Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 883–894 doi: 10.1111/2041-210X.12376

Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages

Carola Gómez-Rodríguez[1,2]*, Alex Crampton-Platt[1,3], Martijn J. T. N. Timmermans[1,4,5], Andrés Baselga[2] and Alfried P. Vogler[1,4]

[1]*Department of Life Sciences, Natural History Museum, London, SW7 5BD, UK;* [2]*Departamento de Zoología, Facultad de Biología, Universidad de Santiago de Compostela, c/Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, Spain;* [3]*Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, UK;* [4]*Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, SL5 7PY, UK; and* [5]*Department of Natural Sciences, Hendon Campus, Middlesex University, London, NW4 4BT, UK*

Summary

1. The biodiversity of mixed-species samples of arthropods can be characterized by shotgun sequencing of bulk genomic DNA and subsequent bioinformatics assembly of mitochondrial genomes. Here, we tested the power of mitochondrial metagenomics by conducting Illumina sequencing on mixtures of >2600 individuals of leaf beetles (Chrysomelidae) from 10 communities.

2. Patterns of species richness, community dissimilarity and biomass were assessed from matches of reads against three reference databases, including (i) a custom set of mitogenomes generated for 156 species (89% of species in the study); (ii) mitogenomes obtained by the *de novo* assembly of sequence reads from the real-world communities; and (iii) a custom set of DNA barcode (*cox1*-5′) sequences.

3. Species detection against the custom-built reference genomes was very high (>90%). False presences were rare against mitogenomes but slightly higher against the barcode references. False absences were mainly due to the incompleteness of the reference databases and, thus, more prevalent in the *de novo* data set. Biomass (abundance × body length) and read numbers were strongly correlated, demonstrating the potential of mitochondrial metagenomics for studies of species abundance.

4. A phylogenetic tree from the mitogenomes showed high congruence with known relationships in Chrysomelidae. Patterns of taxonomic and phylogenetic dissimilarity between sites were highly consistent with data from morphological identifications.

5. The power of mitochondrial metagenomics results from the possibility of rapid assembly of mitogenomes from mixtures of specimens and the use of read counts for accurate estimates of key parameters of biodiversity directly from community samples.

Key-words: beta diversity, biodiversity monitoring, Chrysomelidae, Coleoptera, genome skimming, mitometagenomics, next-generation sequencing, phylobeta

Introduction

The lack of rapid, inexpensive and accurate species identification has been a major limitation for the study of biodiversity and ecosystems, in particular for small-bodied, species-rich groups of animals that constitute most of the known diversity on Earth (Wilson 1988). While specimens can be gathered readily using standardized collecting methods, their further study is frequently impossible due to insufficient taxonomic expertise (the so-called Linnaean shortfall; Brown & Lomolino 1998; Whittaker *et al.* 2005). Identification with molecular markers via DNA barcoding (Hebert, Ratnasingham & deWaard 2003) equally requires a high input of resources (Stein *et al.* 2014). Even greater efforts may be necessary in ecological and evolutionary studies that require information on other facets of biodiversity, such as species abundance or phylogenetic relationships. However, high-throughput sequencing technologies (HTS) offer new avenues for studying complex assemblages (Baird & Hajibabaei 2012; Yu *et al.* 2012; Bohmann *et al.* 2014) and have the potential to go beyond basic species lists, for an integrative approach to biodiversity science (Cristescu 2014).

Recent proposals for HTS for biodiversity studies have used PCR-based 'metabarcoding' for characterizing mixed-specimens samples or environmental samples (Shokralla *et al.* 2012; Gibson *et al.* 2014), but suffer from various shortcom-

*Correspondence author. E-mail: carola.gomez@usc.es

ings, such as overestimation of species richness due to the amplification of nuclear mitochondrial pseudogenes (numts) (Song *et al*. 2008) or underestimation due to the difficulty of amplification with universal primers covering a broad taxonomic range (Deagle *et al*. 2014). Alternatively, PCR-free methods have focused on the assembly of mitochondrial genomes from shotgun sequencing of specimen mixtures (Tang *et al*. 2015; C. Andújar, P. Arribas, F. Ruzicka, A. Crampton-Platt, M.J.T.N. Timmermans, A.P. Vogler, under review; A. Crampton-Platt, M.J.T.N Timmermans, M.L. Gimmel, S.N. Kutty, T.D. Cockerill, C.V. Khen, A.P. Vogler, under review). Mitogenomes are readily amenable to 'genome skimming', by which the high-copy portion of the genome is assembled from libraries sequenced at low depth (Straub *et al*. 2012). When applied to mixtures of total DNA from multiple species, dozens or even hundreds of mitogenomes can be assembled from raw reads (Zhou *et al*. 2013; Gillett *et al*. 2014; A. Crampton-Platt, M.J.T.N Timmermans, M.L. Gimmel, S.N. Kutty, T.D. Cockerill, C.V. Khen, A.P. Vogler, under review). The resulting contigs can be exploited for the well-established uses of mtDNA in species identification (Hebert, Ratnasingham & deWaard 2003), phylogenetics (Cameron 2014) and phylogeography (Avise 2000), whose conclusions are strengthened with the greater amount of data from whole mitochondrial genomes (e.g. Ma *et al*. 2012; Cameron 2014; Gillett *et al*. 2014). Besides, shotgun sequencing approaches would also provide better estimation of species richness in the genomic mixture as the recovery of (single-copy) numts would be suppressed, except perhaps for some genomes with very high loads of recent numts (Bensasson *et al*. 2001). In addition, as no PCR-induced biases are introduced, direct sequencing of biodiversity samples may inform about species abundance (Taberlet *et al*. 2012). Hence, the same biotic sample could be analysed to obtain insights simultaneously on species composition, abundance, genetic variation and evolutionary relationships.

Before mitochondrial metagenomics is widely used for studying unknown biodiversity, key parameters have to be examined carefully. A distinction is needed between 'read-based' analyses that establish the occurrence of taxa by matching sequence reads against a known reference data, such as DNA barcodes or full mitochondrial genomes, and 'contig-based' analyses that aim at the *de novo* assembly of mitogenomes from the mixture of sequence reads. HTS studies to date have usually followed the read-based approach by comparing sequences (usually PCR amplicons) with existing reference libraries, such as the BOLD or SILVA databases (Shokralla *et al*. 2012; Gibson *et al*. 2014). The success of identifications largely depends on the completeness and taxonomic resolution of such databases and their representation of loci. The alternative 'contig-based' approach remains less well-established and so far has been explored only on mixtures of distantly related individuals in particular for phylogenetics (Gillett *et al*. 2014) and for proof-of-principle studies of contrived communities (Zhou *et al*. 2013; Tang *et al*. 2015).

The utility of mitochondrial metagenomics in biodiversity surveys will depend on the degree to which inferred species occurrences are reliable and provide consistent measures of within and between sample diversity. Any errors with species detection would then lead to incorrect conclusions about alpha and beta diversity and the ecological correlates of species diversity that are frequently the focus of biodiversity studies (e.g. Yu *et al*. 2012). Therefore, before mitochondrial metagenomics can be applied to complex specimen mixtures, it is required to explore the performance of this approach for species detection, including the relevance of potential factors that may cause incorrect inferences about species occurrences. For example, the outcome of pooled genome sequencing could be affected by the assembly process itself and the possibility of chimerical contigs or the blanket failure of contig formation in the face of genetic variation, in particular if close relatives are present in the specimen pool.

To test the performance of mitochondrial metagenomics, a controlled analysis is needed against existing data obtained with conventional methods of species inventories. Here, we formalize an approach that combines read-based and contig-based uses of HTS to generate fast and cost-effective reference libraries of mitochondrial genomes against which raw reads from low-coverage sequencing of full communities can be compared, targeting the full mitochondrial genome. We use samples of known specimens previously used to assess the diversity of communities of leaf beetles (Chrysomelidae) in the Iberian Peninsula that have already been studied using morphological identification and DNA barcoding (Baselga, Gómez-Rodríguez & Vogler, 2015). Ten local assemblages with altogether >2600 specimens of 171 species were subjected to shotgun metagenomic sequencing and *de novo* assembly to generate a set of mitochondrial contigs. In parallel, we built a nearly complete reference library of mitochondrial genomes by pooled shotgun sequencing of one representative for each species under uniform DNA concentrations. The 10 communities were then assessed for species richness, abundance and community dissimilarity patterns with the read-based matching against (i) PCR-barcodes from Baselga, Gómez-Rodríguez & Vogler (2015), (ii) the custom-built mitochondrial reference genomes and (iii) *de novo* assembled mitochondrial genomes from the shotgun sequencing of the bulk community samples. The latter reference set is of particular interest because it can be generated directly from the studied community without the requirement for any external sequencing resources, such as a PCR-barcodes database or a purpose-built database of mitochondrial genomes.

Materials and methods

DNA SAMPLES, PCR-BARCODES LIBRARY AND PCR-DERIVED PHYLOGENETIC TREES

DNA samples and PCR-barcodes (*cox1-5′*) come from communities of leaf beetles (Chrysomelidae) that had been sampled using a standardized protocol (beating of vegetation for 20 sampling periods of 30 min) along a north–south transect of the Iberian Peninsula (Baselga, Gómez-Rodríguez & Vogler, 2015). Thus, the present study is based on 2607 morphologically identified specimens from 171 species and 10 collecting sites (from a total of 4533 specimens and 202 species from 20

sites in the original study, Appendix S1). In the following, the existing barcode database will be referred to as 'PCR-barcodes'. Linnaean species and barcode-derived entities were highly congruent, except for two pairs of sister species (Baselga *et al.* 2013). Phylogenetic relationships of these species were inferred from the barcode fragment (Appendix S2a) and from a 4-loci tree (*cox1*-3′, *cox1*-5′, *rrnL* and 18S) based on newly generated data (Appendix S2b).

LIBRARY PREPARATION FOR HTS

Illumina TruSeq libraries were prepared from pooled DNA extracts of single individuals that had been stored at −80°C for 3 years. First, a 'mitochondrial reference library' ('MitoRL') was generated from a single representative of each species in the community set (171 + 5 species from an adjacent locality, see Appendix S1). DNA concentrations were adjusted for uniform read coverage of species in the pool by adding different volumes of DNA extract according to four size classes ('large' = 0·67 μL, 'intermediate' = 2 μL, 'small' = 5 μL and 'very small' = 50 μL; Appendix S3). Secondly, to profile species assemblages in each sampling locality, 10 separate libraries were constructed by pooling DNA extracts from every individual found at that locality at an equal volume of 3 μL per specimen, hereafter called 'local assemblage libraries' ('LocL'). DNA pooling aimed to mimic mass DNA extraction of specimens in bulk. These pools included DNA from 156 to 336 specimens (found at locality HOR and LAS, respectively), representing 27 (HOR) to 67 (ANC) species and between 99 (HOR) and 156 (ADS) *cox1*-5′ haplotypes (Table 1 and Appendix S1).

For the MitoRL, a TruSeq PCR-free library with an insert size of 550 bp was prepared. For each LocL, the DNA mixture was sheared for a TruSeq library aiming an insert size of 800 bp. Sequencing was performed on an Illumina MiSeq sequencer for 2 × 300 bp paired-end reads (MitoRL: 600 cycles, reagent kit version 3) or 2 × 250 bp paired-end reads (LocL samples: 500 cycles, reagent kit version 2). A 117% flow cell on the MiSeq was used for the MitoRL, while each LocL was sequenced on 20% of a MiSeq flow cell (Appendix S3).

ASSEMBLY OF MITOGENOMES, REFERENCE LIBRARIES CONSTRUCTION AND LEAF BEETLE PHYLOGENY

The assembly of mitochondrial contigs followed A. Crampton-Platt, M.J.T.N Timmermans, M.L. Gimmel, S.N. Kutty, T.D. Cockerill, C.V. Khen, A.P. Vogler (under review). In brief, trimmed reads were reduced to putative mitochondrial reads using BLAST (e-value $<10^{-5}$) against a database of mitochondrial genomes of Coleoptera (M.J.T.N. Timmermans, C. Barton, J. Haran, D. Ahrens, L. Culverwell, S. Dodsworth, P.G. Foster, L. Bocak, A.P. Vogler, under review). Mitochondrial reads were *de novo* assembled with NEWBLER 2.7 (identity in overlap = 99%; minimum overlap length = 150 bp) and IDBA-UD 1.1.1 (maximum k-value = 250 bp and minimum k-value = 80 bp). Reads used for IDBA-UD assembly were trimmed to 250 bp. Assembled contigs were filtered again to remove non-mitochondrial DNA contigs using the same procedure as above. Identical or very similar contigs produced by different assemblers were merged with the 'De Novo Assembly' function in GENEIOUS 5.6 (minimum overlap = 500 bp; minimum over-

Table 1. Known species richness and estimates from various high-throughput sequencing (HTS)-based profiling methods. Shown are richness estimates, number of true presences (TP), false presences (FP), false absences (FA), true absences (TA) and profiling success (PS = TP + TA/TP + FP + FA + TA). False absences strictly due to failure of the profiling are differentiated from total number of false absences that includes those that remain undetected due to the incompleteness of the reference library (given in brackets). This affects 24 species missing from the MitoRL, six species from PCR-barcodes and 96 species from DeNovoRL database

	Morpho Richness	MitoRL						PCR-barcodes					
		Richness	PS [%]	TP	TA	FP	FA	Richness	PS [%]	TP	TA	FP	FA
ADS	41	37	93·1	33	129	4	2 (8)	39	94·2	35	127	4	6 (6)
ANC	67	57	92·0	55	105	2	1 (12)	58	93·6	57	104	1	8 (10)
EUM	41	36	94·8	34	131	2	2 (7)	44	94·8	38	125	6	2 (3)
HOR	27	26	97·1	24	145	2	1 (3)	25	97·7	24	144	1	3 (3)
JCB	36	36	96·6	33	135	3	0 (3)	35	97·1	33	134	2	3 (3)
LAS	56	50	92·0	46	114	4	3 (10)	48	91·9	45	113	3	10 (11)
MAC	49	43	95·4	42	124	1	0 (7)	50	95·9	46	119	4	2 (3)
OMA	45	41	94·3	38	126	3	0 (7)	45	95·3	41	123	4	3 (4)
SAN	47	47	95·4	43	123	4	0 (4)	45	96·5	43	123	2	4 (4)
TUE	48	39	92·5	37	124	2	3 (11)	39	94·8	39	124	0	8 (9)

	DeNovoRL					
	Richness	PS [%]	TP	TA	FP	FA
ADS	26	88·3	24	127	2	3 (18)
ANC	38	81·9	37	103	1	2 (30)
EUM	26	89·5	24	129	2	1 (16)
HOR	21	94·7	20	142	1	1 (8)
JCB	27	92·4	25	133	2	0 (11)
LAS	35	83·6	32	111	3	2 (25)
MAC	37	87·7	33	117	4	2 (17)
OMA	26	87·1	25	124	1	2 (21)
SAN	36	87·1	31	118	5	2 (17)
TUE	28	86·0	26	121	2	1 (22)

lap identity = 99%). Contigs of <3000 bp in length or with fewer than two full protein-coding genes were removed because of their presumed low phylogenetic information content. Contigs >15 000 bp were checked for circularity, that is the presence of identical nucleotide sequence on both ends of the contigs, indicating that the mitogenome was sequenced in full. Gene annotations followed A. Crampton-Platt, M.J.T.N Timmermans, M.L. Gimmel, S.N. Kutty, T.D. Cockerill, C.V. Khen, A.P. Vogler (under review) using annotations of tRNA genes to delimit protein-coding regions that were mapped against the *Crioceris duodecimpunctata* mitochondrion (NC_003372). Fragments representing >50% of a given gene were aligned with transAlign (Bininda-Emonds 2005).

For the MitoRL reference set, multiple non-overlapping contigs might represent a single mitogenome ('sibling contigs'), which were identified based on close phylogenetic relationships and identical geographic distributions inferred with the assemblage profiling method explained below (Appendix S4). Taxonomic assignment of contigs was by matches to the PCR-barcodes database (Baselga, Gómez-Rodríguez & Vogler, 2015) with BLAST (e-value = 10^{-5}; identity ≥98%; blast length ≥250 bp). The position in the phylogenetic tree was used to assign names for a small number of species lacking the *cox1*-5′.

Local assemblage libraries were assembled individually ('LocL #1–#10') and combined for all 10 libraries ('CombL'). Contigs from the LocL and CombL libraries were combined into the '*de novo* reference library' ('DeNovoRL'), but due to the complexities of incompletely assembled contigs within and among libraries, only those contigs were retained that included the region of *nad2*, *cox1* and *cox2* (minimum length of 3234 bp, plus any length on either side of this region), to ensure each contig corresponds to a different mitogenome (Appendix S5). Closely related contigs were assessed with the generalized mixed Yule coalescent (GMYC, Pons *et al.* 2006) method for species delimitation (*gmyc* function, *splits* package) on a trimmed matrix (only *nad2*, *cox1* and *cox2*). Multiple representatives of a GMYC group were retained in the DeNovoRL.

For the MitoRL and DeNovoRL reference libraries, ANOVAS were used to evaluate whether the success of species recovery was related to the species average body length, the degree of phylogenetic relatedness in the sample (defined by the number of congeneric species), the intraspecific variability in the combined sample (defined by the number of haplotypes based on sanger-derived *cox1*-5′ [from the PCR-barcodes database], only for DeNovoRL) or the abundance (only for DeNovoRL). All variables were natural-log-transformed. Average body length was extracted from various literature sources or measured from the authors' specimens (see Appendix S3).

In order to build the most comprehensive phylogeny of Chrysomelidae based on mitogenomes, the MitoRL database was complemented by sequences from DeNovoRL for taxa with more complete data in the latter (Appendix S6). This final data set was used to build a maximum-likelihood bootstrap tree (rapid bootstrapping with GTRCAT model, samplings = 100). Trees were constructed with RAxML-HPC2 in Cipres (Miller, Pfeiffer & Schwartz 2010), using a GTRGAMMA model, partitioning by gene and with *Psacothea hilaris* (Cerambycidae) as outgroup.

ASSESSMENT OF CHIMERA FORMATION

As a test system for the error rates in *de novo* assembly of bulk DNA samples, we used contigs in the DeNovoRL and compared them against the complete circular genomes (n = 85) assembled in the MitoRL. Closed-circular genomes showing an exact match of overlapping sequences at both ends of the contig were assumed to be non-chimeric. NEWBLER and IDBA-UD contigs >1000 bp were divided into 200-bp fragments, and for each fragment, a BLAST search (e-value = 10^{-5}) was conducted against the full mitochondrial genomes in the MitoRL. A contig was considered to be chimeric when the fragments had their best BLAST match with two or more different circular reference genomes, indicating that the assembly produced heterospecific contigs.

SPECIES ASSEMBLAGE PROFILING

A presence/absence matrix was built by matching the reads from each local assemblage library (LocL) against the three reference libraries (MitoRL, PCR-barcodes, DeNovoRL) using BLAST (e-value = 10^{-10}, identity ≥98%; blast length ≥150 bp). For hits against MitoRL and DeNovoRL, only protein-coding genes were considered and a stringent criterion for species detection in a locality was applied, requiring that the number of reads for a given species at a site was ≥1% of the total number of reads for that species across all sites. The success of the assemblage profiling was evaluated with a contingency table summarizing the number of true and false presences and absences against the known distribution from morphological counts. The profiling success (PS) was calculated for each local assemblage as PS = TP + TA/(TP + FP + FA + TA), where TP and FP are the number of true and false presences, and TA and FA are true and false absences.

The number of BLAST hits was also used to test whether species biomass is correlated with the number of reads recovered for the species. Biomass was estimated as the product of number of individuals and species body length. The number of BLAST hits was corrected for the length of each mitogenome fragment (protein-coding genes) in the MitoRL and DeNovoRL, and Pearson correlations on the natural-log-transformed values were performed.

ASSESSMENT OF DIVERSITY PATTERNS

The correspondence between alpha diversity estimates based on read profiling and morphological species was assessed with a Pearson correlation for each locality. Congruence in patterns of assemblage dissimilarity (beta diversity) was assessed based on the overall heterogeneity in assemblage composition (multiple-site dissimilarity) and the pairwise pattern of assemblage differences (pairwise dissimilarities). We considered both taxonomic (occurrence and abundance-based) and phylogenetic dissimilarity among localities and differentiated the turnover (βsim) and nestedness-resultant (βsne) components following the framework of Baselga (2010), which was extended to phylogenetic dissimilarity by Leprieur *et al.* (2012) and to differences in species abundance by Baselga (2013). Multiple-site dissimilarity was computed with functions beta.multi or beta.phylo.multi, and pairwise dissimilarity matrices were computed with functions beta.pair (presence/absence data), bray.part (abundance data) or beta.phylo.pair in the betapart package (Baselga & Orme 2012). For analyses of phylogenetic dissimilarity (phylobetadiversity), different phylogenetic trees were used for congruence with the respective reference library used in assemblage profiling: the tree generated from mitochondrial genomes, the barcodes tree (Appendix S2b) and a pruned DeNovoRL tree, with a single tip per GMYC. The 4-loci tree derived by PCR (Appendix S2a) was used for analyses based on morphological species. At each level of analysis, Mantel tests were used to assess the congruence between the dissimilarity matrices based on morphological species and the assemblage profiles based on sequence reads.

Results

MITOCHONDRIAL GENOMES ASSEMBLY AND REFERENCE LIBRARIES BUILDING

The mitogenome reference library (MitoRL) included $25{\cdot}4*10^6$ reads, of which 8·23% were putative mitochondrial reads. A total of 179 contigs matched the minimal length criteria after merging contigs produced by both assemblers, of which 85 contigs were circular mitogenomes and 101 contigs contained all protein-coding genes (see Appendices S7 and S8 for details on assembly results). Further merging of incomplete mitochondrial genomes (sibling contigs) (Appendix S4) resulted in a final composition of the MitoRL of 156 putative mitogenomes that matched known morphological species in all except three cases (Appendix S4). We did not observe differences in the average body length (ANOVA $F_{1,169} = 0{\cdot}032$, $P = 0{\cdot}857$) or number of species in the genus (ANOVA $F_{1,169} = 2{\cdot}75$, $P = 0{\cdot}099$) between the species recovered in the MitoRL library and those that were not (Appendix S9).

Individual libraries (LocL#1–#10) were generated from all specimens encountered at each local assemblage and were composed of around $4*10^6$ reads each, for a combined total of $40{\cdot}2*10^6$ reads, of which only $1{\cdot}94*10^6$ (4·8%) matched mitochondrial genomes. A total of 270 contigs above the minimal length criteria were produced, including 46 circular genomes and 97 genomes containing all protein-coding genes (see assembly details in Appendices S7 and S8). The combined assembly of all 10 libraries (CombL) produced 141 genomes and altogether 411 contigs were obtained from LocL#1–#10 and CombL (Appendix S7). A total of 211 contigs (146 contigs from LocLs and 65 from CombL) were included in a trimmed *cox1*-centred matrix to delimit putative species with the GMYC method. Contigs were clustered into 71 GMYC groups, corresponding to 75 morphological species (Appendix S5). The species missing from the *de novo* assembly were significantly less abundant in the sampled assemblages (ANOVA $F_{1,169} = 81{\cdot}48$, $P < 0{\cdot}001$), had shorter average body length (ANOVA $F_{1,169} = 32{\cdot}972$, $P < 0{\cdot}001$) and harboured less *cox1*-5′ haplotypes ($F_{1,160} = 45{\cdot}123$, $P < 0{\cdot}001$), but there was no significant difference in the number of species per genus included in the sample (ANOVA $F_{1,169} = 2{\cdot}69$, $P = 0{\cdot}102$; Appendix S9).

CHIMERIC CONTIGS ASSESSMENT

Chimera formation in the *de novo* assembly of bulk DNA samples (DeNovoRL) against the *bona fide* full circular genomes ($n = 85$) of the MitoRL assembly (see Materials and methods) resulted in detection of 10 chimeras. Affected contigs were of any length (1218–16 312 bp), they were generated with either assembler, they were obtained by assembling local libraries separately and combined, and crossover points were in protein-coding, rRNA and tRNA genes and in the control region (details in Appendix S10). Chimeric contigs represented 0·31% (LocL) and 0·28% (CombL) of the data. If only long contigs >3000 bp are considered, the proportion of chimeras was larger (0·5% and 1·0%, respectively).

LEAF BEETLE PHYLOGENY BASED ON MITOCHONDRIAL GENOMES

The leaf beetle tree consisted of 171 terminals after contigs from DeNovoRL complemented the MitoRL and sibling contigs were collapsed (Appendix S6). The data set represented 161 morphological species, including six sets of ambiguous sibling contigs (i.e. only present in one locality) that were not concatenated. The tree recovered three major clades within Chrysomelidae (Fig. 1, Appendix S11), including the 'eumolpine clade' composed of Cassidinae, Eumolpinae and Cryptocephalinae (including Clytrini), the 'sagrine clade' represented by Donaciinae and Criocerinae, and the 'chrysomeline clade' represented by Chrysomelinae and Galerucinae (including Alticini).

SPECIES ASSEMBLAGE PROFILING FROM SEQUENCE READS

We inventoried putative species present at each locality based on the reads from each local library against the various reference libraries. Starting with MitoRL, estimated species richness at a local site ranged between 26 at HOR against 27 known species (richness estimation success = 96·3%; PS = 97·1%, i.e. the recovery of species given false absences and false presences) and 57 at ANC against 61 known species (richness estimation success = 91·8%; PS = 92·0%) (Table 1). Most false absences were attributable to the incompleteness of the MitoRL, as no mitogenome was available for 24 species (14% of the total). When the species without reference mitogenome were removed from the calculations, the mean PS rate across the 10 local assemblages was 97·7% ± 1·0 (SD). The remaining false absences ($n = 12$) affected rare species in six genera (Appendix S12), most of which were represented by a single specimen ($n = 7$). This did not prevent a large number of true presences being observed for singleton specimens across all 10 assemblies ($n = 96$; Appendix S13). False presences ($n = 27$) mainly involved species of the genera *Altica* and *Pachybrachis*, and the three contigs that could not be taxonomically assigned and therefore do not add to the count of true presences (Appendix S12).

Conducting the same analysis against the much shorter fragments of the PCR-barcodes library, 1·08% of the mitochondrial reads produced successful BLAST hits. A total of 150 species were recovered (87·7% of species present in the sampling localities), with local richness ranging from 25 (HOR) to 58 (ANC) and the PS rate between 91·9% (LAS) and 97·7% (HOR) (Table 1). False absences were higher ($n = 56$) and were not mostly attributable to the incompleteness of the database, since barcodes were missing for only six species. False absences strictly resulting from incorrect profiling affected 49 species in 19 genera (Table 1, Appendix S12). False presences ($n = 27$) occurred in 11 genera, mainly in *Altica*, *Calomicrus*, *Cryptocephalus* and *Lachnaia* (Appendix S12).

The success of species detection against the DeNovoRL was lower than for the other libraries. The estimated species

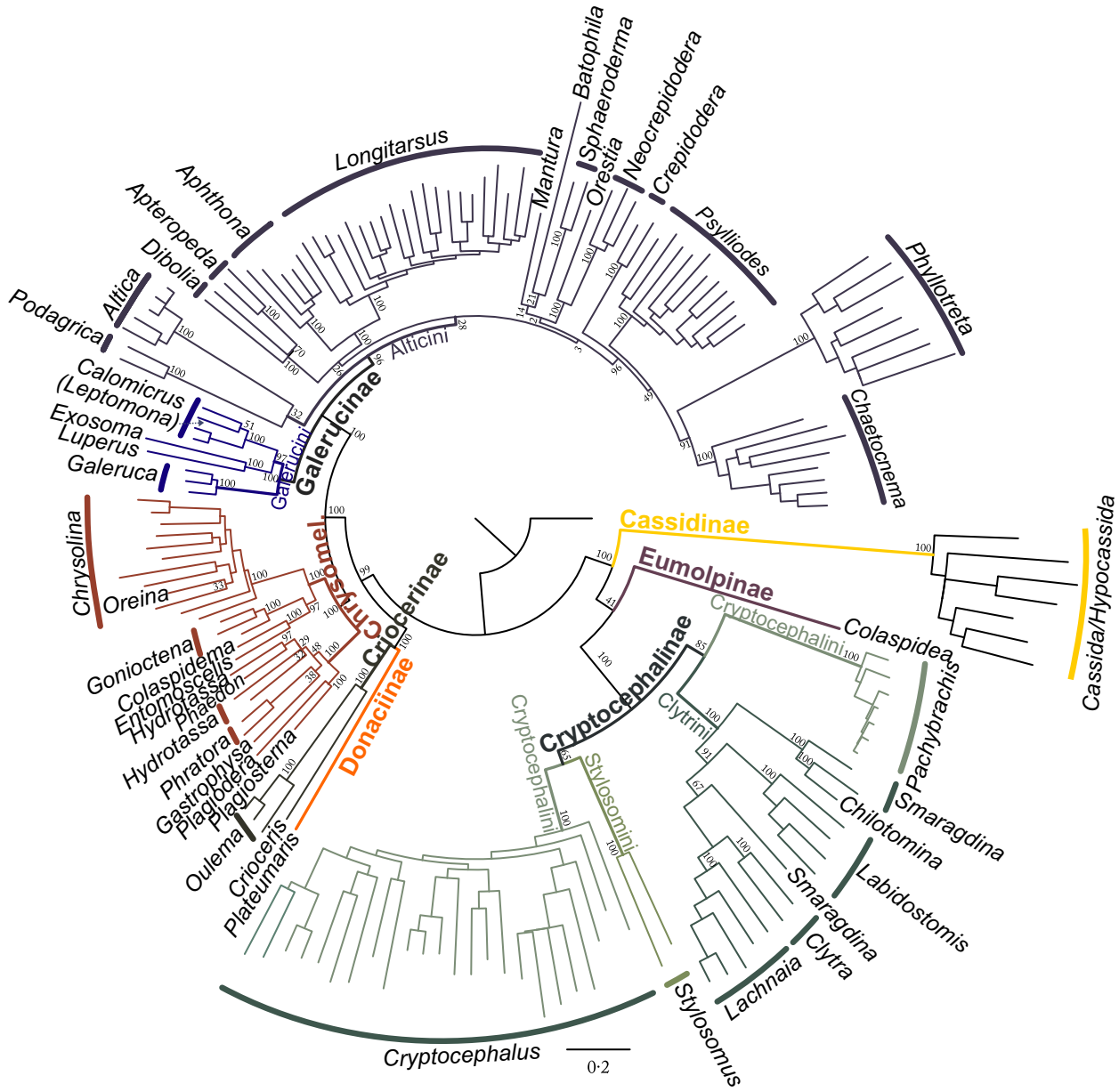

Fig. 1. Maximum-likelihood tree of leaf beetles evolutionary relationships based on mitochondrial genomes. Subfamilies and main tribes are indicated. Node values are bootstrap support values.

richness ranged from 21 (HOR, richness estimation success = 77·8%; PS = 94·7%) to 38 species (ANC, richness estimation success = 56·7%; PS = 81·9%) (Table 1), whereby most false absences were due to the incompleteness of the DeNovoRL, as only 43·8% of morphological species were recovered as GMYC groups. Only 10 false absences were attributable to erroneous BLAST hits (in genera *Altica*, *Hispa*, *Galeruca* and *Psylliodes*), while six cases were associated with the split of a morphological species (*Chrysolina quadrigemina*) into two GMYC groups (Appendix S12). All false presences (n = 23) except one were associated with the three cases where several morphological species had been lumped into a single GMYC group (Appendix S12).

DETERMINING SPECIES ABUNDANCE FROM SEQUENCE READS

Read counts matched to mitogenomes in the MitoRL were highly correlated with the species biomass in all local assemblages combined (Pearson R = 0·80; R^2 = 0·64; P < 0·001; Fig. 2) and in each local assemblage (0·56 [EUM] < Pearson R < 0·86 [OMA]; Table 2). The total number of reads producing BLAST hits for a given species ranged from 3 (*Aphthona euphorbiae*, two specimens, average length = 1·7 mm) to 28 106 (*Lachnaia pubescens*, 57 specimens, average length = 10·5 mm). Read counts for each species were also highly correlated with the species biomass when matched to the PCR-barcodes library, both for the combined assemblage

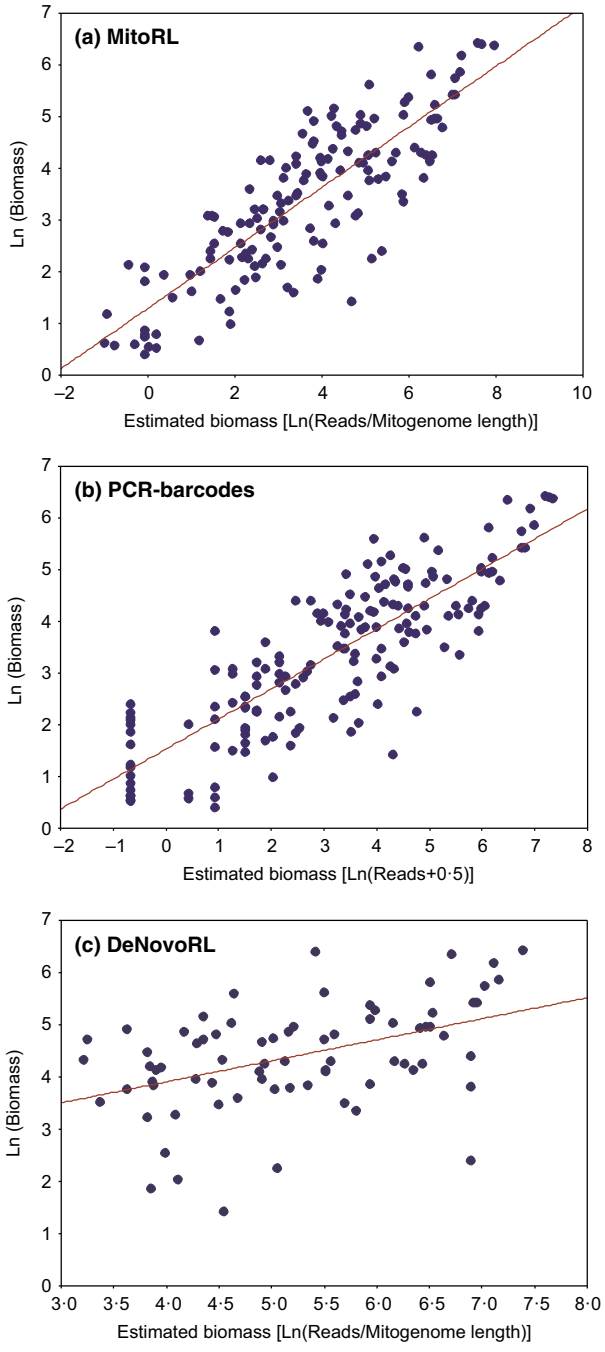

Fig. 2. Scatterplots of the relationship between HTS-based estimated relative biomass and true species biomass in the combined sample. Relative biomass was estimated with three different HTS-based profiling methods. HTS, high-throughput sequencing.

(Pearson $R = 0{\cdot}83$; $R^2 = 0{\cdot}69$; $P < 0{\cdot}001$; Fig. 2) and in each local assemblage (0·54 [EUM] < Pearson $R < 0{\cdot}81$ [OMA], see Table 2). The number of read hits ranged from 1 (*Chaetocnema hortensis, Longitarsus ibericus* and *Oulema rufocyanea*; all with average length ≤3·8 mm and 1 or 2 specimens) to 1516 (*L. pubescens*). Finally, for those species recovered in the DeNovoRL, read counts were correlated with the species biomass in the combined sample (Pearson $R = 0{\cdot}45$; $R^2 = 0{\cdot}20$; $P < 0{\cdot}001$; Fig. 2) but yielded mixed results in the local assemblages (0·35, non-significant [JCB] < Pearson $R < 0{\cdot}82$ [OMA]; Table 2). The total number of read hits for a species ranged from 127 (*Hydrothassa fairmairei*, 10 specimens, average length = 4·4 mm) to 17 858 (*Gonioctena olivacea*, 140 specimens, average length = 4·5 mm).

ASSESSMENT OF DIVERSITY PATTERNS

Correlations between known alpha diversity and HTS-estimated alpha diversity were high and significant for all profiling methods: MitoRL (Pearson $R = 0{\cdot}96$, $P < 0{\cdot}001$), PCR-barcodes (Pearson $R = 0{\cdot}93$, $P < 0{\cdot}001$) and DeNovoRL (Pearson $R = 0{\cdot}84$, $P = 0{\cdot}003$). When comparing observed and HTS-estimated values of dissimilarity, very similar values of multiple-site compositional heterogeneity (Table 3) and high correlations between pairwise matrices (Fig. 3) were observed for the MitoRL and PCR-barcodes libraries, especially for total dissimilarity (β_{SOR}, βsor) and its turnover component (β_{SIM}, βsim). The DeNovoRL exhibited slightly different values of multiple-site heterogeneity and, in the case of pairwise matrices, the lowest correlation in the nestedness component (Pearson $R = 0{\cdot}59$). Phylogenetic dissimilarity was correlated with species dissimilarity (Fig. 3), although both multiple-site values and pairwise dissimilarity values were, in general, lower for phylogenetic dissimilarity analyses (Table 3 and Fig. 3).

Congruent dissimilarity patterns (Pearson $R > 0{\cdot}89$) were also observed in the comparison between the known biomass and the biomass inferred from BLAST hits, except in the case of dissimilarity derived from unidirectional abundance gradients ($d_{BC\text{-}gra}$; 0·23 < Pearson $R < 0{\cdot}38$), which was non-significant in the DeNovoRL profiling (Fig. 4).

Discussion

We show how the recently proposed approach of shotgun sequencing and mitogenome assembly (Zhou *et al.* 2013; Gillett *et al.* 2014; Tang *et al.* 2015; C. Andújar, P. Arribas, F. Ruzicka, A. Crampton-Platt, M.J.T.N. Timmermans, A.P. Vogler, under review; A. Crampton-Platt, M.J.T.N Timmermans, M.L. Gimmel, S.N. Kutty, T.D. Cockerill, C.V. Khen, A.P. Vogler, under review) produces solid estimates of assemblage composition, species abundances and evolutionary relationships. We differentiate two ways of inventorying communities, using either external reference libraries (MitoRL or PCR-barcodes) or *de novo* assembly of mitogenomes from the local assemblages themselves (DeNovoRL). Where external data for the target taxa are not available, building a reference database is a key step. Bulk sequencing from selected species representatives at roughly equal DNA concentration made this step straightforward and recovered 89% of species. The long mitogenomes also provided robust phylogenetic trees and recovered the three major clades of Chrysomelidae first recognized by Gómez-Zurita, Hunt & Vogler (2008), with high support (PP = 100). Within these clades, the tree topology is mostly concordant with Gómez-Zurita, Hunt & Vogler (2008), in spite of the skewed local taxon sampling relative to the world-wide representation of that study. The strong

Table 2. Composition of local assemblages (total number of specimens and average number of specimens per species) and high-throughput sequencing (HTS)-based abundance estimation with different profiling methods. Pearson correlations (*R*) are given for the correlation between known species biomass and estimated biomass ($P < 0{\cdot}001$ in all cases except when indicated). Total number of BLAST hits (TH) and average number of hits per species (AH) are also provided

	Number of specimens		MitoRL			PCR-barcodes			DeNovoRL		
	Total	Per species	*R*	TH	AH	*R*	TH	AH	*R*	TH	AH
ADS	273	6·7	0·709	24 536	481·1	0·747	1862	47·7	0·625	24 501	350·0
ANC	327	4·9	0·683	21 866	308·0	0·727	1562	26·9	0·508	21 984	314·1
EUM	223	5·4	0·564	17 860	343·5	0·538	1384	31·5	0·569[1]	17 801	258·0
HOR	156	5·8	0·812	36 506	960·7	0·754	2654	106·2	0·375[2]	32 700	467·1
JCB	206	5·7	0·614	33 941	722·1	0·649	2518	71·9	0·354[2]	25 877	375·0
LAS	336	6·0	0·779	27 978	458·7	0·753	1976	41·2	0·665	25 838	369·1
MAC	232	4·7	0·728	38 577	665·1	0·632	2516	50·3	0·614	31 606	451·5
OMA	299	6·6	0·864	24 948	430·1	0·808	1846	41·0	0·824	19 853	283·6
SAN	252	5·4	0·744	27 106	459·4	0·687	2210	49·1	0·650	17 562	250·9
TUE	303	6·3	0·802	28 134	574·2	0·681	2510	64·4	0·689	22 780	330·1

[1] $P < 0{\cdot}01$; [2] $P > 0{\cdot}05$.

Table 3. Multiple-site taxonomic dissimilarity values and multiple-site phylogenetic dissimilarities, according to different assemblage profiling methods. Dissimilarities are provided both for the total dissimilarity and for the species turnover and nestedness-resultant components

	Morpho	MitoRL	PCR-barcodes	DeNovoRL
Taxonomic multiple-site				
Total dissimilarity	0·825	0·822	0·813	0·766
Turnover	0·789	0·790	0·779	0·720
Nestedness resultant	0·036	0·032	0·034	0·046
Phylogenetic multiple-site				
Total dissimilarity	0·748	0·740	0·756	0·689
Turnover	0·702	0·691	0·707	0·623
Nestedness resultant	0·047	0·049	0·050	0·066

phylogenetic signal of whole mitogenomes for resolving family-level relationships in Coleoptera (Timmermans *et al.* 2010; Cameron 2014) provides the evolutionary context for integrating ecological and evolutionary analyses.

The alternative approach of library construction straight from the mixed samples is attractive because it does not require additional sequencing, but library completion was only 43·8% of species from nearly twice the amount of sequence data. The lower assembly success and species representation were expected from the variation in read coverage among species of different body sizes and represented in variable numbers of individuals. We found that read coverage varied among species over three orders of magnitude and, as expected, low-abundance species were significantly more likely to lack a mitogenome in the library. A higher sequencing depth for the local libraries would have likely increased the retrieval of contigs for those low-abundance species. The shorter average contig length also complicated the identification of sibling contigs corresponding to non-overlapping portions of a single mitogenome. This problem forced us to retain only contigs of unequivocal orthology centred on the *cox1* region, which reduced the species represented in the library (some of which may have been represented by contigs from elsewhere in the mitogenome).

An additional challenge was the intraspecific variation in the sample. Local assemblages included roughly three times as many different haplotypes (based only on the *cox1*-5′ region) than the number of distinct species, and therefore, the assemblers were confronted with polymorphisms. However, we found that species that were successfully assembled had significantly greater genetic variation in the sample (Appendix S9), probably because potential negative effects of genetic variation on assembly were compensated for by the greater number of reads in those species. Genetic diversity also did not have negative effects at the genus level, as there was no reduced recovery in species-rich genera.

Another potential risk was the assembly of heterospecific contigs, although in our data set chimeras were restricted to a few cases only. Compared against the well-established circular genomes, we only found 10 chimeric molecules, always between species of the same genus. This is not an exhaustive search for chimeras as the reference set of circular genomes is only ~50% complete, and hence may reduce heterologous hits. However, even if the proportion of chimeras for this subset cannot be interpreted in absolute terms, the observed rate suggests that chimera formation in *de novo* assembly from complex natural assemblages would be very low, with no specific preponderance in regard to contig length, assembler used, library type and point of recombination.

SPECIES INVENTORY, ABUNDANCE AND DIVERSITY PATTERNS

Assemblage profiling based on read matches was affected by the completeness of species representation and fragment length in the reference libraries. Given the high inventory completion in the PCR-barcodes and MitoRL libraries, false absences due to missing reference sequences were low, resulting in high profiling success against the reference inventory (Table 1). While minimally affected by missing reference genomes, the analysis based on the PCR-barcodes library had a higher proportion of false presences, compared with the one using the much longer

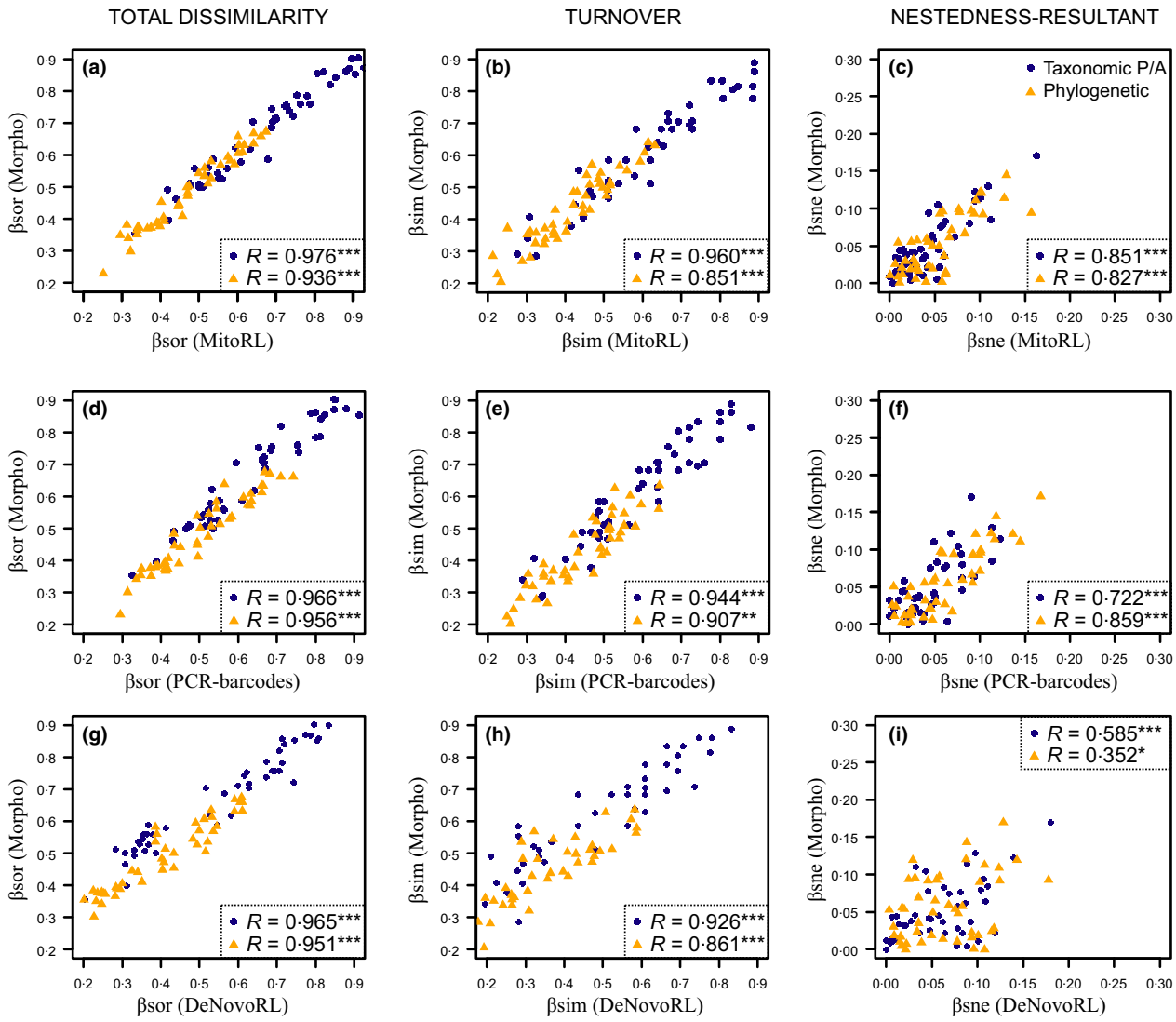

Fig. 3. Scatterplots and Pearson correlation values of the relationship between the pairwise taxonomic (blue) and phylogenetic (orange) dissimilarity values derived from morphological species data (morpho) and the corresponding estimates from HTS-based profiling methods. Dissimilarities were computed as total dissimilarity, βsor (a, d, g), species turnover, βsim (b, e, h), and nestedness-resultant components of dissimilarity, βsne (c, f, i). Significance values are computed with Mantel tests (***$P < 0{\cdot}001$; **$P < 0{\cdot}01$; *$P < 0{\cdot}05$). HTS, high-throughput sequencing.

MitoRL sequences with their greater power for species discrimination. In comparison, the DeNovoRL library was highly incomplete and consequently missed between a quarter to nearly half of local species. The missing species were mostly rare or small-bodied, as numerous reads are needed for the generation of contigs. However, the DeNovoRL method still provided valuable information for comparative purposes, since it usually detected more species where there were more and *vice versa*.

When assessing the relative abundance of individual species, we observed a remarkable correlation between read numbers and biomass, in particular for the MitoRL and PCR-barcodes libraries-based assessments, and a lower correlation with the DeNovoRL approach. However, when assessing individual assemblages, the correlation varied more strongly among the three reference sets, presumably due to stochasticity linked to small sample size. Our assessment may be affected by the crude design of the experiment, which relied on semiquantitative, non-destructive DNA extraction (Baselga, Gómez-Rodríguez & Vogler, 2015) and biomass estimates from average body length, which itself was taken from values given in the literature, rather than from measurements of specimens used. Yet, the close correlation is very encouraging for sequence-based studies of abundances in natural communities, in particular when large numbers of specimens are analysed in bulk.

Regarding taxonomic and phylogenetic dissimilarity among assemblages (i.e. beta diversity patterns), the MitoRL and PCR-barcodes data recovered the reference patterns with great accuracy. The analysis using the DeNovoRL library was likely affected by the taxonomic loss in the assemblage profiling that resulted in an underestimate of overall taxonomic and phylogenetic heterogeneity, in particular in the turnover component (i.e. substitution of one species by another). This finding was expected from the lower detection rate of rare species that

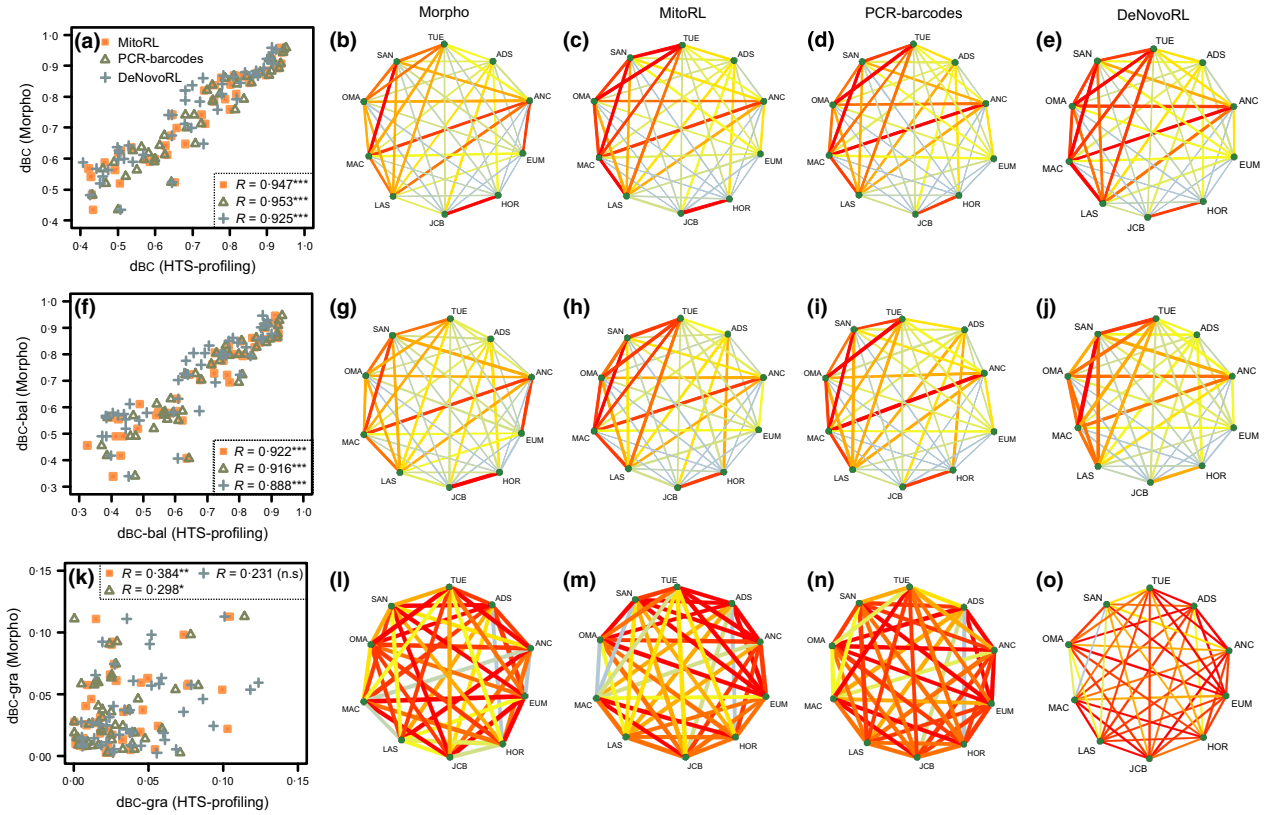

Fig. 4. Scatterplots of the relationship between abundance-based pairwise dissimilarity values derived from morphological species data and values based on HTS-profiling; as well as the respective Pearson correlation values (a, f, k). Dissimilarities values were computed as total abundance-based dissimilarity, measured as the Bray–Curtis index, d_{BC}, (a–e), dissimilarity derived from balanced variation in abundance, d_{BC}-bal, (f–j), and dissimilarity derived from unidirectional abundance gradients d_{BC}-gra (k–o). Significance values are computed with Mantel tests (***$P < 0{\cdot}001$; **$P < 0{\cdot}01$; *$P < 0{\cdot}05$). Networks represent beta diversity patterns based on HTS-estimated abundance (b–e, g–j, l–o), where nodes represent ecological samples and edges represent similarity. Edge width is proportional to assemblage similarity, which is also shown as a colour scale, from blue (low similarity) to yellow (intermediate similarity) and red (high similarity). HTS, high-throughput sequencing.

usually increase dissimilarity between assemblages. Moreover, the DeNovoRL approach poorly captured the dissimilarity pattern derived by unidirectional abundance gradients, implying that even in assemblages with similar composition, this approach was unable to capture the variation in species abundances. Finally, for all approaches, the generally good agreement of inferred dissimilarity with the reference pattern was probably related to the high true dissimilarity of local assemblages in the data set (i.e. 73% of β_{sim} pairwise comparisons >0·50) as the effects of incomplete sampling are expected to be worse in highly similar communities, where we would expect to lose shared species and thus incorrectly increase dissimilarity.

Conclusions

HTS of whole assemblages is likely to transform the analysis of complex biodiversity samples. We show that species richness, abundance and phylogenetic relationships can be reliably estimated from specimen mixtures. Mitochondrial genomes are a unique resource for assembly from mixed metagenomes, while the remaining >90% of the metagenome potentially also provide valuable information on species diversity and comparative genomics, but their assembly is far less straightforward (B. linard, A. Crampton-Platt, M.J.T.N. Timmermans, A.P. Vogler, under review). The leaf beetles studied here are real-world natural assemblages that include closely related species represented by multiple haplotypes, and the analysis of several assemblages permits comparisons of species richness and community dissimilarity. Because they had been well characterized with conventional methods, they provide information about the degree to which imperfect community inventories still recover the true biodiversity patterns. The approach can be applied to any group of organisms with non-recombining organelle genomes, although larger genomes, for example those of chloroplasts, constitute greater challenges, not least due to lower sequence variation that hampers the unequivocal assembly of close relatives. A critical question is the need for a reference library against which to conduct the read-based community profiling. While the added power of an external (or custom-designed) database was very apparent in the current study, the principal biodiversity patterns were already revealed against the (limited) set of contigs generated from the 'specimen soup' itself, at least for those species contributing substantially to the biomass. The good success rate with barcode data is also encouraging, as these databases currently approach three million records (Cristescu 2014). While these sequences

remain poor for phylogenetics (e.g. Klopfstein, Kropf & Quicke 2010), the full mitogenomes provided strongly supported trees even at deep levels and also improved the accuracy of species discrimination.

Acknowledgements

We are grateful for discussions and constructive comments to Carmelo Andújar and to Peter Foster for maintaining NHM computational resources (CTAG). Benjamin Linard provided invaluable bioinformatic support. We also thank Douglas Yu and two anonymous reviewers for comments on a previous version. This work was supported by the NHM Biodiversity Initiative, the Spanish Ministry of Science and Innovation (Grant Number CGL2009-10111 and CGL2013-43350-P to A.B.), Xunta de Galicia (Postdoctoral Fellowship POS-A/2012/052 to C.G.R.), NERC (Postdoctoral Fellowship NE/I021578/1 to M.T.J.N.) and NHM/UCL (PhD studentship to A.C.-P.).

Data accessibility

GenBank Accession Numbers for sanger-derived sequences: KF134544–KF134651 and KF652242–KF656666 (*cox1*-5′), KP762969–KP763108 (*cox1*-3′), KP763109–KP763241 (*rrnL*) and KP762810–KP762968 (18S). Contigs in the MitoRL and DeNovoRL reference libraries are available from the Dryad Digital Repository: http://datadryad.org/resource/doi:10.5061/dryad.3rh21 (Gómez-Rodríguez *et al*. 2015).

References

Avise, J.C. (2000) *Phylogeography. The History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts.

Baird, D.J. & Hajibabaei, M. (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Baselga, A. (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga, A. (2013) Separating the two components of abundance-based dissimilarity: balanced changes in abundance vs. abundance gradients. *Methods in Ecology and Evolution*, **4**, 552–557.

Baselga, A., Gómez-Rodríguez, C. & Vogler, A.P. (2015) Multi-hierarchical macroecology at species and genetic levels to discern neutral and non-neutral processes. *Global Ecology and Biogeography*, **8**, 873–882.

Baselga, A. & Orme, C.D.L. (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Baselga, A., Gómez-Rodríguez, C., Novoa, F. & Vogler, A.P. (2013) Rare failures of DNA bar codes to separate morphologically distinct species in a biodiversity survey of Iberian leaf beetles. *PLoS One*, **8**, e74854.

Bensasson, D., Zhang, D.X., Hartl, D.L. & Hewitt, G.M. (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, **16**, 314–321.

Bininda-Emonds, O.R.P. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, **6**, 156.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., Yu, D.W. & de Bruyn, M. (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, **29**, 358–367.

Brown, J. & Lomolino, M. (1998) *Biogeography*. Sinauer Associates Inc, Sunderland, Massachusetts.

Cameron, S.L. (2014) Insect mitochondrial genomics: implications for evolution and phylogeny. *Annual Review of Entomology*, **59**, 95–117.

Cristescu, M.E. (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, **29**, 566–571.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**, 20140562.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 8007–8012.

Gillett, C.P.D.T., Crampton-Platt, A., Timmermans, M.J.T.N., Jordal, B.H., Emerson, B.C. & Vogler, A.P. (2014) Bulk *de novo* mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J.T.N., Baselga, A. & Vogler, A.P. (2015) Data from: Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, doi:10.5061/dryad.3rh21.

Gómez-Zurita, J., Hunt, T. & Vogler, A.P. (2008) Multilocus ribosomal RNA phylogeny of the leaf beetles (Chrysomelidae). *Cladistics*, **24**, 34–50.

Hebert, P.D.N., Ratnasingham, S. & deWaard, J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B-Biological Sciences*, **270**, S96–S99.

Klopfstein, S., Kropf, C. & Quicke, D.L.J. (2010) An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). *Systematic Biology*, **59**, 226–241.

Leprieur, F., Albouy, C., De Bortoli, J., Cowman, P.F., Bellwood, D.R. & Mouillot, D. (2012) Quantifying phylogenetic beta diversity: distinguishing between 'true' turnover of lineages and phylogenetic diversity gradients. *PLoS One*, **7**, e42760.

Ma, C., Yang, P.C., Jiang, F., Chapuis, M.P., Shali, Y., Sword, G.A. & Kang, L. (2012) Mitochondrial genomes reveal the global phylogeography and dispersal routes of the migratory locust. *Molecular Ecology*, **21**, 4344–4358.

Miller, M.A., Pfeiffer, W. & Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE)*, 14 Nov, 2010, New Orleans, LA, pp. 1–8.

Pons, J., Barraclough, T.G., Gómez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. & Vogler, A.P. (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.

Shokralla, S., Spall, J.L., Gibson, J.F. & Hajibabaei, M. (2012) Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, **21**, 1794–1805.

Song, H., Buhay, J.E., Whiting, M.F. & Crandall, K.A. (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 13486–13491.

Stein, E.D., Martinez, M.C., Stiles, S., Miller, P.E. & Zakharov, E.V. (2014) Is DNA barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the United States? *PLoS One*, **9**, e95525.

Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C. & Liston, A. (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–364.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S. *et al*. (2015) Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, online early, doi:10.1093/nar/gku917.

Timmermans, M.J.T.N., Dodsworth, S., Culverwell, C.L., Bocak, L., Ahrens, D., Littlewood, D.T.J., Pons, J. & Vogler, A.P. (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*, **38**, e197.

Whittaker, R.J., Araújo, M.B., Paul, J., Ladle, R.J., Watson, J.E.M. & Willis, K.J. (2005) Conservation Biogeography: assessment and prospect. *Diversity and Distributions*, **11**, 3–23.

Wilson, E.O. (1988) *Biodiversity*. National Academy Press, Washington, DC.

Yu, D.W., Ji, Y.Q., Emerson, B.C., Wang, X.Y., Ye, C.X., Yang, C.Y. & Ding, Z.L. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L. *et al*. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

Received 30 January 2015; accepted 14 March 2015
Handling Editor: M. Gilbert

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Details on the localities selected in the present study.

Appendix S2. (a) Maximum likelihood tree of PCR-derived *cox1*-5′ sequences for 201 species ('barcodes tree'). (b) Maximum likelihood tree of the evolutionary relationships of 206 leaf beetle species built using PCR-derived sequences of the *cox1*-3', *cox1*-5', *rrnL* and 18S genes ('4-loci tree').

Appendix S3. NGS-Library preparation details.

Appendix S4. Maximum likelihood tree of the evolutionary relationships between contigs assembled in the MitoRL library.

Appendix S5. Maximum likelihood tree of the evolutionary relationships between orthologus contigs assembled in the DeNovoRL library.

Appendix S6. Maximum likelihood tree to identify and collapse sibling contigs after 22 contigs from the DeNovoRL library complemented the MitoRL library.

Appendix S7. Assembly details.

Appendix S8. Relationship between the number of contigs and their length recovered with the different assembly methods.

Appendix S9. Boxplots showing the differences between species recovered and not recovered in the MitoRL and DeNovoRL reference libraries.

Appendix S10. Description of chimeras detected in this study.

Appendix S11. Maximum likelihood tree of leaf beetles evolutionary relationships based on mitochondrial genomes of 161 known morphological species (detailed version).

Appendix S12. Assemblage profiling results.

Appendix S13. Relationship between the abundance of a species in a locality and the number of false absence and true presence cases observed for that species.