low-quality base calls as possible. The first 5 bp of each read were removed and, using Skewer (Jiang, Lei, Ding, & Zhu, 2014), each 3' end was trimmed until the last base had a phred quality 30 or higher.

For the RepeatExplorer analysis of PacBio data, pseudo paired reads of 151 bp with an insert size of 550 bp were cut out of long PacBio reads using custom-made python scripts which depend on the biopython module, http://biopython.org/ (Cock et al., 2009).

RepeatExplorer analyses

RepeatExplorer (https://galaxy-elixir.cerit-sc.cz/, Novák, Neumann, Pech, Steinhaisl, & Macas, 2013) is a pipeline for analysing the repetitive genome content from short-read genome skimming data. It performs an all-to-all comparison and generates clusters of similar reads, which often correspond to particular genomic repeats such as transposable elements or satellites. Mitochondrial genomes and ribosomal RNA genes (rDNA), which are present in high copy numbers, are usually picked up as well.

RepeatExplorer was run twice. The first run was used to assemble a reference sequence for the *Podisma pedestris* mitochondrial genome, from the NextSeq reads. In the second run, 100,000 NextSeq read pairs from each of six individuals (N1-N6) and 150,000 PacBio pseudo read pairs from each of three individuals (P1-P3) were analysed jointly to compare the sequencing methods. The pipeline was supplied with a custom annotation database containing the mitochondrial genome sequence of *Schistocerca gregaria* [Genebank NC_013240.1 (Erler, Ferenz, Moritz, & Kaatz, 2010)] in the first round and with the *Podisma pedestris* rDNA and mitochondrial genome in the second run.

Mitochondrial sequence assembly

Eight RepeatExplorer clusters connected by paired reads (244 - 57 - 230 - 205 - 69 - 85 - 102 - 161) showed sequence similarity to *S. gregaria* mitochondrial DNA. Those clusters overlapping consensus sequences were assembled in Geneious R9, forming a reference to which reads of sample N1 were mapped. High coverage and truncated reads at the control region indicated a duplication, which was then added to the reference. Subsequently, each of the six NextSeq samples' sequencing data were mapped individually using BWA (Li & Durbin, 2009) with the following command line: `bwa mem -t <no of threads> <reference> <(zcat read files)`. For each of the six alignments, 50 % majority rule consensus sequences were created in Geneious. They were annotated automatically using the MITOS WebServer (Version 2 beta, Bernt et al., 2013). Our mitochondrial genome assemblies were checked by re-assembling the short-read data with NOVOPlasty (Dierckxsens, Mardulyn, & Smits, 2016), which produced essentially the same sequences.

Mapping and variant detection

In order to detect individual-specific variants, a second round of mapping was carried out with NextSeq data. Polymorphisms were called using Geneious's function "Find Variations/SNPs" with default settings and a minimum allele frequency set to 0.01. The resulting tables were exported to CSV format and were processed interactively in R 3.3.1 (R Core Team, 2016).

All PacBio reads were aligned to the mitochondrial assembly using the LAST suite (Kielbasa, Wan, Sato, Horton, & Frith, 2011). In brief, the reference genome was masked in regions with GC-content below 10 % and was subsequently converted to a LAST database using the scoring scheme NEAR, preserving all masked regions and additionally masking simple repeats (optimised for high AT-content). Lastal was then run with parameter D set to one thousand times the length of the assembly (corresponding to an e-value of 1e-3 in BLAST). Of the resulting hits, only those with alignment lengths above 100 bp were kept. Shorter ones tended to map to in regions of low complexity, not permitting meaningful conclusions about homology.

Assessment of sequencing bias and genomic proportion of Numts

The output of the comparative (i.e. second) RepeatExplorer run was used to assess sequencing technology-specific bias (see Fig. 2).

Illumina data

Sequencing reads of the mitochondrial-like fraction (ML, see Fig. 1) could either have originated from mitochondrial genomes (M) or from Numts (fractions C and D). Assuming each of the individuals contained the same genomic proportion of Numts, the variation between samples in the relative proportion of ML in the NextSeq data (see Tab. 1) would be attributed to different mitochondrial densities in the extracts (varying proportion of M). One estimate of the proportion of Numt sequences reads can be obtained from the assumption that most mtDNA in any one individual is monomorphic, whereas some of the Numt sequences will be fixed for a different allele (because mtDNA tends to evolve faster than Numts). In this case, the frequency of this Numt allele will be proportional to the relative contribution of Numts. The maximum of the distribution of allele frequencies (shown as horizontal bands of dots in Fig. 4) provides an estimate for the relative contribution of Numts to the data mapping to each mitochondrial assembly.

This assumption is supported by the very close correlation (R^2=0.93, p=0.001) between the maximum allele frequency and the proportion of reads (A) which are not mitochondrial-like.

PacBio data

The estimate of M from the PacBio data was estimated as follows. CCSs aligning only partially (< 95%) were considered Numt-derived. Reads matching along > 95 % of their length were considered full-length matches. For these, alignment error profiles were compared to the reads' phred quality scores. If an alignment contained significantly more

mismatches than expected (5 % confidence interval, one-sided, Bonferroni-corrected for 59 alignments), it was considered a Numt sequence. The remainder of the full-length matches were provisionally classified as mitochondrial, belonging to fraction M.

Tangle plots

Code to reproduce the example shown in figure 5, as well as explanations, and distance computation can be found in the GitHub repository "tangles". See also supplemental information (https://github.com/SBCSnicholsLab/tangles), which contain explanations and another example.

Results

Six mitochondrial genome assemblies

We assembled the mitochondrial genome sequences of six individuals of *Podisma pedestris* (each of our short-read genome-skimming datasets) using contigs produced by the RepeatExplorer pipeline (Novák et al., 2013). RepeatExplorer generates "clusters" of decreasing size corresponding to repetitive DNA sequences in the samples analysed. RepeatExplorer contigs with similarity to the mitochondrial genome of the desert locust, *Schistocerca gregaria*, were merged in Geneious R9 and a 383-pb direct repeat, which had been collapsed, was adjusted manually after mapping each sample's reads back to the respective assembly. To check the reliability of this approach, we re-assembled the mitochondrial genome from each data set with NOVOPlasty, yielding essentially the same

246 sequences, the differences being around the control region; for example NOVOPlasty did not
247 assemble the repeat in 3 cases.
248 Each of our assemblies is 16,008 bp in length. The average mapping depth varies between
249 samples from several hundred to few thousand-fold, which could be due to differences in
250 cellular content of mitochondria between individuals (Tab. 1). All genes typically found in
251 animals were identified using the MITOS WebServer v2beta (Fig. 3A). The gene order is
252 collinear with other grasshopper mitochondrial genomes, and the sequences align readily
253 (see alignment in supplementary file S1).
254 The alignment of all six consensus sequences contains 18 variable sites, five of which show
255 population-specific polymorphisms (Fig. 3B). A neighbour-joining tree shows that each
256 population's individuals have sequences most similar to one other (see tree in
257 supplementary data S2). The PacBio data did not yield enough mitochondria-like sequence
258 to attempt an assembly, see below.
259

260 **The amount of ML data differs between sequencing methods**

261 The proportion of each dataset with similarity to mitochondrial sequences (mitochondrial-
262 like, ML in Fig. 1) was identified by mapping reads back to the assembly (of sample N1). It
263 can be seen from Tab. 1 that the ML fraction in short read data (proportion of reads) is at
264 least one order of magnitude larger than the ML fraction in long-read data (sum of read
265 lengths) in all samples.

High-abundance sequences differ between Illumina and PacBio data

Because of the comparatively low coverage of the genome skimming data generated, it is only possible to compare sequences that are very abundant in the libraries sequenced, such as genomic repeats and organelle DNA. In order to compare the data generated by PacBio and Illumina sequencing, we used RepeatExplorer; an pipeline that generates clusters corresponding to high-abundance sequences. If both Illumina and PacBio sequencing were unbiased in their representation of the DNA found in our samples, then each cluster should contain similar proportions of PacBio and Illumina data, corresponding to the amount of data put into RE. The proportion of short-read data in each of the 300 largest clusters is shown in Fig. 3 (on a logarithmic scale). As is commonly seen, the short-read data contain fewer sequences with higher GC content (Ekblom, Smeds, & Ellegren, 2014). While this bias is between ¼ and 4-fold for most clusters, the mitochondrial-like clusters (circles in Fig. 3) show the most extreme values (at least 16-fold enrichment in our short-read data).

Short reads: Polymorphism in individual-specific alignments

SNPs were called within each alignment of individual-specific short ML reads. The minimum minor allele frequency set to 1 % to avoid erroneous calls due to sequencing errors. All assemblies contain numerous polymorphic sites with low to medium minor allele frequencies, which can be interpreted as variants present in Numts (dots in Fig. 4). The fact that we find appreciable allele frequencies even though we sequenced only a fraction of each genome, strongly suggests that there a multiple Numt insertions present in each sample. The distributions of these allele frequencies are skewed towards 0 with maxima

varying between samples (the extremes are 7 % and 20 % in samples N5 and N6, corresponding to the narrowest and widest band in Fig. 4). Over all samples, there is a correlation between fraction D (distinguishable Numts) and fraction A (without sequence similarity to mitochondrial genomes). The slope of the linear regression represents the genomic proportion of distinguishable Numts in *P. pedestris*. It is 9×10^{-04} ($p=1.04\times10^{-3}$) with a standard error of 1×10^{-4}. As expected, the intercept is not significantly different from zero ($p=0.805$), see Supplemental figure S6.

In total, there are four SNPs with frequencies that are clear outliers from the frequency distribution (shown by arrows in Fig. 4). Interestingly, individuals N1 and N3 from Le Blayeul share one such polymorphism at base pair 7567. These high-frequency variants are presumably the signatures of heteroplasmy (as seen by Mao et al., 2014 in bats).

Tangle plots: Atypical distances between read-pairs as a signature of Numts

Paired reads mapped to the mitochondrial assembly could have originated from mitochondrial genomes (M fraction) or from Numts. Those from the M fraction should show intra-pair distances consistent with the libraries' insert sizes, since the mitochondrial genome is highly conserved. Conversely, Numt sequences may have been subject to insertions, deletions, or rearrangements resulting in longer distances between map locations on the mitochondrial genome or discordant read orientations. The majority of intra-pair differences fell into a distribution with a maximum around 400 bp representing the sequencing insert size (supplemental data S3). There is a second (much shallower) peak

above 15,000 bp resulting from mapping reads generated from circular molecule to a linearised reference.

There is a small subset of read pairs with intermediate mapping distances that might be attributed to Numt sequences containing deletions or rearrangements. Fig. 5 shows each of these intermediate read pairs (those with an intra-pair distance between 1500 bp 14,508 bp) as a line connecting the paired reads' positions resulting in "tangle plots". Interestingly, some of the lines shown cluster together. Given the low sequencing coverage, this strongly suggests the presence of multiple copies of some Numts. Some patterns are shared across multiple samples, but the overall patterns are not population-specific (a linear discriminant analysis failed to assign all individuals to the correct populations, not shown here).

Mapping PacBio reads

PacBio circular consensus sequences (CCS) generated from DNA of three individuals (samples P1-P3) were mapped to the mitochondrial assembly. Out of 297,899 non-redundant reads generated in total, 443 showed similarity to the mitochondrial assembly with a cumulative mapping length of 396,770 bp. Of these, most reads matched the mitochondrial reference only along a part of their length, a pattern expected for short Numts and also chimeric PacBio read (which we expect to be rare). The alignments cover 263,635 bp, which corresponds to 0.027 % of the total CCS data generated. There were only 59 PacBio CCSs matching full-length, of which 41 were sufficiently diverged from the mitochondrial sequence to meet or criterion for classification as Numts. These align along

96,210 bp corresponding to 0.01 % of the (non-redundant) PacBio data generated. The remaining 18 full-length matches could be derived from mitochondrial genomes, but they may well be derived from Numts inserted recently. Covering 36,925 bp, these ambiguous CCSs represent only 9.3 % of the ML fraction in the PacBio data.

Interestingly, the mapping depth of full-length matches has a bimodal distribution. While the 18 ambiguous matches contribute mostly to the first peak, Numt-derived CCSs map to the areas under both peaks (see Fig. 6).

Discussion

Both mitochondrial (M) and Numt (C+D) sequence are generated as side-products of sequencing experiments, analgous to by-catch on fishing trawlers. We investigated these sequences using genome-skimming data (less than ⅓x genomic coverage) from the grasshopper, *Podisma pedestris,* using Illumina's NextSeq and PacBio's RSII platforms with six and three biological replicates, respectively

By-catch differs between sequencing methods

One of the most striking results is that the Illumina data had over 16-fold higher frequency of reads mapping to the mitochondrial clusters than the PacBio data, suggesting that the Illumina protocol produced a correspondingly higher proportion of sequences from the mitochondria (the M fraction), at some point between extraction and data interpretation. This bias cannot be explained by the known general over-representation of sequences with low-GC sequences in Illumina reads (Ekblom et al., 2014), as shown by the deviation of the

mitochondrial clusters from the general trend in Fig. 3. The result is reinforced by the comparable bias shown in the estimates of the proportion of mitochondrial sequences classified as Numts (fraction D/ML). This value is also much higher in the PacBio data than the Illumina. In the PacBio case, the two D categories sum to 91% of ML (the D estimate is obtained from the length of the mitochondrial portion of partially matching PacBio sequence plus the length of diverged full-length matches). In the Illumina data, the D estimates obtained from the frequencies in Figure 4 are much smaller, lying between 7% and 20% of ML.

This enrichment could be due to the greater retention of mitochondrial sequence (fraction M) in the preparation of the short-read libraries. Short-read libraries are usually fragmented and size selected to produce a distribution of fragments around 350-550 bp long, which will include fragments of the mitochondrial genome. On the other hand library preparation for long-read sequencing involves more careful shearing and a subsequent size selection for longer fragments (around 3-4 kbp in this case). This may cause mitochondrial genomes, starting at 16kbp before shearing, to be differentially lost from PacBio libraries while Numts, being part of the nuclear DNA, would be represented in their natural proportion.

Estimating the genomic proportion of Numts

Building on the results presented above, there are two ways of estimating the genomic proportion of Numts in genome skimming data, which are possible even in the absence of a genome assembly. It is shown in Supplemental Information S6 that for our short-read data, there is a good correlation (R^2=0.93) between the proportions of Numt-derived data reads (fraction D) and non-ML data (fraction A). The slope of this regression corresponds to the

374 estimated genomic proportion. It is 0.09 %. This is a lower-bound estimate, because it is
375 based on sequence divergence between Numts and the mitochondrial genomes sequence.
376 The estimate based on PacBio CCSs is somewhat lower; ML CCSs with divergent sequences
377 amount for 0.01% of the PacBio data. Another class of CCSs, which match only along part of
378 their sequence, are likely to represent Numts, too, however there is a small chance that
379 some of them are derived from chimeric SMRT bells. These sequences amount for 0.027% of
380 the PacBio data, giving a total of 0.037%.

381 **Genome size and Numt content**

382 *P. pedestris* has the largest genome of any insect listed in the Animal Genome Size Database
383 (Gregory, 2016, accessed 24 June 2019). Its C-value of 16.93 corresponds to 16.5 Gbp
384 (Doležel, Bartoš, Voglmayr, & Greilhuber, 2003). Consequently, a genomic proportion of just
385 under 0.1 % is equivalent to about a thousand full length mitochondrial genomes inserted
386 into the nuclear DNA (this length of sequence is equivalent to an entire *Drosophila*
387 *melanogaster* or *Arabidopsis thaliana* chromosome). Although this is a surprisingly large
388 number, as a proportion of the total genome it is consistent with estimates from other
389 species. Hazkani-Covo et al. (2010) present estimates of Numt contents for a diverse list of
390 85 species ranging from 0 to 0.25% in multicellular organisms. By contrast RepeatExplorer
391 analyses suggest repeats account for approximately 70 % of the *P. pedestris* genome,
392 including transposable elements able to excise and re-insert themselves, providing a
393 mechanism for copy-number increase.

Tangle plots

In Fig. 5, we show tangle plots, which allow visual comparisons between Numts in multiple samples. The repeated occurrence of the same links within a sample suggest that rearranged mitochondrial sequence has been replicated within a single genome (the low coverage of < ⅓x makes repeated sequencing of the same region unlikely). The similar patterns in different individuals and populations suggest that many of these replicated insertions are fixed or occur at a high frequency. Given that they are unlikely to be functional, it is most plausible that they have spread by genetic drift.

Tangle plots can be used to visualise any short-read data sets mapped to a circular (or tandem-repetitive) reference, see Supplemental Information S7.

Acknowledgements

This research utilised Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded by EPSRC grant EP/K000128/1. Further computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure. HB was funded by a PhD studentship of QMUL's School of Biological and Chemical Sciences. We wish to thank Graham Stone and Duncan Greig for their comments on an earlier version of the manuscript. We would also like to thank Alex Twyford for suggesting the name "tangle plot".

Literature Cited

[dataset]Becher, H., & Nichols, R. A.; 2019; Genome skimming data from the grasshopper Podiums pedestris; DataDryad;Persistent identifier (e.g. DOI)

Anmarkrud, J. A., & Lifjeld, J. T. (2017). Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Molecular Ecology Resources*, *17*(2), 334–341. doi:10.1111/1755-0998.12600

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., … Saunders, N. C. (1987). Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, *18*, 489–522.

Baca, M., Popović, D., Panagiotopoulou, H., Marciszak, A., Krajcarz, M., Krajcarz, M. T., … Nadachowski, A. (2018). Human-mediated dispersal of cats in the Neolithic Central Europe. *Heredity*, *121*(6), 557–563. doi:10.1038/s41437-018-0071-4

Ballard, J. W. O., & Rand, D. M. (2005). The Population Biology of Mitochondrial DNA and Its Phylogenetic Implications. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 621–642. doi:10.1146/annurev.ecolsys.36.091704.175513

Barbrook, A. C., Howe, C. J., & Purton, S. (2006). Why are plastid genomes retained in non-photosynthetic organisms? *Trends in Plant Science*, *11*(2), 101–8. doi:10.1016/j.tplants.2005.12.004

Bensasson, D., Zhang, D.-X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, *16*(6), 314–321. doi:10.1016/S0169-5347(01)02151-6

Bensasson, D., Zhang, D.-X., & Hewitt, G. M. (2000). Frequent Assimilation of Mitochondrial DNA by Grasshopper Nuclear Genomes. *Molecular Biology and Evolution*, *17*(3), 406–415. doi:10.1093/oxfordjournals.molbev.a026320

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., … Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319. doi:10.1016/j.ympev.2012.08.023

Bertheau, C., Schuler, H., Krumböck, S., Arthofer, W., & Stauffer, C. (2011). Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. *Molecular Ecology Resources*, *11*(6), 1056–1059. doi:10.1111/j.1755-0998.2011.03050.x

Blacket, M. J., Semeraro, L., & Malipatil, M. B. (2012). Barcoding Queensland Fruit Flies (Bactrocera tryoni): impediments and improvements. *Molecular Ecology Resources*, *12*(3), 428–436. doi:10.1111/j.1755-0998.2012.03124.x

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, *27*(8), 1767–1780. doi:10.1093/nar/27.8.1767

Brelsford, A., Mila, B., & Irwin, D. E. (2011). Hybrid origin of Audubon's warbler. *Molecular Ecology*, *20*(11), 2380–2389. doi:10.1111/j.1365-294X.2011.05055.x

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

doi:10.1093/bioinformatics/btp163

Dayama, G., Emery, S. B., Kidd, J. M., & Mills, R. E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research*, *42*(20), 12640–12649. doi:10.1093/nar/gku1038

Dierckxsens, N., Mardulyn, P., & Smits, G. (2016). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4), gkw955. doi:10.1093/nar/gkw955

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, *20*(9), 525–7. doi:10.1016/j.tplants.2015.06.012

Doležel, J., Bartoš, J., Voglmayr, H., & Greilhuber, J. (2003). Letter to the editor. *Cytometry*, *51A*(2), 127–128. doi:10.1002/cyto.a.10013

Du Buy, H. G., & Riley, F. L. (1967). Hybridization between the nuclear and kinetoplast DNA's of Leishmania enriettii and between nuclear and mitochondrial DNA's of mouse liver. *Proceedings of the National Academy of Sciences*, *57*(3), 790–797.

Ekblom, R., Smeds, L., & Ellegren, H. (2014). Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics*, *15*(1), 467. doi:10.1186/1471-2164-15-467

Erler, S., Ferenz, H.-J., Moritz, R. F. A., & Kaatz, H.-H. (2010). Analysis of the mitochondrial genome of *Schistocerca gregaria gregaria* (Orthoptera: Acrididae). *Biological Journal of the Linnean Society*, *99*, 296–305. doi:10.1111/j.1095-8312.2009.01365.x

Fernández-Vizarra, E., Enríquez, J. A., Pérez-Martos, A., Montoya, J., & Fernández-Silva, P. (2011). Tissue-specific differences in mitochondrial activity and biogenesis. *Mitochondrion*, *11*(1), 207–213. doi:10.1016/j.mito.2010.09.011

Garber, R. C., & Yoder, O. C. (1983). Isolation of DNA from filamentous fungi and separation into nuclear, mitochondrial, ribosomal, and plasmid components. *Analytical Biochemistry*, *135*(2), 416–422. doi:10.1016/0003-2697(83)90704-2

Gellissen, G., Bradfield, J. Y., White, B. N., & Wyatt, G. R. (1983). Mitochondrial DNA sequences in the nuclear genome of a locust. *Nature*, *301*(5901), 631–634. doi:10.1038/301631a0

Gregory, T. (2016). Animal Genome Size Database. Retrieved from http://www.genomesize.com

Hawlitschek, O., Morinière, J., Lehmann, G. U. C., Lehmann, A. W., Kropf, M., Dunz, A., … Haszprunar, G. (2017). DNA barcoding of crickets, katydids and grasshoppers (Orthoptera) from Central Europe with focus on Austria, Germany and Switzerland. *Molecular Ecology Resources*, *17*(5), 1037–1053. doi:10.1111/1755-0998.12638

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLoS Genetics*, *6*(2), e1000834. doi:10.1371/journal.pgen.1000834

Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003). Barcoding Animal Life: Cytochrome c Oxidase Subunit 1 Divergences among Closely Related Species. *Proceedings: Biological Sciences*, *270*, S96–S99.

Jiang, H., Lei, R., Ding, S.-W., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, *15*(1), 182.

doi:10.1186/1471-2105-15-182
Jordal, B. H., & Kambestad, M. (2014). DNA barcoding of bark and ambrosia beetles reveals excessive NUMTs and consistent east-west divergence across Palearctic forests. *Molecular Ecology Resources*, *14*(1), 7–17. doi:10.1111/1755-0998.12150
Kielbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. doi:10.1101/gr.113985.110
Kim, S.-J., Lee, K. Y., & Ju, S.-J. (2013). Nuclear mitochondrial pseudogenes in *Austinograea alayseae* hydrothermal vent crabs (Crustacea: Bythograeidae): effects on DNA barcoding. *Molecular Ecology Resources*, *13*(5), 781–787. doi:10.1111/1755-0998.12119
Lansman, R. A., Shade, R. O., Shapira, J. F., & Avise, J. C. (1981). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. *Journal of Molecular Evolution*, *17*(4), 214–226.
Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi:10.1093/bioinformatics/btp324
Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., & O'Brien, S. J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, *39*(2), 174–190. doi:10.1007/BF00163806
Mao, X., Dong, J., Hua, P., He, G., Zhang, S., & Rossiter, S. J. (2014). Heteroplasmy and Ancient Translocation of Mitochondrial DNA to the Nucleus in the Chinese Horseshoe Bat (Rhinolophus sinicus) Complex. *PLoS ONE*, *9*(5), e98035. doi:10.1371/journal.pone.0098035
Miraldo, A., Hewitt, G. M., Dear, P. H., Paulo, O. S., & Emerson, B. C. (2012). Numts help to reconstruct the demographic history of the ocellated lizard (Lacerta lepida) in a secondary contact zone. *Molecular Ecology*, *21*(4), 1005–1018. doi:10.1111/j.1365-294X.2011.05422.x
Mohandesan, E., Speller, C. F., Peters, J., Uerpmann, H.-P., Uerpmann, M., De Cupere, B., … Burger, P. A. (2017). Combined hybridization capture and shotgun sequencing for ancient DNA analysis of extinct wild and domestic dromedary camel. *Molecular Ecology Resources*, *17*(2), 300–313. doi:10.1111/1755-0998.12551
Novák, P., Neumann, P., Pech, J., Steinhaisl, J., & Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, *29*(6), 792–793. doi:10.1093/bioinformatics/btt054
R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, *99*(2), 349–64. doi:10.3732/ajb.1100335
Thalmann, O., Hebler, J., Poinar, H. N., Pääbo, S., & Vigilant, L. (2004). Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. *Molecular Ecology*, *13*(2), 321–335. doi:10.1046/j.1365-294X.2003.02070.x
Thalmann, O., Serre, D., Hofreiter, M., Lukas, D., Eriksson, J., & Vigilant, L. (2004). Nuclear

insertions help and hinder inference of the evolutionary history of gorilla mtDNA. *Molecular Ecology*, *14*(1), 179–188. doi:10.1111/j.1365-294X.2004.02382.x

Twyford, A. D., & Ness, R. W. (2016). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12626

van der Valk, T., Lona Durazo, F., Dalén, L., & Guschanski, K. (2017). Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12699

Zhang, D.-X., & Hewitt, G. M. (1997). Insect mitochondrial control region: A review of its structure, evolution and usefulness in evolutionary studies. *Biochemical Systematics and Ecology*, *25*(2), 99–120. doi:10.1016/S0305-1978(96)00042-7

Data Accessibility

Mito sequences aligned – Supplemental data S1

Neighbour-joining tree – Supplemental data S2

Distances of paired reads mapped – Supplemental data S3

RE contigs – Supplemental data S4

matching PacBio reads, SAM format, GZIP-compressed TAR ball – Supplemental data S5

Regression plot – Supplemental data S6

Information on tangle plots – Supplemental data S7 (see also GitHub:

https://github.com/SBCSnicholsLab/tangles)

The original sequencing data will be deposited on Dryad.

Author Contributions

HB and RAN conceived the experiment and collected samples. HB carried out the experiment, analysed the data, and wrote the manuscript. HB and RAN revised the manuscript.

569

570

571

572 # Tables

573

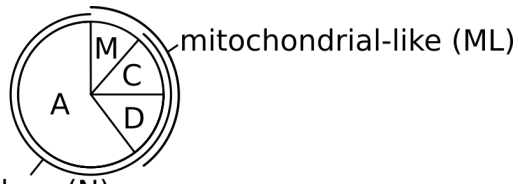
574 Table 1. Overview of the sequencing data generated and fraction ML (mitochondrial-like,

575 mapping to the mitochondrial genome assembly).

NextSeq	Sample	Origin	High-quality reads	ML (% of reads)	Max SNP allele frequency
	N1	Blayeul	25600134	0.48%	15%
	N2	Blayeul	37563526	0.64%	14%
	N3	Blayeul	13034544	1.17%	8%
	N4	Mariaud	35257122	0.97%	9%
	N5	Mariaud	33393734	1.32%	7%
	N6	Mariaud	21813470	0.44%	20%
PacBio	**Sample**	**Origin**	**Reads (bp in CCS)**		**ML reads (bp)**
	PB1	Blayeul	92971 (336152778 bp)		160 (0.04%)
	PB2	Blayeul	97937 (376785844 bp)		131 (0.03%)
	PB3	Bournee	106991 (248262810 bp)		152 (0.06%)

576

577 # Figures



578

579 Figure 1. Fractionation of sequencing data. Of all the sequencing data generated from one
580 sample, some fraction N will originate from the nuclear DNA and some fraction M from the
581 mitochondrial DNA, and other sequences. The nuclear fraction N may contain Numts. Some
582 of these will have sequences divergent from M and hence are easily identifiable (fraction D).
583 Other Numts may be cryptic, e.g. recent insertions, (fraction C). Together, Numts and
584 mitochondrial sequences for the fraction ML “mitochondrial-like”.
585

586

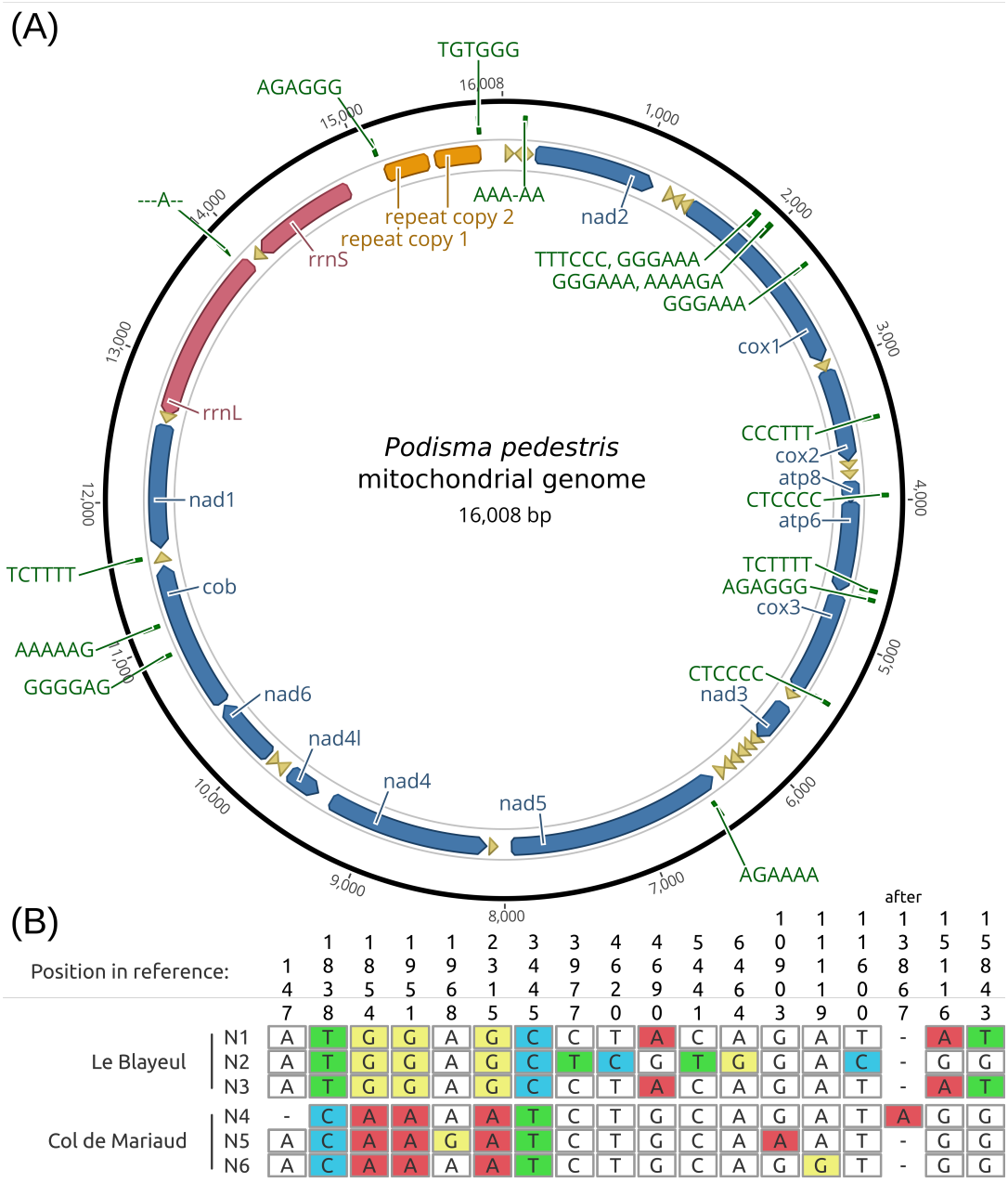


Figure 2. Annotated mitochondrial genome assembly of the alpine grasshopper, *Podisma pedestris*, and between-individual polymorphisms. (A) Shows a representation of the mitochondrial genome assembly. (B) The alignment of the six assemblies contains 18 polymorphic sites. Their positions are indicated above the alignment and in (A). Mismatches are highlighted.

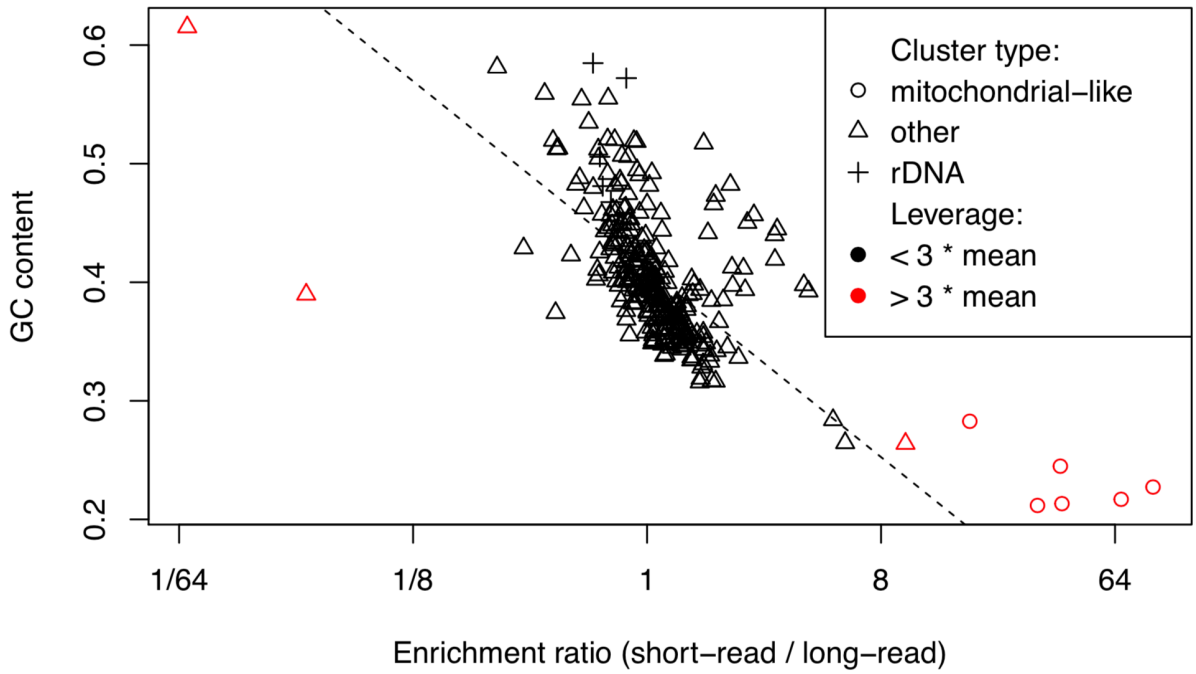


Figure 3. Enrichment for short or long reads plotted against GC-content for 300 RepeatExplorer clusters. Symbols right of the centre indicate high-abundance DNAs enriched in the short-read data. There is a general trend for enrichment in PacBio data for clusters with higher GC-content (regression line). Mitochondrial-like clusters, which show the strongest bias, are indicated as circles, rDNA as crosses. Data points with a leverage greater than three times the mean leverage are shown in red. These were excluded from the final regression (dashed line).

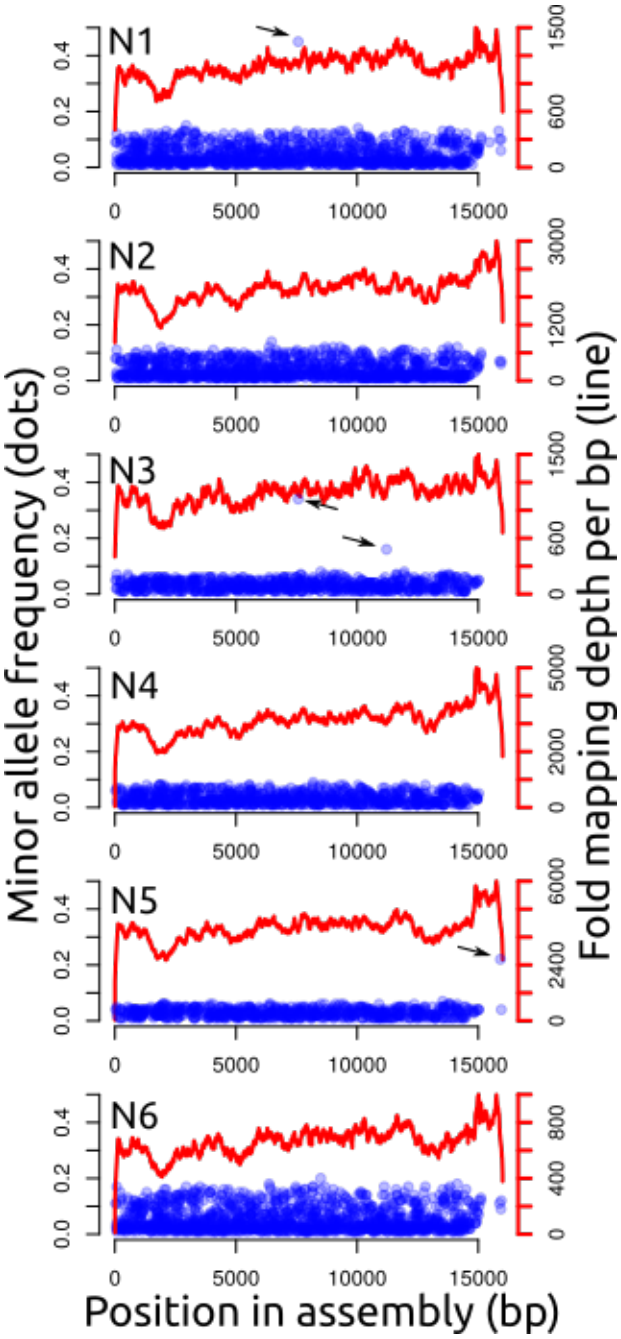
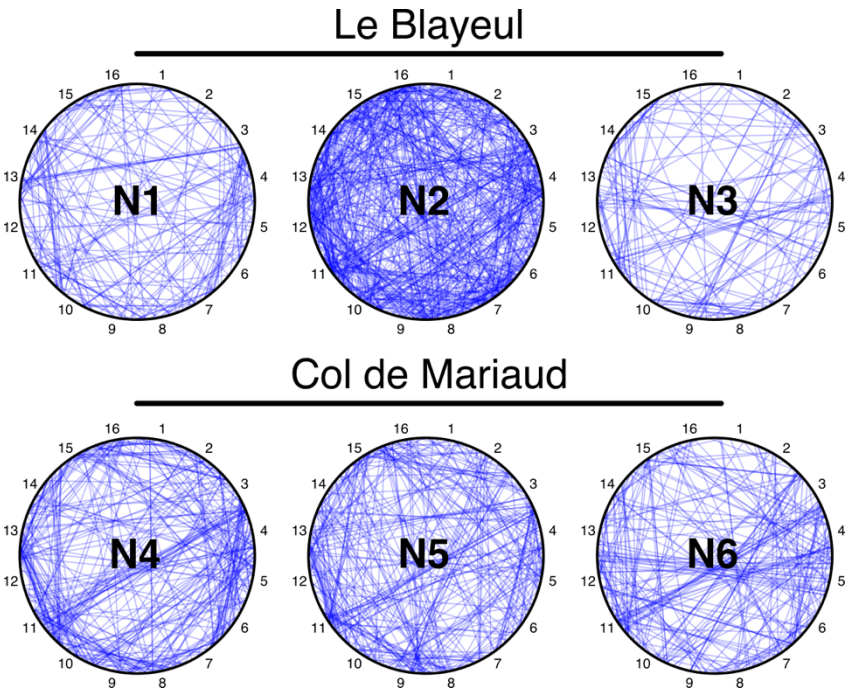


Figure 4. Polymorphisms within alignments. For each individual, the minor allele frequencies and positions of SNPs are shown as dots. Note, the dots generally form bands with widths differing between samples. There are few outliers, marked with arrows, likely indicating heteroplasmy. Individuals N1 and N3 share such a polymorphism at base pair 7567. The lines indicate per-base pair mapping depths.

607

608



609

610 Figure 5. Tangle plots. The position of read pairs mapped with unusual read distances is
611 indicated by a line connecting the read locations. For ease of comparison, 16 positions are
612 labelled on each graph. Note common patterns between individuals. For instance, all
613 individuals in the Col de Mariaud population show several read pairs connecting positions
614 3/4 and 11. All samples of both populations show read pairs connecting segments 3/4 and 7.
615

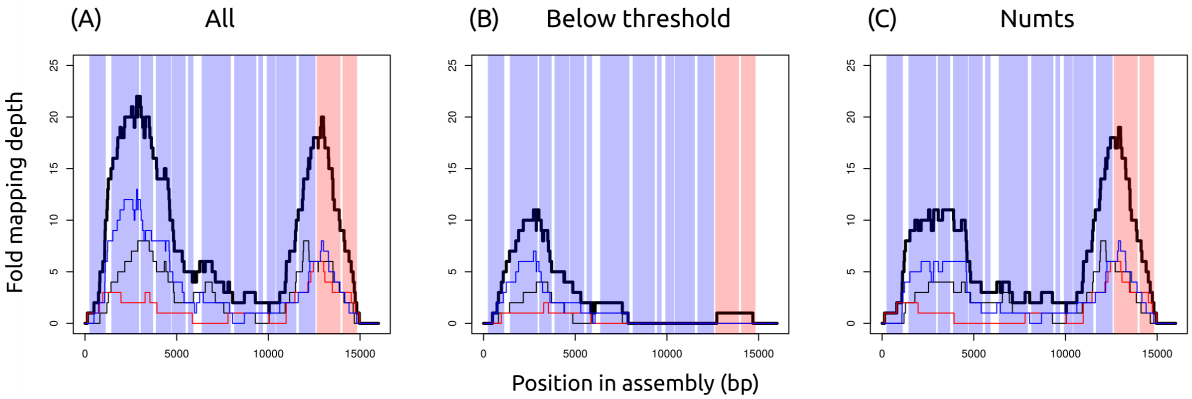


Figure 6. Mapping depths of PacBio CCSs mapping to mitochondria for their full-length. The bold line indicates the total over all three samples, narrow lines represent individual samples. Vertical shaded bands indicate the positions of protein-coding genes and mitochondrial rRNA genes (the two salmon coloured bands on the far right). (A) All CCSs matching full-length. (B) Fraction M or C (either mitochondrial or cryptic Numt) (C) Fraction D (classified as a Numt).