



HAL
open science

Untangling Heteroplasmy, Structure, and Evolution of an Atypical Mitochondrial Genome by PacBio Sequencing.

Jean Peccoud, Mohamed Amine Chebbi, Alexandre Cormier, Bouziane Moumen, Clément Gilbert, Isabelle Marcadé, Christopher Chandler, Richard Cordaux

► To cite this version:

Jean Peccoud, Mohamed Amine Chebbi, Alexandre Cormier, Bouziane Moumen, Clément Gilbert, et al.. Untangling Heteroplasmy, Structure, and Evolution of an Atypical Mitochondrial Genome by PacBio Sequencing.. *Genetics*, 2017, 207 (1), pp.269-280. 10.1534/genetics.117.203380 . hal-01586985

HAL Id: hal-01586985

<https://hal.science/hal-01586985v1>

Submitted on 9 Sep 2024

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.

Public Domain

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

**Heteroplasmy, structure and evolution of an atypical mitochondrial genome
untangled by PacBio sequencing**

Jean Peccoud*, Mohamed Amine Chebbi*, Alexandre Cormier*, Bouziane Moumen*, Clément
Gilbert*, Isabelle Marcadé*, Christopher Chandler†, Richard Cordaux*

*UMR CNRS 7267 Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose,
Université de Poitiers, 96000 Poitiers, France

†Department of Biological Sciences, State University of New York at Oswego, Oswego, NY 13126

Sequence data are available on Genbank (accession numbers to be provided after revision)

19 Short title: PacBio and atypical mitogenomes

20

21 Corresponding author: Jean Peccoud; jean.peccoud@univ-poitiers.fr

22 Laboratoire Ecologie et Biologie des Interactions (EBI) - UMR CNRS 7267

23 Bâtiment B8-B35, 5 rue Albert Turpain TSA 51106 F-86073 POITIERS Cedex 9 tel : +33 (0)5 49 45

24 35 60

25

26

27 Keywords: mtDNA, concerted evolution, crustacean isopods, telomeres, third-generation sequencing

28

29 Abstract

30 The highly compact mitochondrial genome of terrestrial isopods (Oniscidae) presents two unusual
31 features. First, several loci can individually encode two tRNAs, thanks to single-nucleotide
32 polymorphisms at anticodon sites. Within-individual variation (heteroplasmy) at these loci is thought
33 to have been maintained for millions of years because individuals that do not carry all tRNA genes
34 die, resulting in strong balancing selection. Second, the oniscid mtDNA genome comes in two
35 conformations: a ~14kb linear monomer and a ~28kb circular dimer comprising two monomer units
36 fused in palindrome. We hypothesized that heteroplasmy actually results from two genome units of the
37 same dimeric molecule carrying different tRNA genes at mirrored loci. This hypothesis however
38 contradicts the earlier proposition that dimeric molecules result from the replication of linear
39 monomers, a process that should yield totally identical genome units within a dimer. To solve this
40 contradiction, we used the SMRT (PacBio) technology to sequence mirrored tRNA loci in single
41 dimeric molecules. We show that dimers do present different tRNA genes at mirrored loci, hence that
42 covalent linkage, rather than balancing selection, maintains vital variation at anticodons. We also
43 leveraged unique features of the SMRT technology to detect linear monomers closed by hairpins and
44 carrying non-complementary bases at anticodons. These molecules contain the necessary information
45 to encode two tRNAs at the same locus and suggest new mechanisms of transition between linear and
46 circular mtDNA. Overall, our analyses clarify the evolution of an atypical mitochondrial genome,
47 which relies on recombination and where dimerization counterintuitively enabled further mtDNA
48 compaction.

49 Introduction

50 Mitochondria originated from an alphaproteobacterial ancestor which lost most of its genes during its
51 eukaryotic evolution. As a result, animal mitochondrial (mt) DNA is classically considered as a
52 remarkably compact and uniform molecule. Indeed, the typical bilaterian mt genome is a single
53 circular molecule ranging from 15-20 kb in length, which contains 37 genes, including 13 protein-
54 coding genes, two rRNA genes and 22 tRNA genes (Boore 1999). While the majority of bilaterian mt
55 genomes conform to this description, several notable exceptions have been uncovered. Unusual
56 bilaterian mt genomes include multipartite [e.g., Suga *et al.* (2008); Dickey *et al.* (2015)] and linear
57 (Raimond *et al.* 1999) structures, atypical size [e.g., Helfenbein *et al.* (2004); Liu *et al.* (2013)],
58 changes in gene content [e.g., Okimoto *et al.* (1992); Helfenbein *et al.* (2004)], plasticity in gene order
59 [e.g., Singh *et al.* (2009); Gissi *et al.* (2010)] and additional genetic codes [e.g., Watanabe and
60 Yokobori (2011); Abascal *et al.* (2012)]. Because they deviate from the standard model, these mt
61 genomes may constitute ideal systems to further our understanding of mitochondrial biology and
62 evolution in animals, as they can help to address questions of recombination, concerted evolution of
63 mitochondrial loci and non-standard inheritance.

64 The mt genome of terrestrial isopods (Isopoda: Oniscidea) is one of such atypical genomes. It is
65 notable for its compaction. In particular, genes coding transfer RNAs (tRNAs) can partially or fully
66 overlap with protein coding genes (Doublet *et al.* 2015). But one truly unique feature of this genome is
67 the capacity of three tRNA loci to each encode two alternative tRNAs with distinct anticodons, thanks
68 to single nucleotide polymorphisms (SNPs) occurring within the same individual (Marcade *et al.*
69 2007; Doublet *et al.* 2008; Chandler *et al.* 2015). At all three loci, mtDNA shows two different bases
70 at one position of the anticodon, thus making individuals heteroplasmic at these nucleotide positions.
71 This variation appears as a double peak on chromatograms generated by direct Sanger sequencing of
72 PCR amplicons (Marcade *et al.* 2007; Doublet *et al.* 2008; Chandler *et al.* 2015), cut and uncut
73 amplicons on electrophoresis gels after mtDNA digestion by appropriate enzymes (Doublet *et al.*
74 2008) or SNPs among sequences obtained from next-generation technologies (Chandler *et al.* 2015).
75 The same three heteroplasmic anticodon sites have been detected in individuals of two oniscid species,
76 *Trachelipus rathkei* and *Cylisticus convexus* (Chandler *et al.* 2015), each site allowing the encoding of
77 two tRNAs per locus and saving one dedicated tRNA locus. One of these heteroplasmic sites is shared
78 with *Armadillidium vulgare* (Marcadé *et al.* 2007) and a diverse array of terrestrial isopod species
79 (Doublet *et al.* 2008). The presence of these heteroplasmic sites in divergent oniscid lineages suggests
80 that at least some of them have been maintained for millions of generations (Doublet *et al.* 2008).
81 Bottlenecks resulting from the transmission of relatively few organelles to zygotes usually remove
82 heteroplasmy in few generations (Breton and Stewart 2015; Stewart and Chinnery 2015). In these
83 oniscids, however, it is believed that “constitutive” heteroplasmy is maintained by the requirement of
84 all tRNA variants within an animal and possibly even within an individual mitochondrion. This
85 represents the only suspected case of balancing selection acting on organelles (Doublet *et al.* 2008).
86 The hypothesis of constitutive heteroplasmy maintained by balancing selection must however consider
87 another unique feature of the mt genome of terrestrial isopods. This genome is remarkable for
88 presenting two conformations: one linear monomer of ~14kb that represents one unit of mt genome
89 containing the standard bilaterian mt genes, and a circular ~28-kb dimer that is a palindrome
90 composed of two monomers, each representing one genome unit, arranged in a mirrored fashion
91 (Raimond *et al.* 1999; Marcade *et al.* 2007). The presence of dimers, which constitute about half of the
92 mtDNA molecules in *A. vulgare* (Raimond *et al.* 1999), leaves the possibility that both tRNAs of a
93 heteroplasmic site can be encoded by the two genome units of a dimeric molecule, such that a single
94 dimer may encode all tRNAs. The transmission of such dimers would allow faithful inheritance of all
95 vital tRNAs genes to daughter mitochondria and to the progeny, and would ensure good balance of the
96 tRNAs within organelles. This hypothesis implies that the two genome units within a dimer are not
97 completely identical. It therefore conflicts with another formulated hypothesis: that dimers arise from
98 the replication of linear monomers. The extremities of linear monomers contain inverted terminal
99 repeats that are thought to be telomeric hairpins covalently linking the two DNA strands (Doublet *et*
100 *al.* 2013). DNA polymerase would be able to navigate and then replicate the other strand, circularizing

101 the linear monomer into a dimer in the process (Figure 1A). If so, this dimer would be expected to
102 present totally identical genome units.
103 Therefore, the nature of heteroplasmy, the utility of dimeric mtDNA molecules and the possible
104 conversions between the unusual conformations of the oniscid mt genome are entangled issues that
105 must be investigated together. To take on this task, we used long reads generated by the Single-
106 Molecule Real Time (SMRT) sequencing technology from Pacific Biosciences. These reads allowed
107 us to reconstruct the haplotypes, hence the combination of tRNAs encoded by individual dimeric
108 molecules in four oniscid lineages. We specifically investigated whether mt molecules can encode all
109 required tRNAs. In addition to long reads, we took advantage of unique features of the SMRT
110 sequencing technology to identify the conformation of molecules and clarify the conversions between
111 dimeric and monomeric forms of this atypical mitochondrial genome.

112 **Materials and Methods**

113 We examined four terrestrial isopod matriline (table 1): two from *A. vulgare* (named BF and WXf),
114 one from *A. nasatum* and one from *Trachelipus rathkei*. For each matriline, short sequencing reads
115 (Illumina) and long reads (SMRT) have been obtained from the genomic DNA of one or several
116 related individuals (full siblings or first cousins) as part of full-genome assembly projects.

117 **Generation of mitochondrial genome sequences**

118 We aimed at building the dimeric mt genome sequence of each lineage with both units placed head to
119 head. This configuration was chosen to facilitate the use of long reads spanning mirrored anticodon
120 sites, which are much closer to the head-to-head junction than they are to the tail-to-tail junction
121 (Figure 1).

122 The mt genomes of *A. vulgare* BF and *A. nasatum* were reconstructed from contigs generated for other
123 full-genome assembly projects. For each of these lineages, we first retrieved contigs comprising
124 mitochondrial sequences by performing a blastn (Camacho *et al.* 2009) homology search against the
125 nearly complete mitochondrial genome of *A. vulgare* (Genbank accession number EF643519.3). These
126 searches returned several contigs having large portions of the mitochondrial genome sequence in a
127 head-to-head configuration, as expected if these contigs comprised the dimeric form. The contig
128 encompassing the longest homology in such configuration, and the lowest divergence with the
129 reference genome, was selected and trimmed if needed.

130 For consensus sequence polishing, Illumina reads were aligned to the retained contig using Bowtie2
131 version 2.2.9 (Langmead and Salzberg 2012), which was set to the default low sensitivity (“fast”
132 search strategy) and configured to retain only alignments including both reads of a pair. The alignment
133 file was processed with Pilon version 1.18 (Walker *et al.* 2014) to correct potential errors in the
134 mapped reference contig. These two steps were repeated, and the alignment file was inspected for
135 remaining errors with Integrated Genome Viewer version 2.3.92 (Robinson *et al.* 2011). Any error in

136 the consensus sequence was manually corrected with Geneious Pro version 5.4 (Drummond *et al.*
137 2010).

138 For *A. vulgare* WXf, we used the BF mt genome as a reference. We corrected differences with WXf
139 using the same mapping strategy as described above. For *T. rathkei*, we used the reference genome
140 available on Genbank (accession number KR013001.1). As this genome contains a complete unit
141 flanked by short palindromic parts representing the ends of the other genome unit, we generated the
142 expected dimeric form and used it as a reference. We corrected potential errors and differences with
143 our studied lineage as described for the other genomes.

144 From the mapping file generated above, alignments of reads originating from the same DNA fragment
145 were removed with SAMtools rmdup version 1.3.1 (Li *et al.* 2009). Bases at each position were called
146 by SAMtools mpileup. A custom C program was used to convert the pileup file into base counts,
147 discarding bases with quality score lower than 25. Sites where the rarer base was carried by more than
148 20% of the reads, were considered possibly heteroplasmic. At such site, the reference sequence of
149 each lineage was modified to show ambiguities following IUPAC conventions. This was necessary to
150 avoid any bias in the alignment of long reads, which minimizes mismatches at the risk of creating
151 spurious indels.

152 Alignment of long reads

153 We aligned long reads on the corresponding reference genome of each lineage using BLASR version
154 1.3.1 (Chaisson and Tesler 2012) with default settings. Long reads consisted in reads of inserts (Figure
155 2A) and circular consensus (“CCS”), both being generated by the sequencing centers. A CCS is the
156 consensus among reads from the same polymerase read (DNA fragment), and may not be called if not
157 enough reads are present.

158 For subsequent analyses, it was crucial to ascertain the alignment orientation (sequenced DNA strand)
159 of each read mapping across the palindromic genome units. This orientation could be determined as
160 we found that genome units were separated by a short non-symmetrical junction (see Figure 1 and
161 results). We thus retained up to two alignments per read, one per orientation, which we compared to
162 determine the most likely sequenced strand of the junction (see below). We did not simply retain the
163 alignment with best overall score, as sequencing errors may prevent reliable inference of the most
164 likely mapping orientation.

165 All following steps were executed in R 3.3 (R Core Team 2014), with the help of functions from
166 packages GenomicAlignments (Lawrence *et al.* 2013) and Biostrings (Pages *et al.* 2015). Our script
167 was based on the splitting of the BLASR alignment (.bam) file into a matrix of individual bases, in
168 which columns correspond to successive positions of the reference sequence and rows to aligned
169 reads. For all positions of the junction, we counted the frequency of mismatches (including deletions)
170 between the sequence of the reference and that of a read, in each alignment orientation. We did not
171 count insertions in the read as these could not be retained in the matrix. If mismatch frequencies

172 between mapping orientations of a read differed by more than 10%, and if the lowest mismatch
173 frequency did not exceed 25%, we considered the mapping orientation corresponding to that
174 frequency as the correct one. Otherwise, the alignment orientation of a read on the junction was
175 considered undetermined.

176 For each read, we retained the alignment (row of the base matrix) corresponding to the inferred
177 mapping orientation on the junction. If the more likely orientation could not be inferred, we retained
178 the alignment with best mapping score, or selected alignments at random if scores were identical
179 (which was the case for reads not covering the junction).

180 **Determination of mtDNA molecule conformations**

181 To establish the combinations of tRNAs encoded by single dimeric molecules, we used SMRT reads
182 covering mirrored anticodon sites. Counterintuitively, these reads may not all come from mtDNA
183 dimers, due to specificities of the SMRT technology (Figure 2A). Indeed, a linear monomer might be
184 sequenced by the polymerase navigating its natural hairpin telomere just like it would navigate the
185 SMRT bell adapter (Figure 2B). Each resulting read would unite sequences from both DNA strands
186 and cover both genomic units, just like a read sequenced from a dimer across the junction (Figure 2C).
187 We took this possibility as an opportunity to detect linear monomers. We reasoned that a fragment
188 sequenced with two SMRT bells yields reads that alternatively map on opposite orientations, since
189 they come from alternate strands (Figure 2A). This should apply to mtDNA dimers (Figure 2C), but
190 not to linear monomers sequenced with just one SMRT bell, as each read would unite sequences from
191 both strands. Reads from a linear monomer map on the same orientation on the reference (Figure 2B).
192 We thus used the mapping orientation of reads on the junction between genome units to assign
193 fragments as monomers or dimers. The read number (whether it was the first to be sequenced, the
194 second, and so on) was inferred from its coordinates in the parent polymerase read (relative to the
195 other reads) which are part of its name. If solely the first read of a polymerase read spanned the
196 junction, we inferred the most likely mapping orientation of the second read, if present, based on the
197 expectation that the start of a read should map closely to the edge of the region mapped by the
198 previous read (Figure S1).

199 For a fragment to be classified as linear monomer, we further required that the middle of the region
200 mapped by at least one read was no further than 30 bp from the center of the junction, since the hairpin
201 telomere is sequenced at the middle of the read (Figure 2B). The mapping position of the middle of
202 reads can only be informative for “complete” reads, i.e., those starting and ending at a SMRT bell. We
203 defined a complete read as one having its start and end coordinates in the original polymerase read
204 fewer than 70 bp away from those of the previous and next reads, respectively. If the read was the first
205 of the polymerase read, we imposed that its start coordinate in the polymerase read was at most 70 and
206 that at least one of its first 50 bases aligned on the reference genome. CCS were considered as
207 complete reads. The position of the middle of a complete read on the reference genome was

208 designated as the midpoint between start and end positions of its alignment. These were adjusted by
209 adding or subtracting, as appropriate, the lengths of the unaligned “clipped” read parts (which are
210 often zero). These were obtained from the CIGAR of the alignment file.

211 We considered as dimeric fragments those yielding reads aligning to the junction in alternating
212 orientation, regardless of the alignment coordinates. Some fragments spanning the junction but
213 comprising only one reads that could be oriented with certainty were also classified as dimers based
214 on mapping coordinates only. We reasoned that during the sequencing of a linear monomer, the
215 polymerase, after going through the hairpin telomere, either returns to the SMRT bell or ends its
216 polymerization. Either event terminates the read at a position that cannot be further from the center of
217 the telomere than where the start of the read is, assuming that the read starts at the SMRT bell. To
218 classify a fragment as dimer, we thus imposed that the end of the read mapped at a distance from the
219 junction that is at least 100 bp longer than the distance between the junction and the mapping position
220 of the read start. This requirement must be fulfilled by actual mapping positions and by those
221 considering clipped read parts. To exclude first reads (of a polymerase read) that may not start at the
222 SMRT bell, we imposed that such read started at coordinate lower than 70 bp in its parent polymerase
223 read and that its left clipped part was shorter than 50 bp.

224 Establishment of haplotypes carried by mtDNA dimers

225 We then establishing haplotypes, hence tRNA genes carried by mtDNA molecules assigned to dimers.
226 Haplotypes were established by concatenating the bases at heteroplasmic sites in the matrix we
227 generated. Prior to that, we slightly corrected alignments as we found frequent one-bp deletions in
228 reads at these sites, associated with mismatches at the immediate flanking positions. Most mismatched
229 bases corresponded to one of the two possible bases carried by short reads at a heteroplasmic position.
230 We thus swapped the deletion and the mismatched base in the base matrix, which reduced mismatches
231 without altering the original read sequence. We believe that BLASR did not generate the best
232 alignment due to improper management of ambiguities in the reference sequence.

233 All reads of a parent polymerase read may not support the same haplotype, due to sequencing errors
234 that are frequent in SMRT data. In such cases, we selected the haplotype according to four successive
235 criteria: (i) higher number of sites having the bases supported by Illumina data, (ii) presence in the
236 CCS, (iii) higher frequency of the haplotype among reads of the polymerase read, and (iv) fewer
237 mismatches with the most frequent haplotype found across all reads.

238 Identification of fragments with non-complementary bases

239 If a dimer resulted from the replication of a linear monomer, it should present identical genome units
240 and encode a single tRNA type per pair of mirrored loci. As haplotypes clearly contradicted this
241 prediction (see results), we reasoned that the bases forming the two DNA strands of a linear monomer
242 converted into a dimer may not be complementary at the anticodon sites (Figure 1B).

243 To assess base complementarity within linear monomers, whose sequencing reads unite both strands
244 of a molecule (Figure 2B), we compared bases between mirrored heteroplasmic sites covered by the
245 same read. We also looked for base complementary in fragments that were sequenced with two SMRT
246 bells, by comparing bases carried by reads mapping on different mtDNA strands. These fragments
247 include those classified as dimers, and those that do not span the junction between genome units,
248 hence which could not be classified (hereafter called “unclassified” fragments). Unclassified
249 fragments were defined as molecules whose reads all aligned at least 100bp away from the junction
250 between genome units. Alignment positions considered clipped read parts.

251 Rather than a binary value, we developed an index to quantify the complementarity of bases between
252 DNA strands of a molecule (polymerase read) at a given position, as each strand may be sequenced
253 several times. This index ignores all reads carrying rare bases or deletions at this position. We define
254 Bf and Br as the most frequent bases among forward-aligned reads and reversed-aligned reads,
255 respectively. If the two possible bases have equal counts among reads of a given orientation, the most
256 frequent one is chosen at random. We let f be the fraction of forward-aligned reads carrying Bf and r
257 be the fraction of reverse-aligned reads carrying Br . We define our index of complementarity as:

$$258 \quad i = \begin{cases} \frac{f+r}{2}, & \text{if } Br = Bf \\ -\frac{f+r}{2}, & \text{otherwise.} \end{cases}$$

259 Index i varies from -1, if all bases between reads mapped in opposite orientation are non-
260 complementary, to 1 if all are complementary. Intermediate values represent conflicting results
261 between reads mapping in the same orientation. We defined a per-fragment index I that averages i
262 over sites covered by the fragment. To minimize the influence of sequencing errors, values of I that
263 were not obtained from at least two bases per strand were ignored. These two bases may either be
264 sequenced at the same site in two reads from the same strand, or sequenced at two sites in the same
265 read. We considered that a fragment carried non-complementary bases or complementary bases if I
266 was lower than -0.9 or higher than 0.9, respectively. Fragments whose indices fell between these
267 values were not considered. We also compared base complementarity across sites of the same DNA
268 fragment. This analysis relied on the per-site index i and ignored any site whose index differed from -1
269 or 1, or was calculated from less than two reads per strand.

270 [Investigation of recent recombination between mtDNA haplotypes](#)

271 Within-lineage variation between almost identical mtDNA molecules or genome units offered the rare
272 opportunity to assess whether haplotypes could be produced by recombination. To do this, we
273 established the frequencies of two-site haplotypes in reads of insert. We did so to take advantage of
274 fragments from linear monomers, which may not show consistent haplotypes across reads due to non-
275 complementary bases (see results). Estimates of haplotype frequencies considered that a pair of
276 heteroplasmic sites within a unit had two dominant haplotypes corresponding to both genome units,

277 for instance *AG* and *GC*. Recombination can produce haplotypes *AC* and *GG*. Each of these can also
278 result from a sequencing error (or mutation) on a single site of either dominant haplotype. Such
279 haplotype is therefore more likely to emerge by a single sequencing error than a haplotype like *AT*,
280 which can only derive from *AG*. To take that difference into account, we estimated the frequency of
281 haplotype *AT* as

$$\frac{N_{AT}}{\frac{1}{2} \sum_{XX} N_{XX} + \sum_{YY} N_{YY}}$$

283 where N_{AT} is the number of reads carrying haplotype *AT*. *XX* represents any haplotype that can derive
284 from either dominant haplotype with equal probability (hence which has a probability of 0.5 to derive
285 from *AG*) and *YY* is any haplotype that shares at least one base with *AG* (including haplotypes *AT* and
286 *AG*), hence which is considered to derive from it. The frequency of a haplotype that does not share
287 exactly one base with just one of the dominant haplotypes was estimated by dividing the number of
288 reads carrying this haplotype by the total number of reads supporting any haplotype.

289 Data availability

290 Annotated mitochondrial genome sequences are available at Genbank under accession numbers xxx.
291 Sequencing reads that mapped on mitochondrial genomes are available at the NCBI short read archive
292 under accession numbers xxx. (Accession numbers will be provided at revision stage.)
293 File S1 contains the supplementary text and figures S1-S3.

294 Results

295 Mitochondrial genome sequences and polymorphic sites

296 Dimeric genome sequences of all four lineages were successfully reconstructed, including junctions
297 between the heads of genome units. These junctions are 34 to 42-bp long in *Armadillidium* lineages
298 (Figure 3) and their sequences correspond to the “inverted repeats” that have been located near the
299 12S rRNA gene of mtDNA monomers in *A. vulgare* (Doublet *et al.* 2013). These sequences are
300 predicted to form secondary hairpin structures that were suspected to constitute the telomeres of linear
301 monomers (Doublet *et al.* 2013). The location of these sequences at the junctions between genome
302 units in dimers corroborate this hypothesis, under the model of monomer replication shown in Figure
303 1. In *T. rathkei*, only one base separates the heads of genome units (Figure 3). The opposite junction
304 located between the cytochrome b genes of genome units is zero- to three-bp long (not shown),
305 depending on the lineage.

306 The high sequencing depth of Illumina reads aligned to their respective dimeric genomes clearly
307 outlined heteroplasmic sites as SNPs (Figure 4). Three pairs of mirrored SNPs are shared by all
308 lineages (table 2) and correspond to variation at the three tRNA sites previously identified in *T. rathkei*
309 and *C. convexus* (Chandler *et al.* 2015). Only one of these sites (in tRNA Ala/Val, table 2) was

310 previously known to be heteroplasmic in *A. vulgare* and *A. nasatum* (Marcade *et al.* 2007; Doublet *et*
311 *al.* 2008). All lineages present the same two expected bases at very similar frequencies (~50%) at the
312 shared SNPs. The corresponding tRNA genes hence present roughly equal frequencies among
313 sequenced individuals of a lineage, or within the only sequenced animal in the case of *A. vulgare*
314 lineages (table 1). Within-individual variation was systematically observed in previous studies
315 (Doublet *et al.* 2008; Chandler *et al.* 2015) and can safely be extrapolated to all four lineages. Other
316 bases may be found at the shared SNPs (table 2) but their insignificant frequencies, measured in
317 thousandths, can be explained by sequencing errors. The *A. vulgare* WXf lineage shows three
318 additional SNPs that present the same pattern of variation as shared heteroplasmic sites, but those are
319 not located in anticodons of tRNA genes (table 2).

320 Conformation of mtDNA molecules

321 In the three *Armadillidium* lineages, the middle of successive reads that mapped on the same
322 orientation clearly clustered around the center of the junction on the reference genome (Figure S2), as
323 we predicted for reads sequenced from linear monomers (Figure 2B). This observation allowed
324 classifying DNA fragments without ambiguity (table 3). Monomers could not be detected in *T.*
325 *rathkei*, as the one-bp-long junction (Figure 3) was too short to determine the most likely mapping
326 orientation of reads, given the high error rate of SMRT sequences. In this lineage, we could classify
327 some molecules as dimers based on the sole mapping coordinates of reads. The fraction of linear
328 monomers among DNA fragments spanning the junction varies across lineages from ~30% in *A.*
329 *nasatum* to ~77% in *A. vulgare* WXf (table 3). This variability may simply reflect differences in
330 fragment size selection to be sequenced during library preparations, as sequenced fragments classified
331 as linear monomers are considerably shorter than dimers (Figure S3).

332 Asymmetry at anticodon sites of dimeric mtDNA molecules

333 In each *Armadillidium* lineage, reads that spanned all six sites of dimeric molecules indicated the
334 presence of a dominant haplotype (Figure 5). This haplotype (“GCAGGA”) is the same in *A. vulgare*
335 BF and *A. nasatum*. In *A. vulgare* WXf, the dominant haplotype (“AGGACG”) is the reverse of the
336 aforementioned one. We double-checked that the head-to-head junction of the *A. vulgare* WXf
337 reference genome was in the same orientation (strand) as that of the other two lineages. This 42-bp
338 junction between genome units (Figure 3) allows unambiguous orientation of reads, hence of
339 haplotypes. A different dominant haplotype (“ACAGGG” / “GGGACA”) is found in *T. rathkei*. We
340 cannot establish the orientation of this haplotype because reads have almost equal probability of
341 mapping on either strand of the reference genome, as explained previously.

342 Importantly, in each lineage, the prevalent haplotype carries different bases at each pair of mirrored
343 anticodons, and thus represents molecules that encode all six possible tRNAs at these loci. While
344 relatively few sequencing reads spanned all six sites without any apparent sequencing error,
345 asymmetry between genome units of a dimer is confirmed by the more numerous sequenced

346 molecules that covered at least one pair of mirrored anticodons: ~90% of them carry different bases at
347 any pair of sites (Figure 5), a result that extends to the three private SNPs of *A. vulgare* WXf
348 (haplotype counts not shown). Nevertheless, three six-base haplotypes (shown in bold in Figure 5) are
349 symmetric, mirroring the bases found in one of the two genomic units in the dominant haplotype.
350 These symmetric haplotypes are supported by a single sequenced DNA molecule each, all of which
351 were classified as dimers by the mapping coordinates of reads, rather than orientations. As coordinate-
352 based classification may be affected by incorrect *in silico* delineation of reads in raw polymerase reads
353 (Figure 2A), it is not strictly excluded that these haplotypes are in fact carried by linear monomers.
354 Most of the other minor six-site haplotypes differ from the dominant one by just one base (shown in
355 red in Figure 5).

356 Each pair of sites within genome units presents two dominant haplotypes (Figure 6) that are consistent
357 with the ones carried by the two genome units of a dimer (Figure 5). Among minor haplotypes, those
358 that could result from recombination between the dominant ones do not appear to be more frequent
359 than haplotypes that could not (Figure 6). The latter represent reads carrying one base that is very rare
360 or absent from short reads at a SNP (table 2) and most likely result from sequencing errors, which are
361 known to be quite frequent in SMRT sequences. We therefore find no convincing evidence for
362 recombinant haplotypes in the mtDNA molecules of sequenced individuals, as all minor haplotypes
363 could result from sequencing errors.

364 **Linear monomers with non-complementary bases**

365 Estimates of the fraction of linear monomers carrying non-complementary bases varied across
366 lineages (Figure 7), from ~28% in *A. vulgare* BF to ~82% in *A. vulgare* WXf. No dimeric molecule
367 was found to carry non-complementary bases, and we had no reason to expect any. By contrast, 11 to
368 35% of unclassified mtDNA fragments (those of undetermined conformation, see Methods) did
369 present non-complementary bases at heteroplasmic positions (Figure 7). Assuming that non-
370 complementary bases are restricted to monomers, then dividing the fraction of unclassified fragments
371 having non-complementary bases by the fraction of monomers having non-complementary bases
372 provides an estimate of the proportion of linear monomers among unclassified molecules (comprising
373 both monomers and dimers). We thus estimated that ~47% [95% confidence interval (CI): 19-100%]
374 of unclassified fragments are linear monomers in *A. vulgare* BF. These fractions are ~43% (95% CI:
375 36-51%) in *A. vulgare* WXf and ~14% (95% CI: 5-34%) in *A. nasatum*.

376 Looking at base complementarity on a per-site basis, we found that certain DNA fragments spanning
377 at least two sites had a mixture of sites with complementary and non-complementary bases (table 4).
378 These represent 151 fragments in *A. vulgare* WXf, and much fewer fragments in the other two
379 *Armadillidium* lineages. It should be noted that *A. vulgare* WXf comprises much more molecules
380 assigned to monomers (table 3) and was analyzed at three additional SNPs (table 2). We excluded that

381 molecules comprising a mixture of complementary and non-complementary bases resulted from
382 sequencing errors (supplementary text).

383 Discussion

384 Conformations of the atypical oniscid mtDNA

385 SMRT reads, constituted by long sequences from alternative strands of a single DNA molecule,
386 corroborate and extend previous findings on the atypical oniscid mitochondrial genome. The existence
387 of linear monomers and dimers, first inferred from enzymatic digestions (Raimond *et al.* 1999;
388 Doublet *et al.* 2012) and electron transmission microscopy (Raimond *et al.* 1999), is supported by
389 SMRT data in an unanticipated manner. Long reads confirm the existence of linear molecules
390 terminated by hairpins covalently linking their strands, since these hairpins could be navigated by the
391 DNA polymerase during sequencing (Figure 2B). Interestingly, these reads had to be sequenced from
392 ends of linear monomers that were linked to a single SMRT bell (Figure 2B). *A priori*, the technique
393 used to load the SMRT cell with DNA should favor fragments ligated to two SMRT bells: one SMRT
394 bell is attached to a magnetic bead (Magbead) rolling over the cell while the other would attach to the
395 bottom of a Zero-Mode Waveguide (ZMW) where sequencing takes place. Fragments lacking one
396 SMRT bell should not attach to the bottom of the ZMW and remain attached to the Magbead.
397 However, many of the monomer fragments may not have been attached to Magbeads and their smaller
398 size may have helped them populate ZMWs. Due to biases related to SMRT bell ligation and fragment
399 sizes, the frequencies of mtDNA molecule types in living cells cannot be directly inferred from the
400 assignment of sequenced fragments. However, estimates based on fragments with non-complementary
401 bases in “unclassified” fragments, although less direct, should not suffer these biases. All unclassified
402 fragments were linked to two SMRT bells and do not span the junction, nor do they include the hairpin
403 terminating a monomer. We see no feature, aside from the complementarity of bases at heteroplasmic
404 sites, which could differentiate those originating from linear monomers and dimers. Monomers and
405 dimers among unclassified fragments should have the same probability of being sequenced. We hence
406 consider that inferred frequencies of linear monomers reflect those in living individuals, assuming that
407 the procedures of genomic DNA extraction and library preparation did not favor a particular type of
408 molecule over the other. In *A. vulgare* lineages, these frequencies (~47% for BF and ~43% for WXf)
409 corroborate previous studies that estimated the fraction of monomers at ~50% of mtDNA molecules,
410 based on the fluorescence levels of bands in electrophoresis gels (Raimond *et al.* 1999). In *A. nasatum*,
411 the estimated frequency of monomers is three times lower at ~14%. As we have argued, variation in
412 monomer frequencies among species should be biological rather than technical. However, we cannot
413 estimate the extent to which this variation is heritable.

414 Non-complementary bases and conversion between linear and circular mtDNA 415 molecules

416 All three *Armadillidium* lineages carried linear monomers with non-complementary bases. This
417 surprising observation informs on the generation of monomeric and dimeric forms of oniscid mtDNA,
418 and on transitions between these forms. To our knowledge, genomic DNA molecules with non-
419 complementary bases between strands have not been previously reported at such high frequencies.
420 Since they cannot replicate themselves (replication produces complementary strands), they must be
421 formed by another mechanism. We propose that linear monomers with non-complementary bases
422 derive from single-stranded dimers that have renatured with themselves. Palindromic genome units
423 would become strands that are fully complementary, except at sites where the molecule is asymmetric.
424 Self-annealing of palindromic molecules has previously been proposed to explain the fact that
425 experimental denaturation reduces the length of certain digested mtDNA fragments from *A. vulgare*
426 by half (Raimond *et al.* 1999). It may also explain the drop of sequencing depth of short reads near
427 junctions between genome units (Figure 4). Self-annealing would have hindered PCR amplification of
428 palindromic fragments during Illumina library preparation or sequencing. By contrast, our
429 experimental procedure for SMRT sequencing should not have produced single-stranded DNA
430 molecules, since the whole processing of DNA has been (and must be) performed at or below room
431 temperature without denaturing agents. In living cells however, a single-stranded dimer may be
432 produced by DNA replication, during which one strand serves as template while the other strand is
433 lagging, as observed in *Drosophila* (Goddard and Wolstenholme 1980; Joers and Jacobs 2013).
434 Assuming mtDNA replication in oniscids proceeds similarly, we suggest that monomers with non-
435 complementary bases are formed by the annealing of lagging strands of asymmetric dimers before
436 these strands had a chance to serve as replication templates.

437 Since dimers presenting symmetric haplotypes are very rare (Figure 5) if they exist at all, renaturation
438 of single-stranded dimers alone cannot explain the higher fraction of monomers devoid of non-
439 complementary bases. Two other processes have been proposed for monomer formation (Doublet *et al.*
440 *et al.* 2013). One is the replication of monomers into others. Due to its conformation ending with
441 hairpins, a linear monomer should not be able to replicate into two and instead should become a dimer
442 (Figure 1). It has been suggested that monomers may replicate via a rare circular form (Doublet *et al.*
443 2013) for which we found no evidence (supplementary text). The other proposed mechanism is
444 cleavage of a dimer in two monomers, followed by the folding of telomeric hairpins (Doublet *et al.*
445 2013). Our data cannot tell whether this occurs or not. We propose that at least some monomers with
446 complementary bases are molecules in which non-complementary bases may have been corrected by
447 DNA mismatch repair enzymes (Li 2008). Correction of mismatches in some DNA fragments may
448 have been imperfect or stopped at the death of animals, explaining the existence of DNA fragments
449 combining sites with complementary and non-complementary bases (table 4). In that perspective,

450 monomers with fully complementary strands would simply be a by-product of the existing cell
451 machinery. Interestingly, the fraction of monomers with non-complementary bases is much lower in
452 *A. vulgare* BF than in other *Armadillidium* lineages. The relative rate of production of molecules with
453 fully-complementary strands must have been higher in sequenced BF individuals. We cannot yet
454 estimate the extent to which between-lineage differences are heritable or subject to environmental
455 variations.

456 Remarkably, a monomer with non-complementary bases contains the information needed to produce
457 two tRNAs at just one locus, using both DNA strands. To our knowledge, this way of compressing
458 vital information has not been reported to date. The usefulness of these molecules however depends on
459 their rate of conversion into dimers (Figure 1B), because a gene cannot be transcribed from both DNA
460 strands. Complementarity of bases proves informative on this rate of conversion *versus* the origin of
461 dimers from the replication of other dimers. If we assume that sequence variation (base
462 complementarity or symmetry) at just three sites does not significantly alter replication rate, even the
463 slightest amount of replication of monomers into dimers should eventually equalize the frequencies of
464 asymmetric haplotypes among dimers and of molecules with non-complementary bases among
465 monomers (supplementary text). Yet, these frequencies differ, especially in *A. vulgare* BF, whose
466 dimeric haplotypes are almost all asymmetric (Figure 5) whereas less than 30% of monomers have
467 mismatched bases (Figure 7).

468 If some dimers arise from monomers, symmetric dimers must suffer lower fitness to explain their
469 virtual absence within individuals. Under the assumption formulated above, lower fitness of
470 symmetric dimers can only involve the death of organelles lacking the tRNAs that these molecules do
471 not encode. The fraction of dimers deriving from monomers cannot be high in *A. vulgare* BF, or the
472 rate of organelle death explaining the almost complete absence of symmetric haplotypes would be
473 unbearable. We can therefore reasonably conclude that most, if not all, dimers derive from the
474 replication of other circular dimers rather than monomers in this lineage. In other *Armadillidium*
475 lineages where more than 78% of monomers have non-complementary bases (Figure 7), a larger
476 portion of dimers originating from monomers is compatible with the haplotype counts, under a more
477 reasonable fitness cost. Variation in the rate of monomer replication into dimers across lineages is
478 however not required to explain our results (this rate could be very low in all lineages), and is not a
479 parsimonious hypothesis.

480 While the replication of linear monomers into dimers seems to be biochemically plausible given the
481 successful SMRT sequencing of these molecules, this process may be maladaptive since it may
482 produce symmetrical dimers lacking vital tRNA genes. If monomers only rarely replicate, mechanisms
483 must be in place to ensure that organelles inherit dimeric molecules. Further studies assessing the
484 frequencies and production rates of monomers may help to determine whether these molecules are
485 vital or simply a byproduct of the replication of dimers.

486 Heteroplasmy and mitochondrial genome compaction

487 Almost all dimers present asymmetric haplotypes at the three pairs of anticodon sites. Consequently,
488 vital sequence variation between mt genome units is distributed within molecules. This observation
489 constitutes a corner case for the definition of “heteroplasmy”. If the oniscid mt genome is taken as the
490 ~14kb monomer, then within-individual variation at tRNA loci occurs between homologous genomes
491 and could fit the accepted definition. However, if dimeric molecules must be inherited by organelles
492 and cells, as our results suggest, the mt genome of these oniscids should instead be assimilated as the
493 ~28kb dimer, as only it carries all vital genes. Under this view, variation at tRNA loci occurs within
494 the genome, and may not be defined as heteroplasmy *sensu stricto*. With that in mind, some of the
495 variation between homologous tRNA loci is distributed between molecules: linear monomers may
496 present different haplotypes within a lineage and most likely within an individual. However, our
497 results provide little evidence for the replication of these molecules. Between-molecule variation in
498 these oniscids thus likely results from the continuous generation of monomers with complementary
499 bases within individuals rather than from the recurrent death of zygotes that do not inherit such
500 variation. While our findings explain the long maintenance of nucleotide variation at anticodon sites in
501 oniscids (Marcade *et al.* 2007; Doublet *et al.* 2008; Chandler *et al.* 2015), they contradict the
502 hypothesis that this variation was preserved by balancing selection (Doublet *et al.* 2008). The cost of
503 such selection – frequent death or maintenance of a system ensuring that all tRNA genes are
504 transmitted to daughter cells or organelles – may largely outweigh the benefits of mt genome
505 compaction by loss of tRNA loci. Asymmetric dimers that covalently link genomes units encoding
506 different tRNAs avoid such costs. They ensure good balance among tRNA genes in an organelle and
507 minimize the risk of transmitting molecules that do not encode certain tRNAs to organelles or cells.
508 These considerations corroborate previous suspicions that apparent heteroplasmy was permitted by
509 dimerization (Doublet *et al.* 2012; Chandler *et al.* 2015). The ancestral dimeric genome was probably
510 totally palindromic and duplicated all tRNA genes, as supported by the absence of apparent
511 heteroplasmy in certain oniscid species that show dimeric mtDNA, and from all investigated species
512 that do not (Doublet *et al.* 2008; Doublet *et al.* 2012). Then, a tRNA gene must have diverged from its
513 mirrored counterpart (asymmetry) and become identical to another tRNA gene, which gained a third
514 copy in the molecule. This could be achieved by a single point mutation in the anticodon if the two
515 tRNA genes in question were otherwise totally identical, as it is now the case for the three pairs of
516 tRNA genes in the lineages we analyzed (table 2). As long as this asymmetry subsisted, the two
517 mirrored copies among the three could be deleted without compromising viability. Three pairs of
518 mirrored tRNA loci would thus have been lost by the shared ancestor of the species we studied. The
519 evolution toward shorter, asymmetric dimers contributed to the extreme level of compaction of mt
520 genome units in oniscids (Doublet *et al.* 2015) and may have been adaptive if it saved energy for
521 mtDNA replication. Similar mtDNA compaction and tRNA production rates could have been achieved
522 by simply deleting one of the mirrored tRNA genes at several pairs of loci. The evolutionary path

523 taken clearly minimized the asymmetry between genome units (at just one nucleotide per tRNA
524 locus), possibly to sustain transition between mtDNA conformations and recombination (discussed in
525 next section).

526 Interestingly, dimerization appears to have permitted further genome compaction. While this evolution
527 seems counterproductive in terms of space saving, the net increase in molecule size should not be seen
528 as inefficient compaction. An asymmetric dimer may be slightly more efficient in storing genetic
529 content than two standard mtDNA molecules carrying all tRNA genes, and should not require more
530 energy for replication. Since returning to a standard monomeric genome without losing several tRNA
531 genes now requires an improbable chain of mutational events, the maintenance of dimeric genomes
532 tells little about the potential initial benefits of dimerization. These benefits may be revealed by
533 investigating lineages with fully palindromic dimeric genomes (i.e., without apparent heteroplasmy), if
534 any exists.

535 **Recombination and concerted evolution in a dimeric genome**

536 We cannot exclude that rare haplotypes found in each lineage result from sequencing errors. Hence,
537 each individual can be considered as carrying a single asymmetric haplotype at dimeric molecules.
538 The haplotype that is shared by *A. vulgare* BF and *A. nasatum* (Figure 5) may represent an ancestral
539 state that has been maintained since the last common ancestor of both species, ~20 Myr ago (Becking
540 *et al.* 2017). Alternatively, this haplotype may have evolved independently in these two *Armadillidium*
541 lineages. Evolutionary convergence is less parsimonious, considering that eight different haplotypes
542 (2^3 , considering their orientation with respect to the head-to-head junction) can encode all required
543 tRNAs, and all should be equivalent with respect to fitness. Long-term maintenance of a given
544 haplotype is expected, since mutation at one of the asymmetric anticodon sites produces a variant that
545 does not encode all tRNAs and that should be counter-selected. This also applies for a crossing over
546 between different genome units of two dimers. Crossing overs between genome units *within* a dimer
547 would however lead to a new haplotype encoding all required tRNAs. Such event may have occurred
548 in *A. vulgare* WXf, causing an inversion of the region encompassing the head-to-head junction
549 between the two genome units and effectively reversing the haplotype found in the other
550 *Armadillidium* lineages. Another crossing-over may have occurred between the tRNA Leu1/Leu2
551 locus and the two other loci, explaining the haplotype found in *T. rathkei* (Figure 5), a species that
552 diverged from *Armadillidium* ~40 Myr ago (Becking *et al.* 2017).

553 The lack of clear evidence for recombinant haplotypes within lineages consisting of siblings (Figure 6)
554 is not unexpected, since our survey could only have detected those produced very recently. At larger
555 evolutionary scale, recombination between mtDNA molecules most likely occurs in oniscids, like in
556 other lineages such as scorpions (Gantenbein *et al.* 2005), bivalves (Burzynski *et al.* 2003), teleost
557 fishes (Hoarau *et al.* 2002; Tatarenkov and Avise 2007), lizards (Ujvari *et al.* 2007) and even humans
558 (Slate and Gemmell 2004). Recombination in oniscid mtDNA can explain the haplotypes we

559 established, but also the almost perfect identity of genome units within highly divergent species. This
560 concerted evolution reflects two benefits that recombination certainly offers to the peculiar oniscid mt
561 genome. First, adaptive evolution of dimeric mtDNA molecules would be severely constrained
562 without recombination. In the absence of recombination, an adaptive mutation would indeed remain at
563 a “heterozygous” state until the equivalent mutation occurs at the mirrored site of the other genome
564 unit. Second, recombination restricts divergence of mirrored mitochondrial genes that are bound to
565 fulfill the same fundamental function (cellular respiration or mt protein synthesis). Replication of
566 monomers with complementary bases into dimers can also homogenize genome units, thereby offering
567 an adaptive explanation for the existence of linear monomers. We however view this process as
568 deleterious, since it should predominantly yield dimers lacking tRNAs genes, as previously argued.
569 Recombination between dimers may also produce such molecules, but this depends on the specific
570 sites of crossing overs. In this regard, it will be interesting to assess whether related isopods that are
571 not *a priori* constrained by a particular haplotype structure show higher recombination rates.
572 Regardless of the underlying mechanisms, homogenization of genome units of a dimer proceeds at a
573 moderate pace. Indeed, the genome units of *A. vulgare* WXf differ at three private sites (table 3), and
574 other similar sites have been reported in lineages from *C. convexus* and *T. rathkei* (Chandler *et al.*
575 2015). None of the three private WXf mutations are involved in the encoding of alternative tRNAs,
576 and no evidence suggests that variation at these positions is selected. Variation at these sites is simply
577 maintained through the inheritance of the asymmetric dimers carrying it. The accumulation of three
578 asymmetric mutations in *A. vulgare* WXf must have taken thousands of generations, considering the
579 mutation rate of mtDNA in isopods (Becking *et al.* 2017). Relatively long maintenance of asymmetric
580 mutations may have left more time for the loss of tRNA loci, under the evolutionary scenario we
581 described previously. Once these tRNA loci have been lost, variation at mirrored anticodons must
582 have been maintained for millions of generations by the selection of asymmetric molecules in the face
583 of homogenization of genome units.

584 Acknowledgements

585 We thank Isabelle Giraud, Thomas Becking and Lise Ernenwein for animal rearing and preparation of
586 DNA samples used for sequencing. This work was funded by European Research Council Starting
587 Grant 260729 (EndoSexDet) and Agence Nationale de la Recherche Grant ANR-15-CE32-0006-01
588 (CytoSexDet) to R.C., the 2015–2020 State-Region Planning Contract and European Regional
589 Development Fund, and intramural funds from the Centre National de la Recherche Scientifique and
590 the University of Poitiers. C.C. was funded by the National Science Foundation (grant NSF-
591 DEB1453298).

592

593 **Tables**

594

595 **Table 1.** Summary information about the four oniscid lineages used in this study.

	<i>A. vulgare</i> BF	<i>A. vulgare</i> WXf*	<i>A. nasatum</i>	<i>T. rathkei</i>
Matriline source location	Nice, France	Helsingør, Denmark	Thuré, France	Oswego, NY, USA
Illumina data				
Individuals sequenced	1 female	1 female	2 males	5 siblings
Technology		HiSeq 2000, 2×100 bp		HiSeq 2500, 2×250 bp
Sequencing center		Beckman Coulter Genomics		State University of New York at Buffalo
SMRT data				
Individuals sequenced	13 females	7 females	12 males	9 siblings
Technology		PacBio RS II, P6C4 chemistry		
Sequencing center		Genome Québec		University of Delaware

596 *Illumina sequence data were obtained by Leclercq *et al.* (2016). Other sequence data were generated
 597 for ongoing genome assembly projects.

598

599

600 **Table 2.** Location and composition of heteroplasmic sites found in the mtDNA of four oniscid
 601 lineages.

Location	Matriline	Nucleotide position	Base counts (A, C, G, T)
tRNA Leu2 (TAA) / Leu1 (TAG)	<i>A. vulgare</i> BF	9171	11611, 0, 11322, 7
	<i>A. vulgare</i> WXf	9168	7010, 1, 6977, 0
	<i>A. nasatum</i>	9176	4211, 0, 4192, 1
	<i>T. rathkei</i>	9279	3079, 0, 3260, 4
tRNA Gly (TCC) / Arg (TCG)	<i>A. vulgare</i> BF	11601	5, 13516, 12951, 7
	<i>A. vulgare</i> WXf	11604	4, 7659, 7562, 4
	<i>A. nasatum</i>	11605	6, 4352, 4377, 1
	<i>T. rathkei</i>	11718	1, 2272, 2473, 0
tRNA Val (TAC) / Ala (TGC)	<i>A. vulgare</i> BF	12004	12277, 0, 12073, 1
	<i>A. vulgare</i> WXf	12007	7014, 1, 7247, 5
	<i>A. nasatum</i>	12008	4042, 1, 4282, 3
	<i>T. rathkei</i>	12121	2260, 1, 2202, 1
tRNA Gly/Arg	<i>A. vulgare</i> WXf	11606	4, 7375, 2, 7859
nad3 gene*	<i>A. vulgare</i> WXf	11784	7397, 7400, 1, 3
12S rRNA	<i>A. vulgare</i> WXf	13474	0, 6755, 1, 6542

602 Sites that are shared across lineages are designated after the tRNAs they encode depending on the
 603 anticodon (shown in parenthesis in 5' to 3' orientation). Base counts refer to number of mapped
 604 Illumina reads carrying a given base. Positions are given in coordinates of the first genome unit.

605 *Variation at position 11784 involves a change in the nad3 protein sequence.

606
 607

608 **Table 3.** Assignments of sequenced molecules from four oniscid lineages to different conformations
 609 of mtDNA

	Linear monomers	Dimers
<i>A. vulgare</i> BF	174	351
<i>A. vulgare</i> WXf	1834	531
<i>A. nasatum</i>	112	269
<i>T. ratkhei</i>	NA*	79

610 *Molecules could not be assigned as monomers due to the impossibility to reliably infer read
 611 alignment orientation on the reference genome (see text).

612
 613
 614

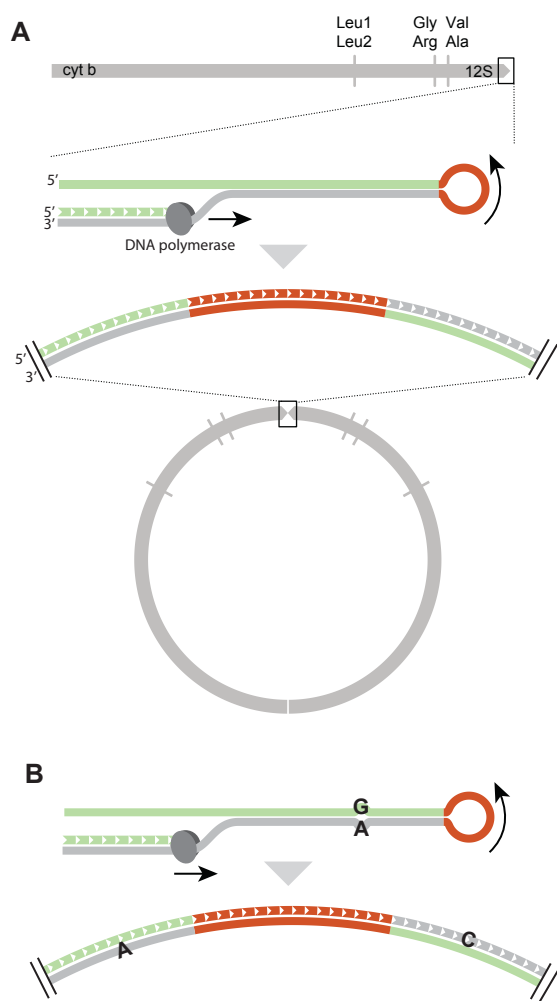
615 **Table 4.** Between-strand base complementarity at heteroplasmic sites within mtDNA fragments
 616 sequenced in three *Armadillidium* lineages.

		All complementary	All non-complementary	Mixed
<i>A. vulgare</i> WXf	Monomers	264	1193	147
	Dimers	33	0	0
	Unclassified	149	105	4
<i>A. vulgare</i> BF	Monomers	93	32	4
	Dimers	10	0	0
	Unclassified	68	9	0
<i>A. nasatum</i>	Monomers	17	57	1
	Dimers	7	0	0
	Unclassified	15	2	0

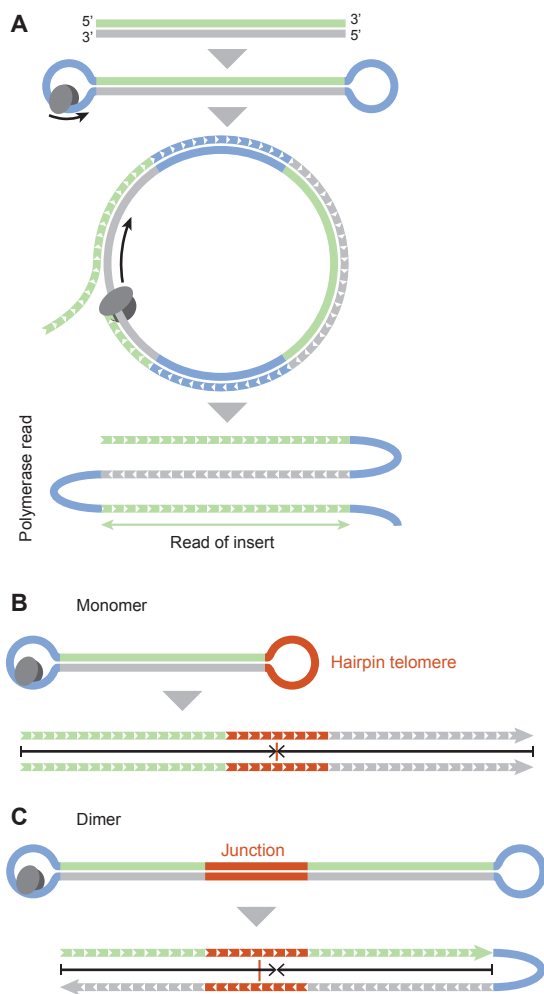
617 Only fragments covering at least two sites for which base complementarity could be assessed are
 618 listed. The “Mixed” category groups fragments having at least one site with complementary bases and
 619 at least another site with non-complementary bases.

620

621 **Figures**



622
 623 **Figure 1.** A) Hypothesized replication of a linear monomeric mtDNA molecule into a circular dimer
 624 in oniscids. A grey arrow represents a genome unit or a monomer. Its “head” is close to the 16S rRNA
 625 gene and its “tail” is close to the cytochrome b gene. Tick marks represent the locations of known
 626 heteroplasmic tRNA loci and indicate the two tRNAs that each can encode. Upon replication, the
 627 telomeric hairpin of a monomer (shown in red) becomes the junction between palindromic genome
 628 units of the circular dimer, each resulting from the replication of a monomer strand. B) Replication of
 629 a linear molecule carrying a pair of non-complementary bases leads to an asymmetric dimer carrying
 630 different bases at the mirrored positions.
 631



632
 633
 634
 635
 636
 637
 638
 639
 640
 641
 642
 643
 644
 645

Figure 2. A) Process of SMRT sequencing. A DNA fragment is bluntly ligated to two SMRT bell adapters (blue) forming hairpins and carrying a DNA polymerase. During sequencing, the newly formed fragment (striped with white arrow heads pointing towards the 3' end) leads to a polymerase read, which is composed of reads of insert (simply called “reads” for short) corresponding to alternative strands of the original fragment and separated by SMRT bell sequences. Reads are oriented as they would align to the green strand in the original fragment. B) SMRT sequencing of a molecule whose telomeric hairpin acts as a SMRT bell. All resulting reads align on the reference (which is the dimeric mitochondrial genome containing the junction between units [shown in red]) on the same orientation, and their middle (the convergence of equally-sized black arrows) corresponds to the center of the junction. C) Sequencing of a dimeric molecule covering the junction between genome units produces reads that align in alternate orientations. The middle of these reads is unlikely to correspond to the middle of the junction. Some drawings are inspired from Fichot and Norman (2013).


```

15,220      15,230      15,240      15,250      15,260      15,270      15,280      15,290      15,300      15,310      15,320
A. vulgare BF  TCCTGTTTTTTCATAGCAGAGGAGGTTAAAGTATAGTTAGGTTAAGGTTAGAAGCCTTTCCCTGT-TTGATACTATACCTTAACCTCCTCTGCTATGAAAAACAGGA
A. vulgare WXf TCCTGTTTTTTCATAGCAGAGGAGGTTAAAGCA-AGTTAGGTTAAGGTTAGAAGCCTTTCCCTGT-CATAATACTATGCCTTACCTCCTCTGCTATGAAAAACAGGA
A. nasatum    TGCTTGTTTTTCAAACAGAGGAGAT-AGAGTA-AGTTAGGTTAAGGTTAGAAGCCTTTCCCTGT-ATAATACT-TACTCT-ATCTCCTCTGTTTGAAAAACAGCA
T. rathkei    GGAAGAAGTGAGGCAGGGGAACGGGTACAGGAGTTTCTAACTCTAATAATTAAATATTAGAGTTAGAAACCTCCTGTACCGGTCCCTCGCTCACTCTTCCA

```

646

647

Figure 3. Alignment between dimeric mtDNA sequences of three *Armadillidium* lineages at the

648

region of the junction between “heads” of genome units (top), and homologous region in *T. rathkei*

649

(bottom). Bases shown in bold font over a grey background constitute the junctions that separate the

650

heads of genome units. The sequences flanking a junction are the reverse complement of each other.

651

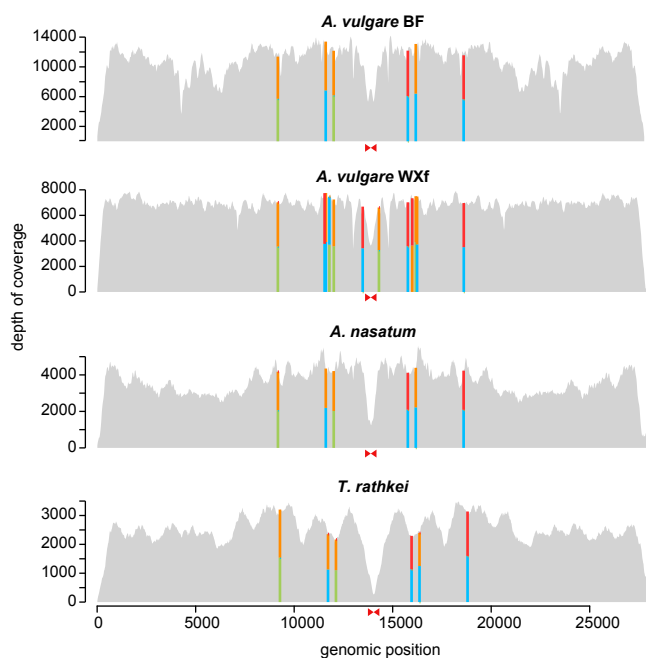
Sequences were aligned by the muscle algorithm (Edgar 2004). The *T. rathkei* region was not aligned

652

as its divergence with the other lineages would have reduced legibility.

653

654



655

656

Figure 4. Sequencing depth of short reads on the mitochondrial dimeric genomes of four oniscid

657

lineages. Colored segments indicate the presence of SNPs, each presenting two bases at very similar

658

relative frequencies (green: adenine, blue: cytosine, orange: guanine, red: thymine). Only SNPs for

659

which the rarer bases are carried by at least 20% of the mapped reads, and whose sequencing depth is

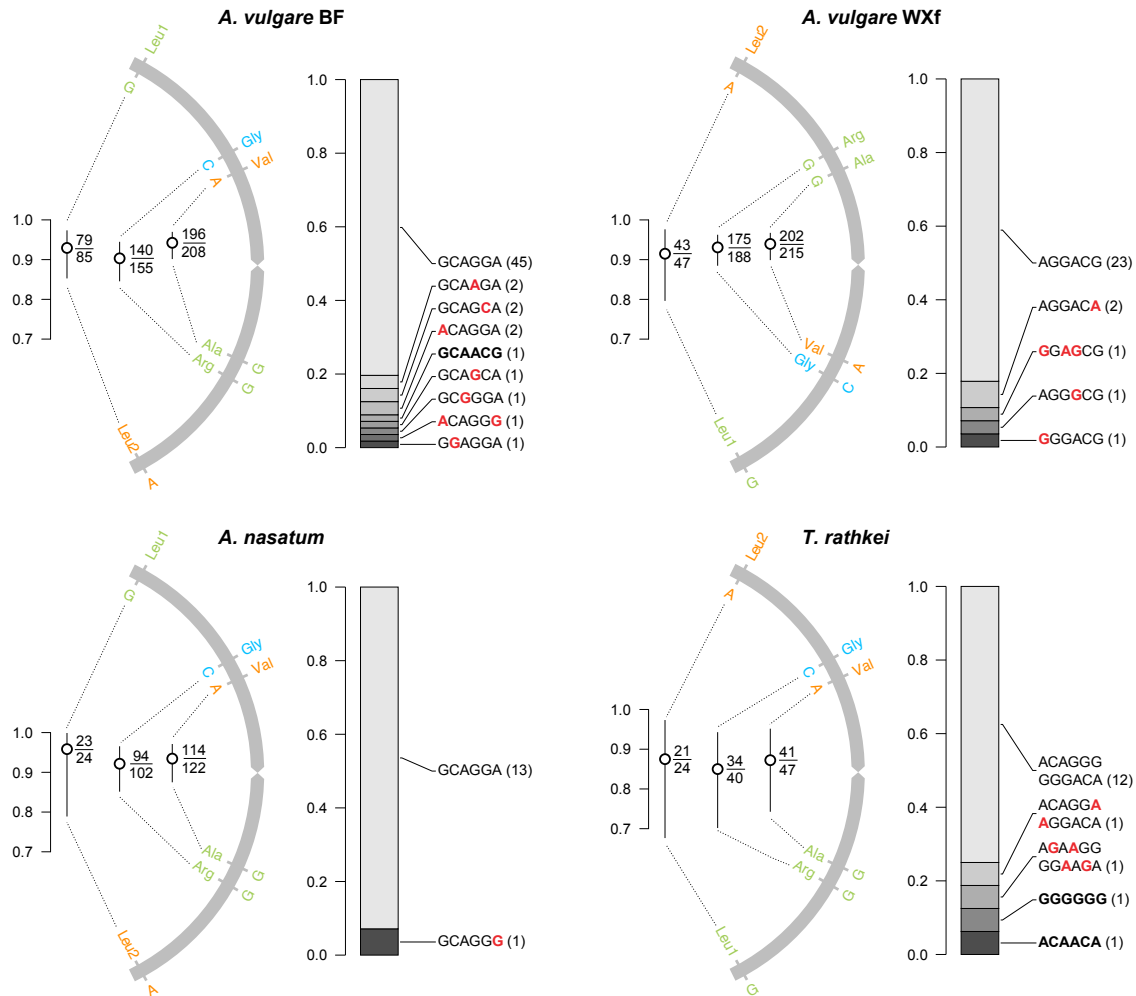
660

higher than 20% of the mean depth are shown. Converging red triangles represent the location of the

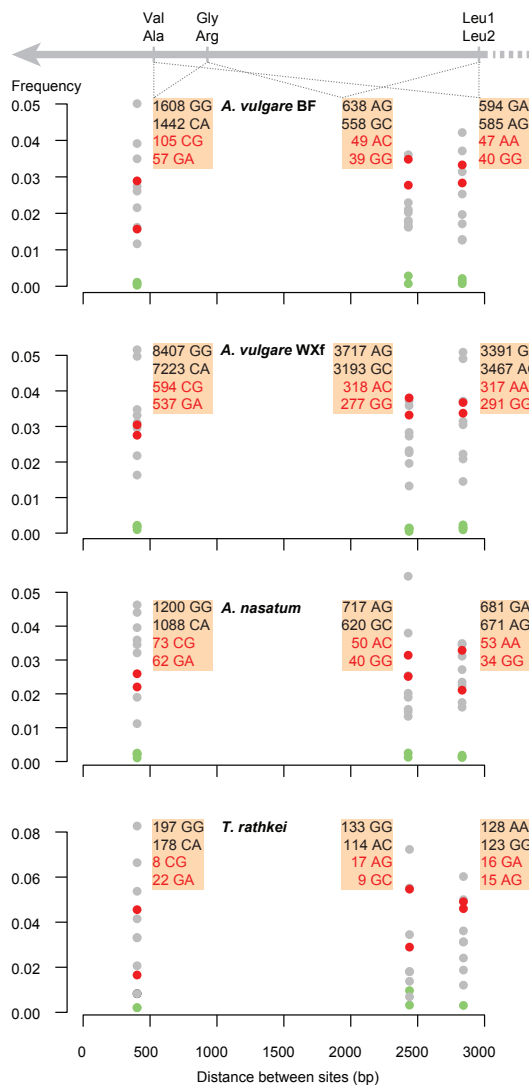
661

head-to-head junctions.

662

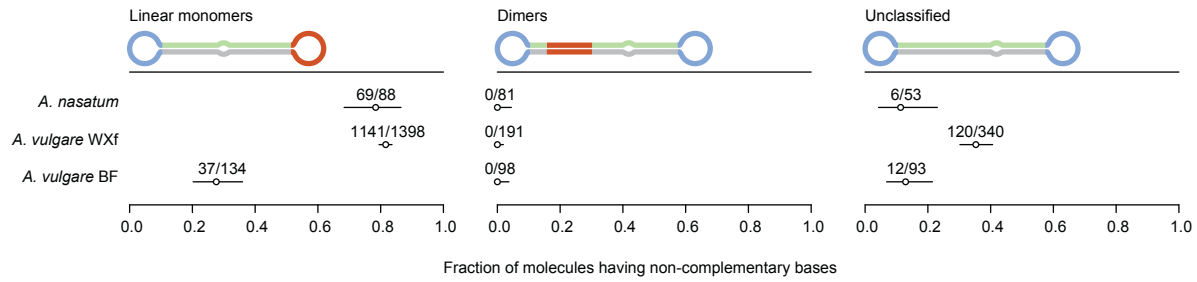


663
664 **Figure 5.** Haplotypes found in dimeric mtDNA molecules at pairs of heteroplasmic anticodon sites in
665 four oniscid lineages. Dimers are shown as converging curved grey arrows as in Figure 1A. Bases of
666 the dominant haplotype are shown on the coding strand of tRNAs for each genome unit, and
667 corresponding anticodons are indicated by the name of tRNAs in front of these bases. For each pair of
668 mirrored site, ratios represent the number of sequenced molecules carrying the dominant two-base
669 haplotype over the number of successfully sequenced molecules at these sites. Error bars represent
670 95% confidence intervals estimated by the Clopper-Pearson method of R package binom (Dorai-Raj
671 2014). Bar plots represent the fraction of sequenced molecules carrying a six-site haplotype, among
672 dimeric molecules that could be successfully sequenced at all sites. To minimize the influence of
673 sequencing errors, molecules that showed deletion or rare bases (table 2) at any of these sites were
674 ignored. Bases in red represent differences from the dominant haplotype of the same lineage.
675 Symmetrical haplotypes, which mirror the bases found in one genome unit, are shown in bold.
676 Numbers in parentheses represent the number of sequenced molecules carrying a haplotype. For *T.*
677 *rathkei*, we merged counts for each haplotype and its mirrored counterpart, as the mapping orientation
678 of reads across the palindromic genome units could not be determined with certainty (see text).
679



680
681
682
683
684
685
686
687
688
689

Figure 6. Frequencies of minor haplotypes at any pair of varying anticodon sites within mitochondrial genome units (represented as in Figure 1A), among long sequencing reads obtained from four oniscid lineages. At any pair of sites, red dots represent the two haplotypes that can, in principle, be generated by recombination between the two dominant haplotypes carried by the two different genome units. Haplotype sequences and read counts are shown in boxes next to points, with potentially recombinant haplotypes appearing in red and dominant ones in black. Grey points represent haplotypes in which one base results from a sequencing error (it is not supported by Illumina data, table 2), and green dots represent haplotypes resulting from two errors. Reads with indel variations at these sites were ignored.



Polymerase read m161013_143320_42158_c101117342550000001823258805011701_s1_p0/117417

11604	11606	11784	12007
G	G	C	A
C	A	T	C
G	G	C	A
C	A	T	C
G	T	C	A
C	C	T	C
G	G	C	A
A	A	T	C
G	G	C	A
A	C	A	C
G	-	-	A
C	A	T	C

690

691 **Figure 7.** Top: frequencies of molecules having non-complementary bases among different types of
 692 mtDNA molecules (see text) in three *Armadillidium* lineages. Red parts of molecules represent hairpin
 693 telomeres or junctions between genome units, and blue parts represent ligated SMRT bell adapters
 694 (see Figure 2). Ratios above points indicate the numbers of molecules with non-complementary bases
 695 over molecules that could be characterized for base complementarity. Error bars represent 95%
 696 confidence intervals. Bottom: example of an “unclassified” mtDNA molecule from *A. vulgare* WXf
 697 having non-complementary bases at the four variable sites it covers. These sites are named after their
 698 genomic positions (table 2). Rows represent successive reads sequenced from complementary strands
 699 (see Figure 2A). Each strand has been sequenced six times, and reads from the reverse strand (in
 700 respect to the reference genome) have been reverse-complemented. Sequencing errors are shown in
 701 grey.

702

703

704 **References**

- 705 Abascal, F., D. Posada and R. Zardoya, 2012 The evolution of the mitochondrial genetic code
706 in arthropods revisited. *Mitochondrial DNA* 23: 84-91.
- 707 Becking, T., I. Giraud, M. Raimond, B. Moumen, C. Chandler *et al.*, 2017 Diversity and
708 evolution of sex determination systems in terrestrial isopods. *Scientific Reports* in
709 press.
- 710 Boore, J. L., 1999 Animal mitochondrial genomes. *Nucleic Acids Res.* 27: 1767-1780.
- 711 Breton, S., and D. T. Stewart, 2015 Atypical mitochondrial inheritance patterns in eukaryotes.
712 *Genome* 58: 423-431.
- 713 Burzynski, A., M. Zbawicka, D. O. F. Skibinski and R. Wenne, 2003 Evidence for
714 recombination of mtDNA in the marine mussel *Mytilus trossulus* from the Baltic.
715 *Mol. Biol. Evol.* 20: 388-392.
- 716 Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST plus :
717 architecture and applications. *BMC Bioinformatics* 10.
- 718 Chaisson, M. J., and G. Tesler, 2012 Mapping single molecule sequencing reads using basic
719 local alignment with successive refinement (BLASR): application and theory. *BMC*
720 *Bioinformatics* 13: 17.
- 721 Chandler, C. H., M. Badawi, B. Moumen, P. Greve and R. Cordaux, 2015 Multiple
722 Conserved Heteroplasmic Sites in tRNA Genes in the Mitochondrial Genomes of
723 Terrestrial Isopods (Oniscidea). *G3-Genes Genomes Genetics* 5: 1317-1322.
- 724 Dickey, A. M., V. Kumar, J. K. Morgan, A. Jara-Cavieres, R. G. Shatters *et al.*, 2015 A novel
725 mitochondrial genome architecture in thrips (Insecta: Thysanoptera): extreme size
726 asymmetry among chromosomes and possible recent control region duplication. *BMC*
727 *Genomics* 16.
- 728 Dorai-Raj, S., 2014 binom: Binomial Confidence Intervals For Several Parameterizations,
729 pp., <https://cran.r-project.org/package=binom>.

730 Doublet, V., Q. Helleu, R. Raimond, C. Souty-Grosset and I. Marcade, 2013 Inverted Repeats
731 and Genome Architecture Conversions of Terrestrial Isopods Mitochondrial DNA. *J.*
732 *Mol. Evol.* 77: 107-118.

733 Doublet, V., R. Raimond, F. Grandjean, A. Lafitte, C. Souty-Grosset *et al.*, 2012 Widespread
734 atypical mitochondrial DNA structure in isopods (Crustacea, Peracarida) related to a
735 constitutive heteroplasmy in terrestrial species. *Genome* 55: 234-244.

736 Doublet, V., C. Souty-Grosset, D. Bouchon, R. Cordaux and I. Marcade, 2008 A Thirty
737 Million Year-Old Inherited Heteroplasmy. *Plos One* 3.

738 Doublet, V., E. Ubrig, A. Alioua, D. Bouchon, I. Marcade *et al.*, 2015 Large gene overlaps
739 and tRNA processing in the compact mitochondrial genome of the crustacean
740 *Armadillidium vulgare*. *Rna Biology* 12: 1159-1168.

741 Drummond, A. J., B. Ashton, S. Buxton, M. Cheung, r. A. Coope *et al.*, 2010 Geneious v5,
742 Available from <http://www.geneious.com/>, pp.

743 Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high
744 throughput. *Nucleic Acids Res.* 32: 1792-1797.

745 Fichot, E. B., and R. S. Norman, 2013 Microbial phylogenetic profiling with the Pacific
746 Biosciences sequencing platform. *Microbiome* 1.

747 Gantenbein, B., V. Fet, I. A. Gantenbein-Ritter and F. Balloux, 2005 Evidence for
748 recombination in scorpion mitochondrial DNA (Scorpiones : Buthidae). *Proc. R. Soc.*
749 *Lond., Ser. B: Biol. Sci.* 272: 697-704.

750 Gissi, C., G. Pesole, F. Mastrototaro, F. Iannelli, V. Guida *et al.*, 2010 Hypervariability of
751 Ascidian Mitochondrial Gene Order: Exposing the Myth of Deuterostome Organelle
752 Genome Stability. *Mol. Biol. Evol.* 27: 211-215.

753 Goddard, J. M., and D. R. Wolstenholme, 1980 Origin and direction of replication in
754 mitochondrial-dna molecules from the genus *Drosophila*. *Nucleic Acids Res.* 8: 741-
755 757.

756 Helfenbein, K. G., H. M. Fourcade, R. G. Vanjani and J. L. Boore, 2004 The mitochondrial
757 genome of *Paraspadella gotoi* is highly reduced and reveals that chaetognaths are a
758 sister group to protostomes. *Proc. Natl. Acad. Sci. USA* 101: 10639-10643.

759 Hoarau, G., S. Holla, R. Lescasse, W. T. Stam and J. L. Olsen, 2002 Heteroplasmy and
760 evidence for recombination in the mitochondrial control region of the flatfish
761 *Platichthys flesus*. *Mol. Biol. Evol.* 19: 2261-2264.

762 Joers, P., and H. T. Jacobs, 2013 Analysis of Replication Intermediates Indicates That
763 *Drosophila melanogaster* Mitochondrial DNA Replicates by a Strand-Coupled Theta
764 Mechanism. *Plos One* 8.

765 Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. *Nat.*
766 *Methods* 9: 357-U354.

767 Lawrence, M., W. Huber, H. Pages, P. Aboyoun, M. Carlson *et al.*, 2013 Software for
768 Computing and Annotating Genomic Ranges. *PLoS Comp. Biol.* 9.

769 Leclercq, S., J. Thézé, M. A. Chebbi, I. Giraud, B. Moumen *et al.*, 2016 Birth of a W sex
770 chromosome by horizontal transfer of *Wolbachia* bacterial symbiont genome. *Proc.*
771 *Natl. Acad. Sci. USA* 113: 15036-15041.

772 Li, G. M., 2008 Mechanisms and functions of DNA mismatch repair. *Cell Res.* 18: 85-98.

773 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence
774 Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.

775 Liu, Y. G., T. Kurokawa, M. Sekino, T. Tanabe and K. Watanabe, 2013 Complete
776 mitochondrial DNA sequence of the ark shell *Scapharca broughtonii*: An ultra-large
777 metazoan mitochondrial genome. *Comparative Biochemistry and Physiology D-*
778 *Genomics & Proteomics* 8: 72-81.

779 Marcade, I., R. Cordaux, V. Doublet, C. Debenest, D. Bouchon *et al.*, 2007 Structure and
780 evolution of the atypical mitochondrial genome of *Armadillidium vulgare* (Isopoda,
781 crustacea). *J. Mol. Evol.* 65: 651-659.

782 Okimoto, R., J. L. Macfarlane, D. O. Clary and D. R. Wolstenholme, 1992 The mitochondrial
783 genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics* 130:
784 471-498.

785 Pages, H., P. Aboyoun, R. Gentleman and S. Debroy, 2015 Biostrings: String objects
786 representing biological sequences, and matching algorithms, pp. The Comprehensive
787 R Archive Network.

788 R Core Team, 2014 *R: A Language and Environment for Statistical Computing*. R Foundation
789 for Statistical Computing, Vienna, Austria.

790 Raimond, R., I. Marcade, D. Bouchon, T. Rigaud, J. P. Bossy *et al.*, 1999 Organization of the
791 large mitochondrial genome in the isopod *Armadillidium vulgare*. *Genetics* 151: 203-
792 210.

793 Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011
794 Integrative genomics viewer. *Nat. Biotechnol.* 29: 24-26.

795 Singh, T. R., G. Tsagkogeorga, F. Delsuc, S. Blanquart, N. Shenkar *et al.*, 2009 Tunicate
796 mitogenomics and phylogenetics: peculiarities of the *Herdmania momus*
797 mitochondrial genome and support for the new chordate phylogeny. *BMC Genomics*
798 10.

799 Slate, J., and N. J. Gemmell, 2004 Eve 'n' Steve: recombination of human mitochondrial
800 DNA. *Trends Ecol. Evol.* 19: 561-563.

801 Stewart, J. B., and P. F. Chinnery, 2015 The dynamics of mitochondrial DNA heteroplasmy:
802 implications for human health and disease. *Nat. Rev. Genet.* 16: 530-542.

803 Suga, K., D. B. M. Welch, Y. Tanaka, Y. Sakakura and A. Hagiwarak, 2008 Two circular
804 chromosomes of unequal copy number make up the mitochondrial genome of the
805 rotifer *Brachionus plicatilis*. *Mol. Biol. Evol.* 25: 1129-1137.

806 Tatarenkov, A., and J. C. Avise, 2007 Rapid concerted evolution in animal mitochondrial
807 DNA. *Proc. R. Soc. Lond., Ser. B: Biol. Sci.* 274: 1795-1798.

808 Ujvari, B., M. Dowton and T. Madsen, 2007 Mitochondrial DNA recombination in a free-
809 ranging Australian lizard. *Biol. Lett.* 3: 189-192.

810 Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: An Integrated
811 Tool for Comprehensive Microbial Variant Detection and Genome Assembly
812 Improvement. PLOS ONE 9: e112963.

813 Watanabe, K., and S.-i. Yokobori, 2011 tRNA Modification and Genetic Code Variations in
814 Animal Mitochondria. Journal of Nucleic Acids 2011: 623095.

815

816