Widespread co-occurrence of two distantly-related mitochondrial genomes in individuals of the leaf beetle *Gonioctena intermedia*

Chedly Kastally and Patrick Mardulyn

Supplementary material

S1. Insect sampling, DNA extraction, MPS

Gonioctena intermedia is a European boreo-montane leaf beetle that feeds exclusively on two trees, *Prunus padus* and *Sorbus aucuparia*. Twenty-four individuals were collected from 5 sampling sites in the Belgian Ardenne on *Sorbus aucuparia* (the only host plant present in the area), between May 2014 and May 2017 (Table S1, Figure S1). Their genomic DNA was extracted from whole insects (2 pupae and 22 adults) using Dneasy Tissue Kit from Qiagen (Hilden, Germany), and following instructions from the manufacturer's protocol.

Table S1: sampling sites in the Ardenne region (see map below)

	Geographical coordinates	Sample size
Sampling site 1	49.826° N, 5.008° E	6
Sampling site 2	49.859° N, 5.019° E	3
Sampling site 3	49.935° N, 4.934° E	3
Sampling site 4	49.971° N, 4.979° E	6
Sampling site 5	49.867° N, 5.222° E	6

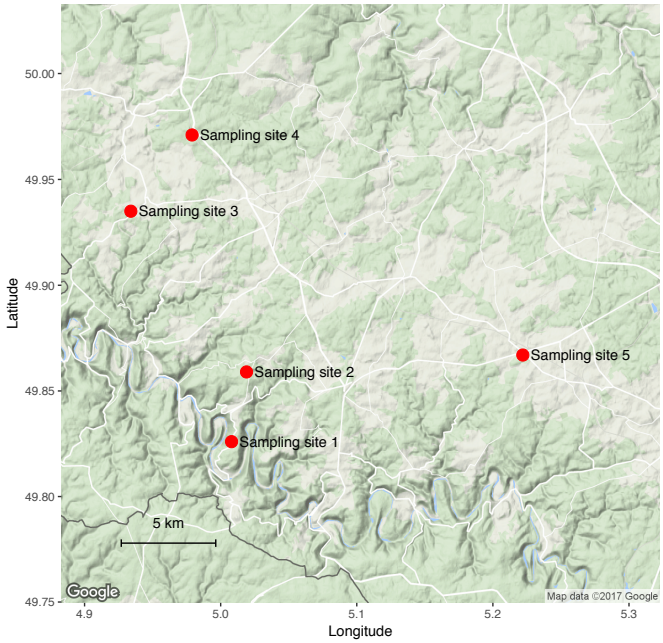

Figure S1: Map of studied region with sampled localities

DNA library preparation and MPS sequencing was performed by the Genomics Core UZ Leuven (gc.uzleuven.be), from 2 µg (Illumina libraries) or 10 µg (PacBio library) of genomic DNA extract. The following MPS datasets were generated, for two individuals (pupae) from sampling site 1, using Illumina technology: (1) ± 87 million 101bp PE reads, and (2) ± 133 million 125 bp PE reads, of 170 bp size inserts, for the first individual, and (3) ± 290 million 250 bp PE reads of 500 pb inserts (PCR-free library), for the second individual. In addition, we obtained 324427 long reads from a Pacific Bioscience (PacBio) sequencer for individual 1. The quality of the two

illumina datasets was controlled using *FastQC* v0.11.2 (www.bioinformatics.babraham.ac.uk/projects/fastqc). PacBio reads were corrected with *proovread* v2.13 (Hackl *et al.* 2014; default settings) using unfiltered Illumina reads generated from the same sample. By default, *proovread* performs quality trimming after error correction, but also provides the error-corrected untrimmed reads. To maintain the full length of the processed PacBio sequences, only the corrected but untrimmed sequences were used for further analysis.

S2. Mitochondrial genome assembly, annotation, and variant calling

Genome assembly was conducted with multiple programs, for comparison: the assemblers *Mira* v4.0.2 (Chevreux *et al.* 1999) and *Novoplasty* v2.2.2 (Dierckxsens *et al.* 2016), and two programs that use external assemblers, *MITObim* v1.8 (Hahn *et al.* 2013; using *Mira* v4.0.2) and *ARC* v1.1.3 (Hunter *et al.* 2015; using *Spades* v3.5.0, Bankevich *et al.* 2012). Each assembler was tested using random subsamples of 15 million Illumina reads. In addition, different subsamples of 7.5 and 30 million reads were tested with *Mira* and *Mitobim*. Each assembly was conducted following the software's default settings and its specific recommendations. *Novoplasty*, *MITObim* and *ARC* are programs specifically designed to assemble organelle genomes, and can use a sequence seed from the targeted genome to start the assembly. The sequence seed provided was a fragment of Cytochrome c Oxydase subunit I (COI) from *G. intermedia* available in *GenBank* under accession number *KJ785974*. Finally, an assembly using the PacBio reads was generated with *Mira*. All assemblies were then aligned pairwise and compared using Blast+ (Camacho *et al.* 2009). The overall size of the control regions estimated from the assemblies were confirmed in both individuals with a long fragment PCR amplification, using the LongAmp Taq PCR kit (*BioLabs*) with one primer located in the small subunit RNA (12s) and another in the tRNA-Met gene (Table S4), followed by agarose gel electrophoresis.

Assemblies of the mt genome from MPS reads resulted in the same sequence regardless of the program used, with the exception of the control region, more challenging to assemble due to its repetitive nature. Only the Mira assembler, either using a combination of 125 bp paired end (PE) illumina reads of 170 bp inserts and Pacific Bioscience long reads (individual 1) or 250 PE illumina reads of 500 pb inserts (individual 2), generated an assembly for this region. After excluding the control region, the assembled genomes from two individuals diverged by 25 nucleotide substitutions and no insertion/deletion.

We annotated the mitochondrial genome using MITOS Web Server (Bernt *et al.* 2013), using the default parameters of the invertebrate preset. The resulting annotation was manually curated and compared to mitochondrial annotations of the closest species found in *GenBank*: *G. aegrota*, *G. leprieuri*, *Diabrotica barberi*, *D. virgifera virgifera*, *Lygus lineolaris* and *Paleosepharia posticata*.

The typical 37 mitochondrial genes were found to be present, in the same order than the *Drosophila* mt genome (Clary and Wolstenholme 1985). Eight of the protein coding genes started with a “Ile” codon and the remaining five started with a “Met” codon. Overall genome A/T content was 78.1% excluding the control region (Table S2).

Table S2: Annotation of *G. intermedia* complete mitochondrial genome.

Gene	Direc.	Position	Anticodon	Anticodon position	Start codon (a.a.)	Stop codon
tRNA-Ile	F	1-74	GAT	30-32		
tRNA-Gln	R	77-145	TTG	115-117		
tRNA-Met	F	145-213	CAT	175-177		
ND2	F	232-1225			ATT(I)	TAA
tRNA-Trp	F	1226-1291	TCA	1258-1260		
tRNA-Cys	R	1284-1346	GCA	1317-1319		
tRNA-Tyr	R	1349-1413	GTA	1383-1385		
COI	F	1406-2948			ATC(I)	TAA
tRNA-L2	F	2949-3013	TAA	2978-2980		
COII	F	3035-3701			ATA(M)	TAA
tRNA-Lys	F	3702-3767	TTT	3732-3734		
tRNA-Asp	F	3768-3833	GTC	3798-3800		
ATPase8	F	3834-3989			ATT(I)	TAA
ATPase6	F	3983-4654			ATG(M)	TAA
COIII	F	4654-5440			ATG(M)	TAA
tRNA-Gly	F	5441-5504	TCC	5471-5473		
ND3	F	5505-5856			ATT(I)	TAA
tRNA-Ala	F	5857-5921	TGC	5886-5888		
tRNA-Arg	F	5921-5984	TCG	5950-5952		
tRNA-Asn	F	5984-6049	GTT	6015-6017		
tRNA-S1	F	6050-6116	TCT	6075-6077		
tRNA-Glu	F	6117-6180	TTC	6146-6148		
tRNA-Phe	R	6179-6242	GAA	6212-6214		
ND5	R	6243-7942			ATT(I)	TAA
tRNA-His	R	7961-8025	GTG	7995-7997		
ND4	R	8026-9358			ATG(M)	TAA
ND4l	R	9352-9633			ATA(M)	TAA
tRNA-Thr	F	9644-9706	TGT	9674-9676		
tRNA-Pro	R	9707-9772	TGG	9742-9744		
ND6	F	9787-10281			ATA(M)	TAA
CytB	F	10287-11418			ATA(M)	TAA
tRNA-S2	F	11419-11486	TGA	11448-11450		
ND1	R	11504-12451			ATA(M)	TAG
tRNA-L1	R	12456-12521	TAG	12492-12494		
16S RNA	R	12556-13197				
tRNA-Val	R	13801-13869	TAC	13839-13841		
12S RNA	R	13870-14635				
Control region	R	14636-18294				

The control region is respectively 3659 bp and 3430 bp long for individuals 1 and 2, and is largely composed of the tandem repetition of respectively 16 and 14 imperfect copies of a fragment varying in length from 116 to 130 bp, a structure similar to the one already described in closely related species (Mardulyn et al. 2003).

We then mapped a random subset of 15 million reads of the corresponding Illumina datasets to the assembly generated for each individual, but excluding the control region, with BWA-MEM v0.7.15 (Li and Durbin 2009; default parameters) and identified single-nucleotide polymorphisms (SNPs) with SAMtools v1.3.1 (Li *et al.* 2009). Quality filters of 0, 20 and 40 were applied with no noticeable difference in the results. Haplotype phasing of each sample was done using two methods: (1) using HapCUT2 (Edge *et al.* 2016; default settings) and (2) based on the read coverage reported for each allele of each identified SNP. HapCUT2 reconstructed each phase along the whole mitochondrial genome for individual 2 (MPS reads of ± 500 pb), but provided phasing for two non-overlapping blocks in the case of individual 1 (MPS reads of ± 170 pb). Both methods of haplotype phasing generated consistent results. The resulting phased haplotypes were then processed again through the MITOS annotation procedure. We checked for the presence of stop codons in sequences from each coding gene. The same procedure was

followed with different sub-samples of the original datasets, of 7.5 and 60 million reads, with and without the duplicated reads filtered out with Fastuniq (Xu *et al.* 2012), with no noticeable difference in the final results.

Mapping illumina reads to our reference assembly resulted in the inferred total number of substitutions between HF and LF haplotypes given in Table S3. Calculated d_N/d_S ratios (Table S4) suggest that most mt coding genes are subject to purifying selection.

Table S3: Pairwise number of polymorphic sites between pairs of mt genome haplotypes sequenced in two individuals using MPS. The overall number of polymorphic sites is 236, using 15 million reads mapped to our assemblies, using BWA-mem v0.7.15 and SAMtools v1.3.1 with a quality filter of 20 (error probability < 0.01).

Haplotypes	Individual 1 HF	Individual 1 LF	Individual 2 HF	Individual 2 LF
Individual 1 HF	0			
Individual 1 LF	186	0		
Individual 2 HF	25	197	0	
Individual 2 LF	164	94	169	0

Table S4: d_N/d_S ratio calculated for each coding gene, between the two haplotypes within each individual, using SequinR (Charif and Lobry 2007).

Gene	D_n/D_s	
	Individual 1	Individual 2
ND2	0.08	0.16
COI	0.07	0.02
COII	0.60	0.20
ATPase8	0.00	0.00
ATPase6	0.21	0.29
COIII	0.15	0.10
ND3	0.00	0.00
ND5	0.07	0.10
ND4	0.11	0.14
ND4l	0.66	0.66
ND6	0.15	0.30
CytB	0.25	0.19
ND1	0.11	0.12

S3. Mitochondrial/nuclear genomic coverage ratio assessment

To evaluate the possibility that one mt variant is of nuclear origin, we have estimated the coverage of both mt variants in MPS data relative to the coverage of the nuclear genome. For this purpose, the subset of reads (15 millions) used previously for variant calling (via mapping of reads to the reference mitochondrial genome) were mapped to 848 nuclear genes previously identified via a preliminary assembly of the nuclear genome, using Discovar-De-Novo (Weisenfeld *et al.* 2014) and BUSCO *v3* (for gene identification; Simão *et al.* 2015). The mapping

resulted in a median estimate of coverage ranging from 24x to 33x for individuals 1 and 2, respectively. In comparison, the coverage of the HF mt haplotype, found in only 50% of the sampled insects, is six times and almost 14 times larger, for individuals 1 and 2, making it very likely of mitochondrial origin.

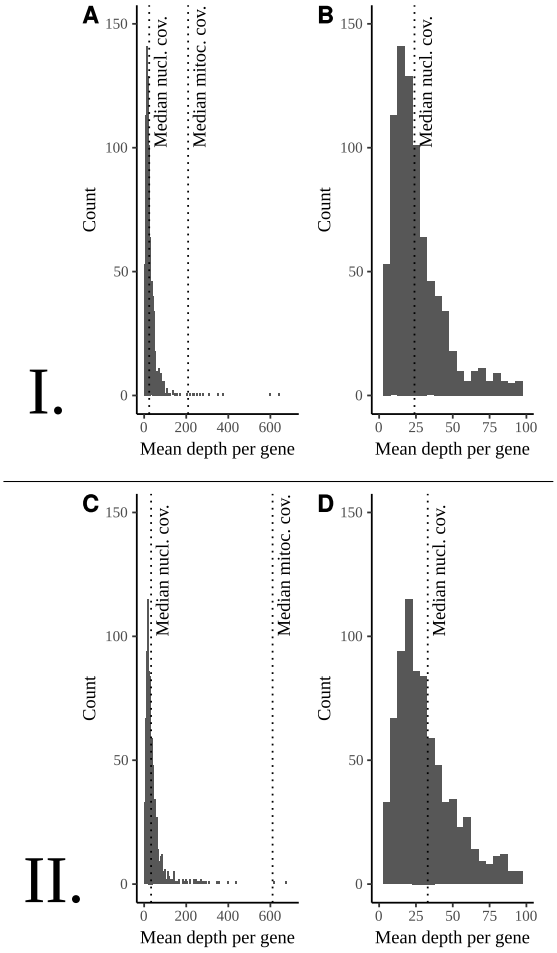

Figure S2. Distribution of mean coverage for 848 nuclear genes, using a subset of 15 million reads for each sampled individual (I, individual 1; II, individual 2). Dashed lines mark, from left to right, the median of the distribution, (I.: 24, II.:33) and the median of the mitochondrial coverage (I.: 209, II.: 610.5). B and D: zoom-in (mean depth per gene between 0 and 100) of distribution in A and C, respectively.

S4. PCR and Sanger sequencing of mt sequences

We have PCR-amplified and sequenced (Sanger Sequencing) a COI fragment for both individuals sequenced via MPS. The sequence obtained for the HF haplotype was identical to the one obtained from MPS reads, in the case of both individuals. Comparing the HF and LF haplotypes from the initial MPS assemblies identified, in this COI fragment, 11 and 15 polymorphic sites (SNPs), respectively, for individuals 1 and 2. For individual 1, Sanger sequencing of the LF haplotype confirmed five unambiguous nucleotide substitutions between HF and LF. Four other

previously identified SNPs were found heterozygous (double peaks on the sequencing electropherograms), suggesting the existence of at least two variants of the LF haplotype in individual 1. Two sites initially suggested as polymorphic by MPS data were in fact monomorphic, and one site that was initially suggested polymorphic for individual 2 but monomorphic for individual 1, was in fact polymorphic for individual 1 as well. For individual 2, Sanger sequencing of the LF haplotype confirmed six unambiguous nucleotide substitutions between HF and LF. Seven other previously identified SNPs were found heterozygous for the two corresponding nucleotides, again, suggesting the existence of at least two variants of the LF haplotype in individual 2. Three sites initially suggested as polymorphic were in fact monomorphic.

Table S5: Primers used for PCR amplifications

Primer sequence	Targeted locus/haplotype	PCR annealing temperature	Origin
AAAGCGACGGGCGATATGTGC	Control region	47°C	This study
TAACCTTYATAAATGGGGTATG	Control region	47°C	Mardulyn et al 2003
TATCTATGTTCAGCAGGAGGAAGC	COI fragment, HF haplotype	65°C	This study
GTTTTTTGATCCGGCAGGTGGG	COI fragment, HF haplotype	65°C	This study
TATCTATGTTCAGCAGGAGGAAGT	COI fragment, LF haplotype	65°C	This study
GTTTTTTGATCCGGCAGGTGGA	COI fragment, LF haplotype	65°C	This study
GGAGCTCCTGATATAGCWTTYCC	COI fragment, universal	52°C	Simon et al 1994
TCCAATGCACTAATCTGCCATATTA	COI fragment, universal	52°C	Simon et al 1994

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers, *Annals of the Entomological Society of America*, **87**, 651-701

All PCRs conducted on additional individuals generated a PCR product, which was subsequently sequenced with an ABI 3730 automated sequencer, using the BigDye Terminator V 3.1 Sequencing Kit (Applied Biosystems). Both strands were sequenced for each PCR product. With 11 individuals, the HF-specific and LF-specific PCRs produced a HF and LF sequence, respectively, and the third PCR (universal primers) produced the same HF sequence as in the first PCR. This pattern was interpreted as a signature of a heteroplasmic HF/LF genotype. With 12 individuals, all three PCRs (HF-specific, LF-specific, universal) produced the same LF sequence. This pattern was interpreted as a signature of a homoplasmic LF genotype. With one individual, all three PCRs produced the same HF sequence. This pattern was interpreted as a signature of a homoplasmic HF genotype. The distribution of these genotypes among sampling sites is shown in Figure 2.

From these results, we can infer that (1) the two specific primer pairs will only amplify their intended target haplotype if present, but in case of its absence, will amplify the other haplotype present, and (2) that the universal primer pair will preferentially amplify the HF haplotype when both HF and LF are present, to the extent that the LF haplotype will not be detected by sequencing the PCR product. The "universal" primer pair was proposed by Simon et al (1994) to amplify COI in different insect orders, and it is worth noting that their sequences correspond exactly to the sequences of the corresponding priming sites on both haplotypes HF and LF (no mutation). Preferential amplification of HF can therefore not be explained by mutations occurring at the priming sites. Consequently, we can infer that the detection limit for identifying a haplotype variant using Sanger sequencing is > 20% (since the proportion of LF variants in individuals 1 and 2 was estimated to ± 20% with MPS data). Also, since more than one LF variant was detected (double peaks on the sequencing chromatograms) in individuals 1 and 2 when specifically amplifying the LF variant, it is likely these variants have a frequency > 20%.

Comparing the LF haplotype sequence to sequence variation over the entire species distribution (Quinzin and Mardulyn, 2014) shows that it is more closely related to more distant haplotypes (h27 and h34) found in the Ural Mountains (eastern Europe) and in Finland (northern Europe) than they are to the HF haplotype (previously detected in the Ardennes region).

Phylogenetic relationships among haplotype sequences were inferred under the maximum parsimony with PAUP v. 4b10 (Swofford 2003). The haplotype network of Figure 2 is drawn from the single inferred most parsimonious tree.

Tajima's D (Tajima 1989), a statistic that was initially developed to detect selection, was calculated from all COI sequences and resulted in a significantly positive value (1.78, $p < 0.04$), that can be interpreted as a signal of balancing selection. Other interpretations are also possible however, as specific demographic histories (mostly a population bottleneck) and/or population structure can also generate positive D values, and this pattern of variation alone is not sufficient to conclude that balancing selection occurs.

S5. Heteroplasmy, numt, female sperm storage or cross-sample contamination?

Apparent intra-individual mtDNA polymorphism can indicate co-existence of two different mtDNA variants (heteroplasmy), or the presence of one or more copy of a mt sequence inside the nuclear genome (numt) (e.g., Zhand and Hewitt 1996). We are confident that individual heteroplasmy, rather than the presence of numts, explains our observations. First, we found 11 insects with only the LF haplotype, and one insect with only the HF variant. If LF sequences were of numt origin, we would have expected to find the HF variant, i.e., the mt sequence in that case, in all individuals (and vice versa if the HF sequence belonged to a numt). Second, the translation of the inferred DNA sequences into amino-acid sequences for all protein-coding genes of both mt genomes (HF and LF) did not reveal a single stop codon disrupting the protein sequence, which would have been expected in the case of a numt that is not constrained by selection (pseudogene). Finally, MPS data gathered for two individuals showed that reads from

both HF and LF haplotypes had a much larger coverage than a sample of 848 presumably single copy genes from the nuclear genome (see section S3, "Mitochondrial/nuclear genomic coverage ratio assessment"), as expected if of mt origin.

In addition, two strong arguments can be put forward against the hypothesis that our observations result from cross sample contamination. First, cross contamination between the two first individuals for which genomic DNA was extracted, those sequenced via MPS, would result in the presence of haplotype from individual 1 in the extract of individual 2, and vice versa. We should have found the same two haplotypes in both individuals under this hypothesis. Our observations show in fact 4 different haplotypes for the two individuals. In other words, the two haplotypes identified in one individual are not found in the other individual (and vice-versa), which allow us to rule out cross-contamination. Second, MPS data allowed us to quantify the relative amount of each haplotype, and we estimated the ratio of one haplotype to the other being ± 4:1 (80% haplotype HF, 20% haplotype LF, in both cases). If this was explained by cross-contamination, it would mean we had transferred a really large amount of DNA extract from one sample to the other (at least as much as $1/5^{th}$ of a DNA extract, and up to 4 times the original amount in the case where haplotype HF would be the contaminant), which seems highly unlikely. Cross-contamination among samples, while possible if the experimenter is not careful (and we were careful), typically involves the transfer of only minute amounts (aerosol droplets) of one sample to another, and is mainly a problem for downstream PCR (MPS library construction for individual 2 was PCR free).

Finally, our observation of intra-individual mtDNA polymorphism cannot be attributed to an artefact of sperm storage by females, because heteroplasmy was also detected in pupae and male adults. Indeed, we extracted genomic DNA from pupae (2 individuals) to acquire the NGS data analysed in this study. Heteroplasmy was detected in both individuals, with MPS and Sanger sequencing (COI). To specifically verify this hypothesis, we also extracted genomic DNA from 8 additional individuals, all males, from another locality sampled in May 2017 (coordinates 49.858° N, 5.151° E), and genotyped them (COI) using the procedure described in S4. In this case, heteroplasmy was detected in 3 males out of 8 (genotype LF/HF, versus genotype LF for the others).

References

Bankevich A, Nurk S, Antipov D *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology,* **19**, 455–77.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF (2013) MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution,* **69**, 313–319.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Charif D, Lobry JR (2007) SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In "*Structural approaches to sequence evolution: molecules, networks, populations", (eds: U Bastolla, M Porto, HE Roman, M Vendruscolo), pp* 207-232, Springer, Berlin Heidelberg.

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Proceedings of the German Conference on Bioinformatics (GCB)*, **99**, 45–56.

Clary DO, Wolstenholme DR (1985) The mitochondrial DNA molecule of Drosophila yakuba: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution*, **22**, 252-271

Dierckxsens N, Mardulyn P, Smits G (2016) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, **45**, e18.

Edge P, Bafna V, Bansal V (2016) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, **27**, 801-812.

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47- 50.

Hackl,T, Hedrich R, Schultz J, Förster F (2014) Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, **30**, 3004–11.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads–a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**, e129.

Hunter SS, Lyon RT, Sarver BAJ, Hardwick K, Forney LJ, Settles ML (2015) Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. bioRxiv. http://biorxiv.org/lookup/doi/10.1101/014662

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–60.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Mardulyn P, Termonia A, Milinkovitch MC (2003) Structure and evolution of the mitochondrial control region of leaf beetles (Coleoptera: Chrysomelidae): a hierarchical analysis of nucleotide sequence variation. *Journal of Molecular Evolution*, **56**, 38–45.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210-3212.

Swofford DL (2003) PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods), v. 4b10. Sinauer Associates, Sunderland, Massachusetts, USA.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585-595.

Weisenfeld NI, Yin S, Sharpe T, *et al.* (2014) Comprehensive variation discovery in single human genomes. *Nature Genetics*, **46**, 1350–1355.

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S (2012) FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE*, **7**, e52249.

Zhang DX, Hewitt GM. 1996 Nuclear integrations: challenges for mitochondrial DNA markers. *TREE* **11**, 247–251.