Using next-generation sequencing to improve DNA barcoding? Lessons from a small scale study of wild bee species of Halictidae

Gontran Sonet[1*], Alain Pauly[1], Zoltán T. Nagy[1], Massimiliano Virgilio[2], Kurt Jordaens[2,3], Jeroen Van Houdt[4], Sebastian Worms[5], Marc De Meyer[2] and Thierry Backeljau[1,3]

[1] Directorate Taxonomy and Phylogeny (JEMU), Royal Belgian Institute of Natural Sciences, 29 Vautierstraat, B-1000 Brussels, Belgium.

[2] Department of Biology (JEMU), Royal Museum for Central Africa, 13 Leuvensesteenweg, B-3080 Tervuren, Belgium.

[3] Evolutionary Ecology Group, University of Antwerp, 1 Universiteitsplein, B-2610 Antwerp, Belgium.

[4] Genomics Core KULeuven – UZLeuven, Gasthuisberg O&N1, 49 Herestraat - box 602, B-3000 Leuven, Belgium.

[5] Institute of Life Sciences, Université catholique de Louvain, 4-5 Croix du Sud, B-1348 Louvain-la-Neuve, Belgium.

* Corresponding author: Gontran Sonet, gontran.sonet@naturalsciences.be

Short title: Applying NGS to study halictid bees

Running title: Applying NGS to study halictid bees

Abstract

The parallel sequencing of targeted amplicons is a scalable application of next-generation sequencing (NGS) that can advantageously replace Sanger sequencing in certain DNA barcoding studies. It can be used to sequence different PCR products simultaneously, including co-amplified products. Here we explore the interest of this approach by simultaneously sequencing five markers (including the DNA barcode and a diagnostic marker of *Wolbachia*) in 12 species of Halictinae that were previously DNA barcoded using the method of Sanger. Sequences were obtained at a moderate cost from fresh bees with success rates of 74-100% depending on the DNA fragment. They did not enable a better species delineation but permitted to improve the phylogeny of the group, detect *Wolbachia* infections and characterize haplotype variants. We provide guidelines for selecting NGS or Sanger sequencing depending on the goals of future studies and whether variant haplotypes are expected or not.

Keywords

NGS; heteroplasmy; numt; *smaragdulus*; *Wolbachia*

1. Introduction

DNA barcoding is a standardised and widely used method to identify specimens at the species level using a restricted set of short DNA fragments (Hebert et al. 2003). Developed in 2003, the standard DNA barcoding protocol relies on Sanger sequencing, but with the development of next-generation sequencing (NGS) technologies, new methods involving the parallel sequencing of targeted amplicons were proposed to improve or complement the standard DNA barcoding pipeline (Batovska et al. 2017; Wilkinson et al. 2017). These methods, applied as "next-generation DNA barcoding" (Shokralla et al. 2014) or "targeted amplicon sequencing" (Bybee et al. 2011) enable the analysis of mixtures of DNA fragments that are co-amplified during PCR or obtained by pooling different PCR products. In bee systematics, these methods can be used to (1) sequence several loci at a moderate cost and improve phylogenies that are only inferred from COI sequences and (2) investigate the presence of cytoplasmic endosymbiotic bacteria such as *Wolbachia* Hertig 1936 emend. Weiss 1974 (Breeuwer & Werren 1993; James et al. 2002; Hiroki et al. 2004; Raychoudhury et al. 2010) that are frequently detected in Halictidae and can affect the transmission of the mitochondrial genome (Smith et al. 2012). They can also be used to (3) detect variants in the PCR products that can be due to heterozygosity, heteroplasmy – variants of mitochondrial DNA – or nuclear COI pseudogenes – nuclear copies of mitochondrial DNA, (numts) – (Buhay 2009); all of these can affect gene trees in Hymenoptera (Magnacca & Brown 2010; Cristiano et al. 2012).

Here we implemented the parallel sequencing of targeted amplicons to (1) re-sequence COI, (2) sequence three additional nuclear gene fragments and (3) sequence a fragment

of the *Wolbachia* outer surface protein gene in 12 Halictinae species that were recently studied by DNA barcoding using Sanger sequencing (Pauly et al. 2015). These species belong to *Halictus* (*Seladonia*) Robertson 1918 [or *Seladonia* depending on the classification considering *Seladonia* as a subgenus (Michener 2007) or a genus (Pesenko 1999, 2004)] and include five species belonging to the *H. smaragdulus* Vachal 1895 [or *S. smaragdula*] species complex. While COI data strongly supported the delineation of these five species, they neither fully resolved their phylogenetic relationships (Pauly et al. 2015). Our small-scale NGS implementation was evaluated for its efficiency to assemble a multilocus dataset and its usefulness toimprove phylogenetic resolution, assess the presence of *Wolbachia* and detect variants in the PCR products.

2. Material and methods

Sampling and DNA sequencing

We sampled 21 specimens (Table I) representing five of the six species of the *smaragdulus* complex and seven closely related species showing the smallest interspecific p-distances at COI with respect to the *smaragdulus* complex (Pauly et al. 2015). One species of the complex, *H. Cretellus* **(**Pauly & Devalez 2015) is only known from Crete (Pauly et al. 2015; Schmidt et al. 2015) and was not sampled here. Most specimens were collected after 2011 and two specimens >40 years ago (AP030 in 1973 and AP048 in 1890). All specimens were captured with a net, killed with ethyl acetate and stored in absolute ethanol. Genomic DNA was extracted from one middle leg using the NucleoSpin Tissue Kit (Macherey-Nagel, Germany). We targeted five gene fragments, including three nuclear markers that were previously used for phylogenetic

analysis in hymenopterans, viz. wingless (wnt1), white (w) and a hippo gene (HOG7036-02) for putative serine/threonine kinase, exons 1-2 (Danforth et al. 2004; Kawakita et al. 2008; Gibbs et al. 2012; Hartig et al. 2012). The 5' end of the COI gene (standard barcode for animals) was sequenced and used for both phylogenetic reconstruction and to assess the presence of COI pseudogenes, heteroplasmy and contaminant *Wolbachia* COI. Finally, the *Wolbachia* outer surface protein (wsp) was used to survey the presence of *Wolbachia* in the bee specimens.

Two consecutive PCRs performed using the Multiplex PCR Kit (QIAGEN, The Netherlands) allowed the amplification of six DNA fragments from the five targeted genes (including two overlapping fragments for COI) and the addition of specimen specific molecular identifiers (MIDs) and the Illumina adapters of the TruSeq Custom Amplicon kit (Illumina, USA). PCR products were pooled together and sequenced in one lane of a MiSeq Sequencing System flow cell (Illumina, USA) using the paired-end protocol of the Reagent Nano Kit v. 2 (2x250 bp). Additional details on the protocol are provided in the Online Resource 1.

Data analysis

MiSeq data were demultiplexed and cleaned using Trimmomatic v. 0.32 (Bolger et al. 2014) and AlienTrimmer v. 0.4.0 (Criscuolo & Brisse 2013). Paired-end reads were then assembled with PEAR v. 0.9.6 (Zhang et al. 2014) and NextAllele (O'Neill et al. 2013) was used to identify the reads obtained for each targeted fragment and to get the consensus sequences .

Reads obtained for wsp were used to identify *Wolbachia* haplotypes using the *Wolbachia* wsp typing module of the *Wolbachia* multilocus sequence typing (MLST)

system (Baldo et al. 2006), a central depository of *Wolbachia* bacterial and host information (Jolley & Maiden 2010). Heteroplasmy (for COI), heterozygosity (for nuclear genes) and undesired co-amplified products (paralogues or contaminants) were investigated using assemblies with sequencing depth (number of reads per position) >20. For these assemblies, we calculated the average rate of substitution per base and used Geneious v. 10.2.3 (Kearse et al. 2012) to examine all variant nucleotide characters showing a relative frequency >10%, a value known to be much higher than sequencing error rates reported for different DNA library preparations and sequencing with the Illumina (Illumina, USA) platform (Schirmer et al. 2016).

Phylogenetic analyses were repeated on different datasets in order to compare topologies and resolutions obtained with separate gene fragments, with nuclear data and with the concatenations of all gene fragments: COI (21 specimens x 658 bp), wnt1 (14 x 383), w (21 x 384), HOG7036-02 (19 x 417), the concatenations of the three nuclear fragments (17 x 1184), the four fragments (13 x 1842), the four fragments allowing one missing fragment per specimen (17 x 1842) and the three fragments obtained for the higher number of specimens (COI, w and HOG7036-02, 17 x 1459). In order to assess the added value of including nuclear fragments to a COI phylogeny, we tested a last COI dataset including only specimens used in the concatenated datasets (17 x 658). We extracted unique haplotypes using the R packages APE (Paradis et al. 2004) and pegas (Paradis 2010). When alternative haplotypes were observed for the same individual, all haplotypes were searched on GenBank using BLAST (Boratyn et al. 2013) and phylogenetic analyses were repeated with the different haplotypes. Sequences of two outgroup taxa, one Halictidae, *Dufourea novaeangliae* (Robertson 1897), and one Apidae, *Apis* sp., were retrieved from GenBank (Table I). Neighbour-joining trees were

constructed in MEGA with pairwise deletion and 1000 bootstrap pseudo-replicates. Maximum parsimony (MP) trees were searched using the R package phangorn (Schliep 2011), using the parsimony ratchet heuristic method (Nixon 1999), with characters of equal weights, gaps considered as missing data and using 500 non-parametric bootstrap replicates. For Bayesian phylogeny inference (BI), best partition scheme (see Online Resource 1) and best-fit substitution models were estimated using PartitionFinder v. 1.1 (Lanfear et al. 2014). BI analyses were performed with MrBayes v. 3.2.6 (Ronquist et al. 2012) and two parallel runs with four chains each were run for five million generations, with unlinked nucleotide substitution parameters for each data partition. Every 1000th generation was sampled, and the first 25% of the trees were discarded ("burn-in"). Convergence was monitored and average standard deviation of split frequencies was <0.01 after five million generations. Additional details on data analysis are provided in the Online Resource 1.

3. Results

Data collection

Overall, 260214 reads (paired and unpaired) were assigned to the targeted gene fragments (Figure 1). Read quality scores (Phred) ranged from 28 to 40 (mean values between 38 and 39 depending on the specimens). Numbers of reads per specimen obtained for each DNA fragment (Table I) ranged from zero (for old museum specimens) to 21336 reads (for w in AP031). The COI consensus sequences were identical to the COI sequences obtained by Sanger sequencing (Pauly et al. 2015) when sequencing depth was ≥5. Hence, we considered that a sequencing depth <5 was not reliable and consensus sequences were never considered when sequencing depth was

<5. Following this rule, haplotypes were obtained for the fresh specimens with success rates of 100% for COI (aligned length of 658 bp) and w (385 bp), 95% for HOG7036-02 (417 bp) and 74% for wnt1 (383 bp). Concerning the older museum specimens collected in 1890 (Greece) and 1973 (Iran), sequencing depth was always <5 except for w of AP048 (Table I). We also obtained *Wolbachia* COI for specimen AP001 (see below). The COI alignment comprised 105 variable sites overall and 88 within the species complex (interspecific p-distances from 2.3 to 12.5%). The nuclear data (three markers) comprised 36 variable sites overall and 7 within the complex (interspecific p-distances from 0 to 2.6%).

Phylogenetic analyses

The phylogenetic relationships within the *H. smaragdulus* complex were entirely resolved with the highest support in the BI (Figure 3) using the concatenation of all DNA fragments (COI, wnt1, w and HOG7036-02) and including either 13 specimens (no missing data) or 17 specimens (missing data for maximum one marker). Topologies obtained with BI and MP were consistent for all concatenated datasets and variant haplotypes (see below) affected neither the topology, nor the support of the trees. Phylogenies obtained using COI only were slightly less resolved (see support in Figure 3). Those solely based on nuclear data (both separate and concatenated datasets) only supported a few nodes outside the species complex (Online Resources 2). The only nodes that were never resolved concerned the relationships among *H. seladonia*, *H. lucidipennis* and the clade grouping *H. subauratus* with *H. subauratoides*.

Wolbachia infection

Wolbachia sequences of wsp were obtained in eight out of the 21 specimens, with 14 to 831 reads per specimen (Table I). The eight wsp positive specimens belonged to five species (Table I): *H. cephalicus* (2 detections/2 specimens), *H. seladonia* (2/2), *H. subauratus* (1/1), *H. smaragdulus* (2/2) and *H. gemmellus* (1/2). All haplotypes queried in the *Wolbachia* MLST database provided a perfect match with *Wolbachia* sequences of the supergroup A, a clade of *Wolbachia* strains commonly found in Hymenoptera (Casiraghi et al. 2005; Ros et al. 2009; Gerth et al. 2011). Five different sequences of the hypervariable region 1 (HVR1) of wsp, coded as numbers 1, 11, 13, 51 and 53 in the *Wolbachia* MLST database were observed. One or two different HVR1 sequences were detected per specimen. We observed mainly HVR1: 11 in *H. cephalicus*, HVR1: 51 in *H. seladonia* and HVR1: 11 and 1 in *H. smaragdulus* (Figure 2). *Wolbachia* COI was only sequenced in one specimen, AP001, which was also positive for wsp. No exact match was found for this sequence in the MLST database but best matches in GenBank were 99% similar (100% sequence coverage) and mostly comprised *Wolbachia* COI from hymenopterans.

Detection of variant haplotypes

The average rate of substitution per base varied from 0.002 to 0.004 depending on the assemblies. Variant nucleotide characters were found in 10% to 50% of the reads of wnt1 (in 6 specimens), w (1) and COI (3) (Table II). Two variant characters were observed with relative frequencies of 0.45 and 0.50 in wnt1 and w, respectively and within one single specimen (AP027, *H. lucidipennis*). Other variant characters found in wnt1 with a frequency of 0.11 were situated at the end of the reads (Table II). Finally,

199 the variant characters found in COI occurred in 10-23% of the reads of *H. lucidipennis*
200 (8 positions) and both specimens of *H. seladonia* (17 and 26 positions). Most of them
201 (49/51) corresponded to synonymous substitutions and were observed with a high
202 sequencing depth and in good quality reads. The intra-individual p-distances among
203 these haplotypes were 0.26% for nuclear genes and ≤4% for COI (0.2-2.7% within
204 AP027 and 0.2-4.0% within AP055). These values were within the range of
205 interspecific distances measured here (0-2.6% for nuclear genes and 2.3-12.5% for COI)
206 but below the distance to the nearest-neighbours of these specimens (>1.1% for nuclear
207 data and >7.8% for COI). The use of the alternative haplotypes did neither affect the
208 results of our phylogenetic reconstructions nor the best matches in GenBank. No variant
209 was observed for *Wolbachia* COI.

4. Discussion

Parallel sequencing of PCR amplicons is most effective when limited sequence data are targeted per specimen (Mamanova et al. 2010; Grover et al. 2012). This is the case for DNA barcoding or small-scale multilocus phylogenetic analyses. In comparison with Sanger sequencing, it can improve the sequencing sensitivity (fewer false negatives) and accuracy by enabling the simultaneous detection of co-amplified products such as homologues, paralogues and contaminants (Grover et al. 2012; Shokralla et al. 2014) with a moderate cost (Bybee et al. 2011). Below we evaluate the added value of the protocol applied here compared to standard DNA barcoding using Sanger sequencing. Data collection and cost-efficiency

Success rate of parallel amplicon sequencing is expected to highly depend on the PCR amplification. For COI (the only marker that was sequenced both by NGS and Sanger), the usage of NGS did not produce a more complete dataset than with Sanger sequencing since COI could only be obtained from fresh specimens in both cases. The low sequencing depths obtained here for both older museum specimens were not considered reliable. The total cost of this analysis was 1500 € (excluding VAT and labour cost, and rounded to a multiple of 5). The cost associated to the usage of NGS was 1205 € and corresponded to the second PCR and the MiSeq sequencing run. For comparison, sequencing the same PCR products (5 x 21 amplicons) using Sanger sequencing was estimated to cost 545 € (5.2 € per bidirectional read). However, targeting the same number of DNA fragments in 96 samples would become more cost-efficient with NGS (1330 € for 300 Mb output to 1900 € for 7 Gb output) than with Sanger (1955 €). A

more uniform molarity of the PCR products and a selection of the Illumina reagent kit in accordance with the number of samples processed can further improve this cost-efficiency. The labour cost was also superior here (1 person month) than for Sanger data analysis (0.5 person month) but the analysis pipeline developed here can be reused to analyse other projects. On the basis of these estimations, we expect the usage of NGS to be more cost-efficient when more than five markers (DNA fragments < 450 bp) have to be sequenced for more than 100 samples, particularly if several projects using the same approach are planned.

Phylogeny

The lack of resolution of the trees exclusively constructed with nuclear data prevented us to check the species delineation obtained with COI. In contrast, some deeper nodes were only resolved in the analyses combining COI and the three nuclear gene fragments (Figure 3). With this dataset, the two clades identified by morphology (Pauly et al. 2015), viz. (*H. phryganicus*, *H. smaragdulus*) and ((*H. orientanus*, *H. submediterraneus*) *H. gemmellus*), were supported in our phylogeny. The Halictidae comprises thousands of species that are often difficult to identify morphologically and whose taxonomy is regularly being refined using COI sequence data. Although COI data provide good support for most morphologically described halictid species (Schmidt et al. 2015), some groups like *Lasioglossum* (*Dialictus*) are more problematic (Gibbs 2017). It is therefore valuable to consider additional loci both for a better species delineation and for a better understanding of interspecific phylogenetic relationships (Danforth et al. 2011). Obviously, the set of loci analysed here was not useful for species delineation but it clarified the evolutionary history of the species studied.

Wolbachia infection

The detection of the wsp gene in more than one third of the specimens sampled reveals a high prevalence of *Wolbachia* in the group under study. Although *Wolbachia* infections were observed previously for the genus (Gerth et al. 2011), these are the first records for the *H. smaragdulus* species complex. In most infected individuals (5/8), two different HVR1 sequences were detected. This is also in agreement with previous studies revealing the co-occurrence of more than one *Wolbachia* sequence type in insects (Breeuwer et al. 1992; Mercot et al. 1995; Perrot-Minnot et al. 1996). We observed the same HVR1 sequence type in conspecific specimens (HVR1:11 in both *H. cephalicus* and both *H. smaragdulus* specimens and HVR1:51 in both *H. seladonia* specimens). This can result both from vertical and horizontal transmission of the bacteria. Data are insufficient to assess if the presence of *Wolbachia* has influenced the evolutionary history of the species complex (Baldo et al. 2005, 2006; Simões et al. 2011; Uday & Puttaraju 2012). Our results confirm that *Wolbachia* COI can be unintentionally sequenced with PCR primers that are routinely used in Metazoa (Smith et al. 2012) and that a parallel sequencing approach provides good quality results when different DNA fragments are co-amplified.

Detection of variant haplotypes

The average rates of substitution per base calculated for each assembly were within the expected range of sequencing error rates reported for amplicon sequencing with the Miseq Illumina platform (Schirmer et al. 2016). They were two orders of magnitude below the threshold of 10% used here to detect variants. Variant haplotypes observed with relative frequencies of 0.45 and 0.5 in two nuclear fragments (wnt1 and w) of one

279 specimen (*H. lucidipennis*) correspond to biallelic heterozygosity. The other variants
280 observed with a frequency of 0.11 at the end of the wnt1 reads more probably
281 correspond to sequencing errors. Indeed, the uneven distribution of sequencing errors
282 along sequencing reads can explain some more frequent sequencing errors (Schirmer et
283 al. 2016). Concerning COI, the reads obtained for three specimens (both specimens of
284 *H. seladonia* and *H. lucidipennis*) showed eight to 26 variant nucleotide characters (10-
285 23% of the reads). These variants are not cross-contaminants because they are different
286 from the COI haplotypes sequenced in the other individuals. They are also unlikely
287 numts or sequencing errors because most substitutions (49/51) are synonymous and
288 none are responsible for a stop codon. They are more probably due to heteroplasmy.
289 Heteroplasmy was already reported for Hawaiian *Hylaeus* (*Nesoprosopis*) Perkins 1899
290 (Magnacca & Brown 2010). These variant haplotypes did neither affect the phylogenies
291 nor the BLAST results because both species investigated here (*H. seladonia* and *H.*
292 *lucidipennis*), are relatively divergent from their closest known species. However, the
293 intraindividual divergences observed here (up to 2.7% and 4%) are in the range of
294 interspecific divergences in Halictidae (Pauly et al. 2015; Gibbs 2017) and could affect
295 results of DNA barcoding analyses involving closely related species (Magnacca &
296 Brown 2010). Detecting such variants is therefore essential in DNA barcoding.
297 Concerning the detection of numts, we did not observe stop codons or shifts in the
298 reading frame but we cannot totally exclude that nuclear copies were amplified
299 preferentially. In this regard, our approach does not offer more guaranties than Sanger
300 sequencing as it also relies on the PCR amplification of small DNA fragments and can
301 be biased by different amplification efficiencies. Sequencing the whole mitochondrial
302 genome represents a better solution to detect numts (Nelson et al. 2012).

Conclusion

The parallel sequencing of targeted amplicons, as applied here, can advantageously replace DNA barcoding in two cases: when a multilocus dataset has to be assembled for a considerable number of specimens and when variant haplotypes are expected in the sampling. Indeed our experiment was useful to construct a multilocus dataset consisting of DNA barcodes (COI) and three nuclear gene fragments with a cost efficiency that is estimated to become interesting compared to Sanger sequencing when more than 100 specimens are investigated. Our experiment also enabled the detection of variant COI haplotypes (with intraindividual divergences in the range of interspecific distances in Halictidae) and mixed sequence types of the intracellular bacteria *Wolbachia*. This relatively cheap application of NGS may therefore be useful in bee systematics, when these cases are encountered. **Acknowledgements**

This study was funded by the Belgian Science Policy (BELSPO) and supported by the FWO Research Community W0.009.11N ‘‘Belgian Network for DNA Barcoding” (BeBoL). Sequencing and library preparation was performed at the Genomics Core of KU Leuven (Belgium) with the help of Sigrun Jackmaert. We would like to thank the valuable suggestions of the two anonymous reviewers.

Authors’ contributions: GS, ZTN and TB conceived the research and wrote the article. GS and SW collected the data and performed the analyses. AP, ZTN, MV, KJ, JVH and MDM contributed to the interpretation of the data. All authors revised the text and the figures and approved the final manuscript.

References

Baldo, L., Hotopp, J.C.D., Jolley, K.A., Bordenstein, S.R., Biber, S.A., Choudhury, R.R., Hayashi, C., Maiden, M.C.J., Tettelin, H. & Werren, J.H. (2006). Multilocus sequence typing system for the endosymbiont Wolbachia pipientis. *Appl. Environ. Microbiol.*, 72, 7098–7110.

Baldo, L., Lo, N. & Werren, J.H. (2005). Mosaic nature of the Wolbachia surface protein. *J. Bacteriol.*, 187, 5406–5418.

Batovska, J., Cogan, N.O.I., Lynch, S.E. & Blacket, M.J. (2017). Using Next-Generation Sequencing for DNA Barcoding: Capturing Allelic Variation in ITS2. *G3: Genes|Genomes|Genetics*, 7, 19–29.

Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J. & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*

Breeuwer, J.A.J., Stouthamer, R., Barns, S.M., Pelletier, D.A., Weisburg, W.G. & Werren, J.H. (1992). Phylogeny of the cytoplasmic incompatibility microorganism in the parasitoid wasp of the genus $\backslash$emph{Nasonia} (Hymenoptera: Pteromalidae) based on 16S ribosomal DNA sequences. *Insect. Mol. Biol.*, 1, 25–36.

Breeuwer, J.A.J. & Werren, J.H. (1993). Cytoplasmic incompatibility and bacterial

density in Nasonia vitripennis. *Genetics*, 135, 565–574.

Buhay, J.E. (2009). “COI-like” sequences are becoming problematic in molecular systematic and DNA barcoding studies. *J. Crustac. Biol.*, 29, 96–110.

Bybee, S.M., Bracken-Grissom, H.D., Haynes, B.D., Hermansen, R.A., Byers, R.L., Clement, M.J., Udall, J. a, Wilcox, E.R. & Crandall, K. a. (2011). Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biol. Evol.*, 3, 1312–23.

Casiraghi, M., Bordenstein, S.R., Baldo, L., Lo, N., Beninati, T., Wernegreen, J.J., Werren, J.H. & Bandi, C. (2005). Phylogeny of Wolbachia pipientis based on gltA, groEL and ftsZ gene sequences: Clustering of arthropod and nematode symbionts in the F supergroup, and evidence for further diversity in the Wolbachia tree. *Microbiology*, 151, 4015–4022.

Criscuolo, A. & Brisse, S. (2013). AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*, 102, 500–506.

Cristiano, M.P., Fernandes-Salomão, T.M. & Yotoko, K.S.C. (2012). Nuclear mitochondrial DNA: an Achilles’ heel of molecular systematics, phylogenetics, and phylogeographic studies of stingless bees. *Apidologie*, 43, 527–538.

Danforth, B.N., Brady, S.G., Sipes, S.D. & Pearson, A. (2004). Single-copy nuclear genes recover Cretaceous-age divergences in bees. *Syst. Biol.*, 53, 309–326.

Danforth, B.N., Cardinal, S., Praz, C., Almeida, E. a. B. & Michez, D. (2011). The impact of molecular data on our understanding of bee phylogeny and evolution.

369 *Annu. Rev. Entomol.*, 58, 120830113030002.

370 Gerth, M., Geißler, A. & Bleidorn, C. (2011). infections in bees (Anthophila) and
371 possible implications for DNA barcoding. *Syst. Biodivers.*, 9, 319–327.

372 Gibbs, J. (2017). DNA barcoding a nightmare taxon: assessing barcode index numbers
373 and barcode gaps for sweat bees. *Genome*.

374 Gibbs, J., Brady, S.G., Kanda, K. & Danforth, B.N. (2012). Phylogeny of halictine bees
375 supports a shared origin of eusociality for Halictus and Lasioglossum (Apoidea:
376 Anthophila: Halictidae). *Mol. Phylogenet. Evol.*, 65, 926–939.

377 Hartig, G., Peters, R.S., Borner, J., Etzbauer, C., Misof, B. & Niehuis, O. (2012).
378 Oligonucleotide primers for targeted amplification of single-copy nuclear genes in
379 apocritan Hymenoptera. *PLoS One*, 7.

380 Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003). Biological
381 identifications through DNA barcodes. *Proc. R. Soc. London. Ser. B, Biol. Sci.*,
382 270, 313–321.

383 Hiroki, M., Tagami, Y., Miura, K. & Kato, Y. (2004). Multiple infection with
384 Wolbachia inducing different reproductive manipulations in the butterfly Eurema
385 hecabe. *Proc. Biol. Sci.*, 271, 1751–1755.

386 James, a C., Dean, M.D., McMahon, M.E. & Ballard, J.W.O. (2002). Dynamics of
387 double and single Wolbachia infections in Drosophila simulans from New
388 Caledonia. *Heredity (Edinb).*, 88, 182–189.

389 Jolley, K.A. & Maiden, M.C.J. (2010). BIGSdb: Scalable analysis of bacterial genome

variation at the population level. *BMC Bioinformatics*, 11, 595.

Kawakita, A., Ascher, J.S., Sota, T., Kato, M. & Roubik, D.W. (2008). Phylogenetic analysis of the corbiculate bee tribes based on 12 nuclear protein-coding genes (Hymenoptera : Apoidea : Apidae). *Apidologie*, 39, 163–175.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647–1649.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C. & Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.*, 14, 82.

Magnacca, K. & Brown, M. (2010). Mitochondrial heteroplasmy and DNA barcoding in Hawaiian Hylaeus (Nesoprosopis) bees (Hymenoptera: Colletidae). *BMC Evol. Biol.*, 10, 174.

Mercot, H., Llorente, B., Jacques, M., Atlan, A. & Montchamp-Moreau, C. (1995). Variability within the Seychelles cytoplasmic incompatibility system in Drosophila simulans. *Genetics*, 141, 1015–1023.

Michener, C.D. (2007). *The Bees of the World*. 2nd Editio. Johns Hopkins University Press, Baltimore.

Nelson, L. a, Lambkin, C.L., Batterham, P., Wallman, J.F., Dowton, M., Whiting, M.F., Yeates, D.K. & Cameron, S.L. (2012). Beyond barcoding: a mitochondrial genomics approach to molecular phylogenetics and diagnostics of blowflies

412 (Diptera: Calliphoridae). *Gene*, 511, 131–42.

413 Nixon, K.C. (1999). The parsimony ratchet, a new method for rapid parsimony analysis.
414 *Cladistics*.

415 O'Neill, E.M., Schwartz, R., Bullock, C.T., Williams, J.S., Shaffer, H.B., Aguilar-
416 Miguel, X., Parra-Olea, G. & Weisrock, D.W. (2013). Parallel tagged amplicon
417 sequencing reveals major lineages and phylogenetic structure in the North
418 American tiger salamander (Ambystoma tigrinum) species complex. *Mol. Ecol.*,
419 22, 111–129.

420 Paradis, E. (2010). pegas: an R package for population genetics with an integrated-
421 modular approach. *Bioinformatics*, 26, 419–20.

422 Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of Phylogenetics and
423 Evolution in R language. *Bioinformatics*, 20, 289–290.

424 Pauly, A., Devalez, J., Sonet, G., Nagy, Z.T. & Boevé, J.L. (2015). DNA barcoding and
425 male genital morphology reveal five new cryptic species in the West Palearctic bee
426 Seladonia smaragdula (Vachal, 1895) (Hymenoptera: Apoidea: Halictidae).
427 *Zootaxa*, 4034, 257–290.

428 Perrot-Minnot, M.J., Guo, L.R. & Werren, J.H. (1996). Single and double infections
429 with Wolbachia in the parasitic wasp Nasonia vitripennis: Effects on compatibility.
430 *Genetics*, 143, 961–972.

431 Pesenko, Y.A. (1999). Phylogeny and Classification of the Family Halictidae Revised
432 (Hymenoptera: Apoidea). *J. Kansas Entomol. Soc.*, 72, 104–123.

433 Pesenko, Y.A. (2004). The phylogeny and classification of the tribe Halictini with
434 special reference to the Halictus genus-group (Hymenoptera: Halictidae).
435 *Zoosystematica Ross.*, 13, 83–113.

436 Raychoudhury, R., Grillenberger, B.K., Gadau, J., Bijlsma, R., van de Zande, L.,
437 Werren, J.H. & Beukeboom, L.W. (2010). Phylogeography of Nasonia vitripennis
438 (Hymenoptera) indicates a mitochondrial-Wolbachia sweep in North America.
439 *Heredity (Edinb).*, 104, 318–326.

440 Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S.,
441 Larget, B., Liu, L., Suchard, M. a & Huelsenbeck, J.P. (2012). MrBayes 3.2:
442 efficient Bayesian phylogenetic inference and model choice across a large model
443 space. *Syst. Biol.*, 61, 539–42.

444 Ros, V.I.D., Fleming, V.M., Feil, E.J. & Breeuwer, J.A.J. (2009). How diverse is the
445 genus Wolbachia? Multiple-gene sequencing reveals a putatively new Wolbachia
446 supergroup recovered from spider mites (Acari: Tetranychidae). *Appl. Environ.*
447 *Microbiol.*, 75, 1036–1043.

448 Schirmer, M., D'Amore, R., Ijaz, U.Z., Hall, N. & Quince, C. (2016). Illumina error
449 profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC*
450 *Bioinformatics*, 17, 125.

451 Schliep, K.P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27, 592–
452 593.

453 Schmidt, S., Schmid-Egger, C., Morini??re, J., Haszprunar, G. & Hebert, P.D.N. (2015).
454 DNA barcoding largely supports 250 years of classical taxonomy: Identifications

455 for Central European bees (Hymenoptera, Apoidea partim). *Mol. Ecol. Resour.*, 15,
456 985–1000.

457 Shokralla, S., Gibson, J.F., Nikbakht, H., Janzen, D.H., Hallwachs, W. & Hajibabaei,
458 M. (2014). Next-generation DNA barcoding: using next-generation sequencing to
459 enhance and accelerate DNA barcode capture from single specimens. *Mol. Ecol.*
460 *Resour.*, n/a-n/a.

461 Simões, P.M., Mialdea, G., Reiss, D., Sagot, M.F. & Charlat, S. (2011). Wolbachia
462 detection: An assessment of standard PCR Protocols. *Mol. Ecol. Resour.*, 11, 567–
463 572.

464 Smith, M.A., Bertrand, C., Crosby, K., Eveleigh, E.S., Fernandez-Triana, J., Fisher,
465 B.L., Gibbs, J., Hajibabaei, M., Hallwachs, W., Hind, K., Hrcek, J., Huang, D.W.,
466 Janda, M., Janzen, D.H., Li, Y., Miller, S.E., Packer, L., Quicke, D.,
467 Ratnasingham, S., Rodriguez, J., Rougerie, R., Shaw, M.R., Sheffield, C., Stahlhut,
468 J.K., Steinke, D., Whitfield, J., Wood, M. & Zhou, X. (2012). Wolbachia and DNA
469 barcoding insects: Patterns, potential, and problems. *PLoS One*, 7.

470 Uday, J. & Puttaraju, H.P. (2012). Comparative analysis of Wolbachia surface protein.
471 *Bioinformation*, 8, 711–715.

472 Wilkinson, M.J., Szabo, C., Ford, C.S., Yarom, Y., Croxford, A.E., Camp, A. &
473 Gooding, P. (2017). Replacing Sanger with Next Generation Sequencing to
474 improve coverage and quality of reference DNA barcodes for plants. *Sci. Rep.*, 7,
475 46040.

476 Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. (2014). PEAR: A fast and accurate

Illumina Paired-End reAd mergeR. *Bioinformatics*, 30, 614–620.

Tables

Table I List of specimens and GenBank accession numbers, counts of reads and sequencing depth obtained for each gene fragment

Table II Characterization of variant nucleotide characters found with a relative frequency >10% in the assemblies used for phylogenetic analyses (COI, wnt1 and w). No variant >10% relative frequency was observed for HOG7036-02

Figures

Fig. 1 Number of reads obtained throughout the procedure

Fig. 2 Types of hypervariable region 1 (HVR1) identified in the *Wolbachia* surface protein (wsp) gene fragment surveyed in this study. Values indicate number of front end reads matching a HVR1 type

Fig. 3 Phylogeny of the *Halictus (Seladonia) smaragdulus* species complex inferred using a Bayesian analysis (BI) of the concatenated dataset including gene fragments of the cytochrome *c* oxidase subunit I (COI), wingless (wnt1), white-like (w) and hippo for putative serine/threonine kinase (HOG7036-02). Posterior probabilities >0.95 are given at nodes. Other analyses (BI, MP and NJ) based on COI or on the concatenations of three and four DNA fragments supported the same nodes with posterior probabilities >0.95 (BI) and bootstrap values >70% (MP) except for [A]: BI based on COI only, [B]: MP based on COI only, [C]: BI based on three genes (COI, HOG7036-02, w), [D]: MP based on three genes and NJ based on four genes and [E]: NJ based on COI only. Analyses exclusively based on nuclear data were much less resolved and are presented as Online

Resources 2. Specimens in which the *Wolbachia* surface protein gene was detected are tagged with a *. Scale bar represent expected changes per site. Wnt1 was not obtained for AP002, AP014, AP065 and AP083 and their position in the tree (represented with double lines) was inferred from the concatenation of the three other gene fragments

Supplementary material

Online Resource 1: Description of the DNA library preparation and data analysis.

Online Resource 2: Unrooted phylogenetic trees constructed with neighbour-joining (NJ) using nuclear data from each separated gene fragment (A-C) and their concatenation (D). Values at node represent the bootstrap values obtained with NJ and stars indicate nodes that are supported by parsimony (with bootstrap values >70) and Bayesian inference (with posterior probability >0.95).