Received Date : 20-May-2016
Accepted Date : 28-Sep-2016
Article type : Research Article
Editor : M. Gilbert

Intraspecific genetic variation in complex assemblages from mitochondrial metagenomics: comparison with DNA barcodes

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

[1] Department of Life Sciences, Natural History Museum, London, SW7 5BD, United Kingdom

[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, United Kingdom

[4] Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, SL5 7PY, United Kingdom

[5] Department of Natural Sciences, Hendon Campus, Middlesex University, London, NW4 4BT, United Kingdom

* Author for correspondence:

Carola Gómez-Rodríguez

Departamento de Zoología, Facultad de Biología, Universidad de Santiago de Compostela, c/ Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, Spain

E-mail: carola.gomez@usc.es, Phone: +34 881813278

Running title: Mitometagenomics: Intraspecific genetic variation

Abstract

1. Metagenomic shotgun sequencing, using Illumina technology, and *de novo* genome assembly of mixed field-collected samples of invertebrates readily produce mitochondrial genome sequences, allowing rapid identification and quantification of species diversity. However, intraspecific genetic variability present in the specimen pools is lost during mitogenome assembly, which limits the utility of 'mitochondrial metagenomics' for studies of population diversity.

2. Using 10 natural communities (>2600 individuals) of leaf beetles (Chrysomelidae), DNA variation in the mitochondrial *cox1*-5' 'barcode' was compared for Sanger sequenced individuals and Illumina shotgun sequenced specimen pools.

3. Generally, only a single mitochondrial contig was assembled per species, even in the presence of intraspecific variation. Ignoring ambiguity from the use of two different assemblers, the *cox1* barcode regions from these assemblies were exact nucleotide matches of a Sanger sequenced barcode in 90.7% of cases, which dropped to 76.0% in assemblies from samples with large intra and interspecific variability. Nucleotide differences between barcodes from both data types were almost exclusively in synonymous 3rd codon position, although the number of affected sites was very low, and the greatest discrepancies were correlated with poor quality of Sanger sequences.

4. Unassembled shotgun reads were also used to score single nucleotide polymorphisms and to calculate intraspecific nucleotide diversity (pi) for all available populations at each site. These values correlated with Sanger sequenced *cox1* variation but were significantly higher.

5. Overall, the assemblage-focused shotgun sequencing of pooled samples produced nucleotide variation data comparable to the well-established specimen-focused Sanger approach. The findings thus extend the application of mitochondrial metagenomics of complex biodiversity samples to the estimation of diversity below the species level.

Keywords: Mitochondrial metagenomics, mitometagenomics, genome skimming, Chrysomelidae, intraspecific variability, population genetics, DNA barcodes

Introduction

High-throughput shotgun sequencing (HTS) of mixtures of specimens and the bioinformatic reconstruction of mitochondrial genomes (mitochondrial metagenomics or mito-metagenomics, MMG) can greatly speed up biodiversity inventories (Tang *et al.* 2014; Crampton-Platt *et al.* 2015; Gómez-Rodríguez *et al.* 2015). The methodology allows the rapid identification of the biological entities present in a complex sample. MMG is a 'genome skimming' (Straub *et al.* 2012) procedure, which takes advantage of the fact that the mitochondrial genome is present in numerous copies per nuclear genome and thus is amenable to *de novo* genomic assembly at fairly low sequencing depth (Zhou *et al.* 2013; Tang *et al.* 2014; Crampton-Platt *et al.* 2015). The characterisation of mixed communities is also possible with PCR-based metabarcoding (e.g., Ji *et al.* 2013), but MMG in addition provides information on the relative abundance of species and allows reliable inference of phylogenetic diversity (Gómez-Rodríguez *et al.* 2015)

The assembly from mixtures combines the shotgun reads such that separate mitogenomes are reconstructed for multiple species (Crampton-Platt *et al.* 2016). Thus, the assembly software is required to separate multiple orthologous DNA fragments originating from different species. This poses a challenge for metagenomic assembly of bulk biodiversity samples, which contain closely related species and haplotypes present in unequal numbers and whose sequence divergence may be close to the sequencing error rate (Nagarajan & Pop 2013). Assemblers typically collapse the genetic variants present in a specimen mixture into a single sequence at (approximately) the level of species (Crampton-Platt *et al.* 2016). However, the nucleotide variation at SNPs may be inconsistently resolved if the assembler combines reads from two or more closely related haplotypes, and thus the 'consensus' contig may represent a recombinant sequence that does not correspond to a genuine haplotype

present in the population. A related concern is that collapsing of genetic variation at the contig level limits the use of these sequences for the quantification of intraspecific diversity, unlike Sanger-based DNA barcoding, and thus constrains the application of MMG to questions about species-level diversity.

Contigs from MMG have been shown to be highly consistent with Sanger-derived sequences in the absence of intraspecific variability (Gillett *et al.* 2014). Equally, chimerical mosaic contigs produced by the *in silico* recombining of genomes of different species are rare under these conditions, in particular when the assembly uses reads of 250-300 bp in length (Gómez-Rodríguez *et al.* 2015). However, the assembly inevitably becomes more complicated when applied to natural communities containing genetic variants and closely related species. As a limited number of discrete contigs is assembled from the primary reads, it is not clear how the resulting products correspond to haplotypes actually present in the population. Comparing the MMG sequences with Sanger-derived haplotypes may reveal potential artefacts in the assembly of these data. In addition, intraspecific variability in a natural community may be investigated based on the primary shotgun reads, which hold complete information on both the interspecific and intraspecific genetic variability of the sample, although the individual reads are affected by greater noise from read errors than the assembled contigs. Previous studies have illustrated the use of shotgun data as a cost-effective technique for population genetic analysis, including the estimation of allele frequencies (e.g., Futschik & Schlötterer 2010; Zhu *et al.* 2012; Gautier *et al.* 2013) and nucleotide diversity (e.g., Hellmann *et al.* 2008; Futschik & Schlötterer 2010; Ferretti, Ramos-Onsins & Pérez-Enciso 2013). These studies were based on whole-genome sequencing of pools of individuals of a single species at a high depth of coverage (see review in Schlötterer *et al.* 2014), suggesting that population genetics parameters can also be

estimated for each species in mixed assemblages if focused on mitogenomes present in high copy numbers.

This study makes side-by-side comparison of datasets obtained for the same specimens and mixed species assemblies using individual-based Sanger sequencing and bulk-specimen lllumina sequencing. Gómez-Rodríguez *et al.* (2015) conducted an MMG study on 10 local communities of leaf beetles (Chrysomelidae) from the Iberian Peninsula (>2600 individuals). This dataset encompasses large intraspecific and interspecific variability, with a few dozen species per locality, most of which are represented by numerous individuals and multiple haplotypes (based on the *cox1*-5' fragment) which were sequenced in bulk. Sanger sequences (hereafter PCR-barcodes) for each individual in the mixed community samples were available from an earlier study (Baselga, Gómez-Rodríguez & Vogler 2015). Sequencing with either approach arrived at very similar conclusions about species richness and beta diversity, but the study made no attempt to address the question about intraspecific genetic variation. Here we address (a) the sequence similarity of PCR-based haplotypes in a population with the assembled consensus haplotypes obtained from those same specimens, and (b) the prospect of using shotgun read mapping to obtain measures of genetic variation comparable to estimates from conventional barcoding methods. The results are critical for the utility of shotgun reads obtained from a 'biodiversity soup' in population and community genetics, and for the wider scope of MMG studies below the species level.

MATERIALS AND METHODS

PCR-barcodes library and MMG-sequences

Sanger sequenced PCR-barcodes (*cox1*-5') from an existing study of leaf beetle (Chrysomelidae) communities from the Iberian Peninsula (Baselga, Gómez-Rodríguez & Vogler 2015) included 4533 specimens from 20 local assemblages that were successfully sequenced. All specimens had been identified to a Linnaean species according to taxonomic monographs of Warchalowski (2003) and Petitpierre (2000) (see Baselga, Gómez-Rodríguez & Vogler 2015). In addition, species delimitation was conducted based on the *cox1*-5' sequences using the Generalized Mixed Yule Coalescent (GMYC) method (Pons *et al.* 2006).

The corresponding Illumina sequences were obtained, in a previous study, from a subset of ten local assemblages, totalling 2607 specimens (for 88.9% of which a PCR-barcode is available) in 171 species that exhibited 1089 different PCR-barcodes (haplotypes) (Gómez-Rodríguez *et al.* 2015). Mitogenome assemblies from various combinations of these specimens had been obtained from shotgun sequencing of specimen mixtures following a standard protocol, which involved these main steps (see Gómez-Rodríguez *et al.* 2015): (i) Illumina TruSeq libraries preparation from pooled DNA extracts of single individuals; (ii) sequencing on an Illumina MiSeq sequencer; (iii) assembly with Newbler 2.7 and IDBA-UD 1.1.1, followed by removal of non-mitochondrial and short (<3000 bp) contigs; (iv) merging of identical or very similar contigs produced by different assemblers through re-assembly in Geneious 5.6 (minimum overlap = 500bp; minimum overlap identity = 99%; minimum gaps per read = 1%); (v) taxonomic assignment of contigs via Blast matches to the PCR-barcodes database. From these reconstructed mitogenomes we here only consider the *cox1*-5' barcode

regions (655bp) which are the assembled consensus of numerous shotgun reads and hereafter will be referred to as MMG-barcodes.

Using the above methodology, Gómez-Rodríguez *et al.* (2015) built three different sets of mitochondrial genomes. First, a 'mitogenome reference library' (MitoRL) was generated by shotgun sequencing of a concentration-adjusted mixture of one representative for each of the 176 focal species (including five species added from another locality), which was sequenced on 117% of the MiSeq flow cell. The assemblies of mitogenomes from this pool are considered a benchmark for the performance of MMG under no intraspecific variation. Second, a 'local library' (LocL 1 to 10) was generated by shotgun sequencing and *de novo* assembly from each of 10 local assemblages, each of which included between 156 and 336 specimens drawn from between 27 and 67 Linnaean species and 99 to 156 known *cox1* haplotypes. Each LocL was sequenced on approximately 20% of a MiSeq flow cell (see details in Gómez-Rodríguez *et al.* 2015). This set informs us about MMG performance in the presence of intraspecific and interspecific genetic variation in a natural community and under variation in abundance and biomass of species, which ranged from 1 to 25 individuals per species of 1.5 to 10.5 mm in length. Third, an additional assembly was conducted on the combined reads from all LocL pools to generate a 'combined library' (CombL), which further increased the intraspecific and interspecific nucleotide variation and difference in abundance/biomass among species in the mixture but also increased the number of reads per species available for assembly.

Nucleotide discrepancies between PCR-barcodes and MMG-barcodes

All PCR-barcodes and MMG-barcodes were aligned with transAlign (Bininda-Emonds 2005) and a matrix of uncorrected pairwise distances between both types of sequences was computed with the *dist.alignment* function in seqinr (Charif & Lobry 2007). This distance

matrix was transformed into a discrepancy matrix that summarizes the number of nucleotide discrepancies between two sequences. This step was needed because the use of multiple assemblers produced ambiguous base calls (see below) that complicate the estimate of distances, while the discrepancy matrix gives a direct account of the number of sites differing between two sequences.

Subsets of the discrepancy matrix were extracted for each morphological species and mitochondrial contig set as needed to test discrepancies between: (a) the MMG-barcode and the corresponding PCR-barcode of the exact specimen used for library construction in MitoRL; (b) the MMG-barcode of a given morphological species produced in a given LocL assembly and all PCR-barcodes of that same species found in the corresponding local assemblage; (c) the MMG-barcode of a given morphological species produced in the CombL assembly and all the PCR-barcodes of that morphological species in the full database. In all cases, the smallest number of discrepancies was selected as best match. Only barcodes with length >524 bp, corresponding to 80% of the *cox1*-5' fragment, were considered for this comparison, and 99.2% of the sequences in the PCR-barcodes database fulfilled this criterion.

MMG-barcodes exhibiting at least one discrepancy compared to their most similar PCR-barcode ('discrepant MMG-barcodes') were further explored to establish the nature of these changes that may be due to an incomplete PCR-barcode library or sequencing errors in either data set. The quality of PCR-barcodes was assessed in Geneious as the number of bases with high quality (phred score >30) in the consensus sequence after trimming the low quality ends. Phred scores were computed with CodonCode Aligner v6. The quality of shotguns reads was assessed for the full *cox1*-5' dataset as well as for each discrepancy pair. To build the *cox1*-5' reads data set, the shotgun library was searched against the database of PCR-barcodes using BLAST (e-value < 10^{-5}). In the cases where discrepancies were observed between a MMG-barcode and the closest PCR-barcode, the *cox1*-5' reads data set was also

run against the PCR-barcodes of those morphological species, in order identify the reads for those species and estimate their quality. The read quality of these shotgun reads was then assessed by establishing the proportion of high quality bases (phred > 30) with the Bioconductor package qrqc (Buffalo 2012) (see details in Appendix S3).

Finally, for each morphological species, the polymorphic sites in the *cox1*-5' region were identified from all available PCR-barcodes (hereafter 'PCR-polymorphisms') using a custom R script. The assessment of polymorphic sites from both PCR-barcodes and MMG-barcodes allowed the identification of 'MMG-polymorphisms', i.e. base positions that are invariant in the PCR-barcodes but show a different nucleotide in the MMG-contigs. Only species with more than one PCR-barcode could be assessed for these sites because of the need to identify PCR-polymorphisms.

Assessment of genetic diversity from shotgun reads

Nucleotide diversity (Nei 1987) from PCR-barcodes was computed with the *nuc.div* function in the *pegas* package (Paradis 2010) in R (R Development Core Team 2015). To compute nucleotide diversity from shotgun reads, the reads from each LocL were mapped in Geneious 8.0 (custom sensitivity settings: minimum overlap: 100bp, maximum mismatches per read = 3%, minimum overlap identity = 97%) against consensus sequences of each morphological species that had been computed from the PCR-barcodes of that local assemblage using the *consensus* function in the seqinR package (Charif & Lobry 2007). Geneious mapping allows setting a maximum divergence value in read mapping, which is a convenient feature when reads from different species are in the pool. Mapped reads were exported as BAM file and nucleotide diversity (Tajima's Pi) subsequently computed for a single window of 655 bp (length of barcode fragment) using the variance-sliding perl script of Popoolation (Kofler *et al.* 2011) for all *cox1*-5' fragments that had a coverage of >2 reads for >80% (minimum

count=1) of their nucleotide positions. The same analyses were also conducted using GMYC-groups instead of morphological species. This eliminated discrepancies from read mapping against the PCR-barcodes within some Linnaean species whose haplotype diversity exceeded the threshold set for read mapping (3%).

Repeated Measures ANOVA (RM ANOVA) on natural-logarithm transformed values of average nucleotide diversity, for each morphological species and local assemblage, was used to assess whether the value of average nucleotide diversity obtained from PCR-barcodes was significantly different from the value obtained from shotgun reads. The dataset used (PCR-barcodes or shotgun reads) was considered as within-subjects factor and locality as between-subjects factor. Additionally, a Pearson correlation on natural-logarithm transformed values of both datasets was conducted. Statistical analyses were done in Statistica 7.0.

RESULTS

Nucleotide discrepancies between PCR-barcodes and MMG-barcodes

MMG-barcodes were obtained by extracting the *cox1*-5' regions from the complete or partial mitogenome contigs that had been assembled previously from three shotgun data sets, derived from: (a) the singular representatives of each species (MitoRL), (b) all specimens in each of 10 localities, assembled separately (LocL 1 to 10), and (c) all specimens in the dataset, assembled altogether (CombL) (see Materials and Methods). Where contigs produced with the IDBA and Newbler showed slight differences in the re-assembly with Geneious, a IUPAC ambiguity code was introduced in those positions. The re-assembly generally produced a single consensus contig for each species containing the *cox1*-5' fragment, except for nine cases (18 contigs out of 402) that recovered two contigs per morphological species (Appendix S1). In four of these cases, the PCR-barcodes showed very high within-species genetic distance (between 4.1% and 8.4%) and may correspond to different cryptic species.

The final number of contigs used for the analysis of sequence variation in the *cox1*-5' region included 128 MMG-barcodes in MitoRL, corresponding to 128 morphological species; 177 MMG-barcodes in LocL, corresponding to 76 morphological species represented by 1668 sequences in the PCR-barcode database; and 82 MMG-barcodes in CombL, corresponding to 80 species represented by 1624 sequences in the PCR-barcode database.

'Ambiguous sites' represented by IUPAC ambiguity codes and introduced in the Geneious re-assembly (see Material and Methods) affected 3.1% (MitoRL), 15.2% (LocL) and 13.5% (CombL) of all MMG-barcodes, and in the majority of cases only one ambiguous site (in 655 positions) was observed per MMG-barcode (Table 1, Figure 1). Allowing for these 'ambiguous sites', the proportion of identical MMG-barcodes and PCR-barcodes was 90.7% (MitoRL), 78.8% (LocL) and 76.0% (CombL). Among the divergent pairs, the average number of discrepancies ('discrepant sites') with the closest matching PCR-barcode ranged from 2.7 (CombL) to 7.8 (LocL) of the 655 positions (Table 1 and Figure 1b). These averaged numbers were greatly inflated by a few outliers differing by numerous sites (Figure 1b). These cases with more discrepancies showed notably lower sequence quality in the PCR-barcodes (proportion of high quality bases [phred >30] ranging from 8.1% to 99%) than the average for the full dataset (average proportion of high quality bases [phred >30] = 92.2 ± 10.5 S.D., see details in Appendix S2). In contrast, the quality of MMG reads across all discrepancy cases was very high (Figure 1b) and similar to the quality observed for the full *cox1*-5' reads dataset (proportion of high quality bases = 90%, see details in Appendix S3).

We further investigated species exhibiting variation in PCR-barcodes (≥2 different PCR-barcodes required) but differing from MMG-barcodes for that species (number of cases = 60). In most cases, the discrepancies in the MMG-barcodes were observed at base positions

that also showed variation (i.e. polymorphisms) within the PCR-barcodes for that morphological species. However, in 15 cases (6 in MitoRL, 7 in LocL and 2 in CombL) the MMG-barcodes exhibited nucleotide differences at sites that were otherwise invariable in the PCR-barcode alignment for that morphological species. Additionally, unique polymorphisms were also observed in six MMG-barcodes that had not been initially identified as 'discrepant MMG-barcodes' because they did not have a corresponding PCR-barcode from the same locality. These MMG-barcodes could be used for the assessment of MMG-polymorphisms, for a total of 66 cases, since PCR-barcodes for the same morphological species were available from other localities and thus allowed the identification of PCR-polymorphisms. The number of MMG-polymorphisms (SNPs unique to the MMG-barcodes) ranged from 1 to 12 (mean = 3.8; median = 1; S.D. = 4.0) and they mostly occurred in the 3rd codon position (total number of MMG-polymorphisms: 1st codon position = 7; 2nd codon position = 2; 3rd codon position = 70; Figure 2). The number of SNPs unique to the MMG-barcodes was inversely related to both the relative success rate of the Sanger-barcoding (the number of PCR-barcodes divided by the number of specimens for a given morphological species in the full dataset) and to the total number of PCR-barcodes available for the species (Figure 2b). This type of 3rd position SNP appeared to be concentrated in species from a few genera, including *Cryptocephalus* (n=49 SNPs), *Altica* (n=6), *Aphthona* (n=6) and *Labidostomis* (n = 5).

Assessing nucleotide diversity within populations using shotgun reads

Nucleotide diversity was computed for each 'population', defined as all the individuals of a Linnaean species at a sampling locality. An example of measures of nucleotide diversity in MMG and Sanger sequence data is shown for populations of five species widespread across the ten study sites (Figure 3). Nucleotide diversity from PCR-barcodes was computed for 316

out of 457 populations, corresponding to 95.0% of PCR-barcodes in the dataset, with a mean number of populations in a locality = 31.6 ± 7.6 (S.D.). Those 'populations' not included were from sites with only one specimen representing a species at a locality (n=119 populations) or only one or no PCR-barcode sequence available for the species due to Sanger sequencing failure (n = 22 populations). Nucleotide diversity based on shotgun reads was computed for 186 populations (40.8% of the dataset, mean number of populations per locality = 18.7 ± 4.3 [S.D.]). The absence (n = 109) or insufficient number (n=162) of mapped reads greatly reduced the number of available populations. The analysis was also conducted based on GMYC delimitation, rather than grouping of haplotypes based on assignment to a Linnaean species, which produced 474 local populations, of which nucleotide diversity data could be obtained for 320 populations from PCR-barcodes (93.3% of the dataset) and for 186 populations from shotgun reads (39.2% of the dataset).

Nucleotide diversity values from PCR-barcodes and shotgun reads were highly correlated both when assessing morphological species (Pearson $R = 0.724$, $p < 0.001$, $n = 180$) or GMYC groups (Pearson $R = 0.776$, $p < 0.001$, $n = 178$) (Figure 4). However, the nucleotide diversity was significantly higher when calculated based on shotgun reads than PCR-barcodes (RM ANOVA within-subjects effect; morphological species: $F_{1, 170} = 12.39$, $p < 0.001$; GMYC data: $F_{1, 168} = 34.01$, $p < 0.001$; Figure 5a, c). Locality or the interaction between locality and sequencing method had no significant effect ($p>0.05$; Figure 5b, d). Note that when the natural-logarithm transformation is not applied, Pearson correlation of nucleotide diversity for PCR-barcodes and shotgun reads was high in the analyses of GMYC-groups (Pearson $R = 0.802$, $p < 0.001$, $n = 178$) but only moderate in the analyses of Linnaean species (Pearson $R = 0.524$, $p < 0.001$, $n = 180$), due to the presence of some clear

outliers in the latter, which were affected by large genetic distances within morphologically defined species (see Appendix S4 for details).

Discussion

We show that assembly from shotgun sequences of mixed communities generally produces a single contig per species, effectively removing the intraspecific variation present in complex samples. While this characteristic simplifies the MMG approach for the rapid biodiversity assessment of specimen 'soups' (Crampton-Platt *et al.* 2015), it precludes the use of the assembled contigs for studies of population genetic variability, as with the current technology it is not possible to reconstruct the different haplotypes of a single species by means of differential assembly. Intraspecific differentiation was detectable to some extent when different haplotypes of the same species were assembled from different bulk samples, as seen in the LocL analyses that recover 177 different MMG-barcodes assigned to 76 morphological species. In contrast, a mapping approach allowed capturing the full genetic variability of the LocL samples. This approach required reference sequences representing the species in each LocL sample, which we based on the consensus sequence of the PCR-barcodes for each morphological species. However, the read-based approach was affected by low rates of read mapping to the barcode region in many LocL populations, limiting our analyses to just 40.8% of populations across the 10 samples. The sequencing depth for each LocL library was based on 20% of a flow cell corresponding to approximately 4,000,000 reads (Gómez-Rodríguez *et al.* 2015), of which only between 0.05% (MitoRL) and 0.1% (LocL) of the reads were identified as *cox1*-5' with our BLAST procedure. Data volume used here was not sufficient for an exhaustive analysis at the population level although it had proven useful for community level analyses (Gómez-Rodríguez *et al.* 2015). As a minimum, this study would probably have greatly benefitted from increased sequencing volume (e.g. 50% of a Miseq

flow cell), even if this does not guarantee a 2.5-fold increase in the number of populations that could be included in the dataset (i.e. 100% of the populations). Low sequencing depth (Futschik & Schlötterer 2010; Anderson, Skaug & Barshis 2014) and unequal contributions of DNA of each individual (Futschik & Schlötterer 2010; Ferretti, Ramos-Onsins & Pérez-Enciso 2013; Gautier *et al.* 2013) have been recognised as problems for calculating population parameters from pooled sequenced data. The efficacy of current mapping software may also present a challenge for these analyses (see Schlötterer *et al.* 2014; Crampton-Platt *et al.* 2015). However, in the present case, different read mappers gave similar results (see example in Appendix S5), indicating that robust information for population and community assessments can be gained at low sequencing depth.

While direct mapping of reads revealed numerous SNPs, genome assembly removed most of this variation by collapsing all haplotypes whose divergence was below ~4% (Appendix S1). In the absence of intraspecific variability in the MitoRL analysis, the two types of assemblers produced differences in only 4 of the 128 contigs (3.1%). Moreover, the consensus sequence (ignoring ambiguous sites) from both assemblers was identical to the PCR-barcode of the actual specimen used in MMG sequencing in 90.7% of the cases and most of the discrepant haplotypes differed by a single nucleotide only (Figure 1b). The greatest discrepancies were explained by low quality in the PCR-barcodes, and it is conceivable that other differences are also due to errors in the PCR-barcodes. The MitoRL assembly is a natural benchmark for assessing differences between Illumina and conventional Sanger sequencing without being compromised by intraspecific variability in the assembly. Our analysis shows that the MMG approach produces sequences of a quality equivalent to Sanger sequences even when closely related species are present in the sample.

When facing intraspecific variation, as exemplified by the LocL and CombL assemblies, the two assemblers resolved the variability in shotgun reads in slightly different ways, which produces ambiguities in 15.2% and 13.5% of the consensus sequences in either library, respectively, after re-assembly in Geneious, although most assemblies again differ in only one or two positions of the 655-bp barcode region (Figure 1a). Notably, the differences between the two assemblers do not increase from the intermediate level of diversity in the LocL assemblies to the more complex CombL, neither when considering the proportion of ambiguous contigs or the number of ambiguous sites. We thus conclude that the genome assemblers, even when using the two principal approaches for contig building (i.e. de Bruijn Graph approach in IDBA-UD or the Greedy Graph-based assembly in Newbler; see details of assembly approaches in Miller, Koren & Sutton 2010), reproducibly remove intraspecific genetic variation in the mixtures, while interspecific variation is maintained. This property supports the re-assembly of contigs from multiple assemblers in MMG analyses, as it is currently recommended to obtain longer contigs and to detect possible assembly errors (Crampton-Platt *et al.* 2016).

Ignoring the ambiguities from different assembly procedures, we then assessed the conflict between PCR-barcodes and MMG-barcodes, which revealed discrepancies in as many as 21.2% and 24% of contigs in LocL and CombL, but again these discrepancies affected mostly a single or a small number of nucleotides. In general, in the cases with a larger number of discrepancies (e.g. *Clytra quadripunctata*, see details in Appendix S1) there was a tendency towards lower quality in the PCR-barcodes, and thus errors in the Sanger data may also explain many of the single nucleotide differences. In addition, genuine haplotypes detected by the MMG assembly may be missing from the PCR-barcode dataset due to low barcoding success rate that correlates with the presence of MMG-specific SNPs (see Figure

2b), although PCR-barcodes were available for 88.9% of individuals and thus are unlikely to be missing many of the existing haplotypes. Failed barcodes or poor quality may also be caused by heteroplasmy that would lead to both mixed Sanger signal and the detection of multiple variants in Illumina sequencing differing primarily in synonymous 3rd codon positions, consistent with our findings of low Sanger sequencing success (Figure 2b).

An alternative explanation for the large number of 'discrepant' 3rd codon positions, and perhaps the greatest concern for the methodology, is the potential for generating artefactual haplotypes through the *in silico* recombination among true variants present in the mixture, i.e. the formation of chimeras of closely related haplotypes to produce sequences that are non-existent in the specimen mixture, even if the SNP variation itself is genuine. The errors from MMG assemblies therefore may be different from those in Sanger barcodes, as the latter result mainly from low-quality base calls or mixed signals, possibly resulting from heteroplasmy. The possibility of 3rd codon shuffling is supported because it is much less evident in the MitoRL assembly composed of one individual per species, and thus this form of chimera formation is evident only when very closely related haplotypes are present in the LocL and CombL libraries. At this stage we cannot resolve these artefactual haplotypes from true variation, e.g. due to heteroplasmy. A striking observation was that 3rd codon position 'discrepant' haplotypes were evident mainly in a small number of genera, including *Cryptocephalus*, *Altica*, *Aphthona,* and *Labidostomis*. It is well known that heteroplasmy is increased in certain lineages (e.g. Pons *et al.* 2011), and these genera may be affected by this phenomenon, but equally they may also constitute local assemblages of closely related haplotypes highly prone to *in silico* shuffling. The problem would be aggravated in genome skimming of other markers such as nuclear rRNA genes, which include variable and conserved regions that increase these effects of chimera formation, but if reference sequences

are available, the read mapping approach (against the most variable regions) remains fully effective.

Finally, due to the collapse of variation in the assembly, estimates of population genetic diversity are only possible with the unassembled reads. SNPs observed in reads mapped against the PCR-barcodes were used for estimating nucleotide diversity and showed a good correspondence with the scores obtained from the PCR-barcodes themselves. However, nucleotide diversity estimated from shotgun reads was significantly higher than estimated from PCR-barcodes, an issue that may be related to Illumina sequencing errors. Yet, even if absolute values differ, both estimates are highly correlated and this enables comparative studies among different populations or localities using shotgun reads for nucleotide diversity estimation. For example, a preliminary analysis of the climatic correlates of average nucleotide diversity in each locality suggests high nucleotide diversity across the ten locations is correlated with warm climate, cold winters and a marked wet season. Although ten localities is a limiting sample size for biological interpretations, this result is found using nucleotide diversity estimates from shotgun reads and PCR-barcodes alike (see Appendix S6). A species-per-species analysis may reveal additional insights into the structure of nucleotide diversity, or it will allow the estimation of correlates between species diversity and genetic diversity (see Vellend 2003; Baselga, Gómez-Rodríguez & Vogler 2015) from the same read mixtures. Notably, inflated values of PCR-barcode diversity were evident in some cases (see *Exosoma lusitanicum* in Figure 3), which may be due to the fact that population genetic parameters are calculated for differently defined entities. The read mapping approach relies on the match to haplotypes which may be grouped based on external criteria, such as membership in a morphologically delimited Linnaean species, or may be based on the sequence data themselves, e.g. the GMYC groups from PCR-barcodes. We observed that the

correspondence of genetic diversity from read mapping improves when GMYC groups are used, rather than the morphological species, due to the threshold for divergence used for read mapping (3% in this case) that may eliminate cryptic species exhibiting greater divergence from consideration. The genome assembly from mixed intraspecific variants has a similar effect, by collapsing only closely related haplotypes and thus producing a consensus for an entity that roughly corresponds to the species level. Thus, appropriate delimitation of species entities, along with sufficient sequencing depth and accurate read mapping, will be the cornerstones for successful analyses of intraspecific variation with MMG. The high quality of the Illumina data and the established potential for MMG below the species level now overcomes the final hurdle to a holistic analysis of community diversity, in addition to the assessment of species diversity, phylogenetic placement of taxa, their abundance, and now the accurate measurement of genetic diversity for most species in the community sample.

Acknowledgments

This work was supported by the NHM Biodiversity Initiative, a NHM/UCL PhD studentship (to ACP), a NERC Postdoctoral Fellowship NE/I021578/1 (to MJTNT), the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) (grants CGL2013-43350-P and CGL2016-76637-P) and Xunta de Galicia (postdoctoral fellowship POS-A/2012/052 to C.G.R.).

Data accessibility

GenBank accession numbers for Sanger derived sequences: KF134544 - KF134651 and KF652242 - KF656666 (*cox1*-5'). MMG-contigs are available from the Dryad Digital Repository: http://datadryad.org/resource/doi:10.5061/dryad.3rh21

References

Anderson, E., Skaug, H. & Barshis, D. (2014) Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology,* **23,** 502–512

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,* **24,** 873-882.

Crampton-Platt, A., Timmermans, M.J.T.N., Gimmel, M.L., Kutty, S.N., Cockerill, T.D., Khen, C.V. & Vogler, A.P. (2015) Soup to Tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution,* **32,** 2302-2316.

Crampton-Platt, A., Yu, D.W., Zhou, X. & Vogler, A.P. (2016) Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience,* **5:15**.

Charif, D. & Lobry, J.R. (2007) Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. *Structural approaches to sequence evolution: Molecules, networks, populations* (eds U. Bastolla, M. Porto, H.E. Roman & M. Vendruscolo), pp. 207-232. Springer Verlag, New York.

Ferretti, L., Ramos-Onsins, S.E. & Pérez-Enciso, M. (2013) Population genomics from pool sequencing. *Molecular Ecology,* **22,** 5561-5576.

Futschik, A. & Schlötterer, C. (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics,* **186,** 207–218.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C. & Estoup, A. (2013) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology,* **22,** 3766–3779.

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) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution,* **6,** 883-894.

Hellmann, I., Mang, Y., Gu, Z., Li, P., Vega, F.M., Clark, A.G. & Nielsen, R. (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Research,* **18,** 1020-1029.

Ji, Y.Q., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X.Y., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters,* **16,** 1245-1257.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C. & Schlotterer, C. (2011) PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *Plos One,* **6,** e15925.

Miller, J.R., Koren, S. & Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics,* **95,** 315-327.

Nagarajan, N. & Pop, M. (2013) Sequence assembly demystified. *Nature Reviews Genetics,* **14,** 157-167.

Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York:.

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

Petitpierre, E. (2000) *Coleoptera, Chrysomelidae I*. Museo Nacional de Ciencias Naturales, Madrid.

Pons, J., Barraclough, T.G., Gomez-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.

Pons, J., Fujisawa, T., Claridge, E.M., Savill, R.A., Barraclough, T.G. & Vogler, A.P. (2011) Deep mtDNA subdivision within Linnean species in an endemic radiation of tiger beetles from New Zealand (genus *Neocicindela*). *Molecular Phylogenetics and Evolution,* **59,** 251-262.

R Development Core Team (2015) *R: A language and environment for statistical computing. URL https://www.R-project.org/*. R Foundation for Statistical Computing, Vienna, Austria.

Schlötterer, C., Raymond Tobler, Kofler, R. & Nolte, V. (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics,* **15,** 749–763.

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.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A. & Zhou, X. (2014) Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research,* **42,** e166.

Vellend, M. (2003) Island biogeography of genes and species. *American Naturalist,* **162,** 358-365.

Warchalowski, A. (2003) *Chrysomelidae. The leaf-beetles of Europe and the Mediterranean area.* Natura optima dux Foundation, Warszawa.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. & Huang, Q. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience,* **2:4**

Zhu, Y., Bergland, A., González, J. & Petrov, D. (2012) Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. *Plos One,* **7(7),** e41901.

Supporting information

Appendix S1. Problematic contigs

Description of cases where two or more different contigs are assembled for the same species.

Appendix S2. Sequence quality and length in PCR-barcodes per locality

Boxplots showing the sequence length and the percentage of bases with high quality (phred score > 30) in the PCR-barcodes of the different study localities.

Appendix S3. Sequence quality in MMG-barcodes

Boxplots showing base quality (phred score) according to its position in the read. Data is also shown for the different libraries: 'mitogenome reference library' (MitoRL) and 'local library' (LocL).

Appendix S4. Relationship between nucleotide diversity estimated from PCR-barcodes and nucleotide diversity estimated from shotgun reads

Scatterplot of the relationship between nucleotide diversity estimated from PCR-barcodes and estimated from shotgun reads. Data is shown for Linnaean species (morphological species) and GMYC-groups.

Appendix S5. Comparison between read mappers

Example of the number of mapped reads by two different mapping software (Geneious and bbmap) in the ADS locality.

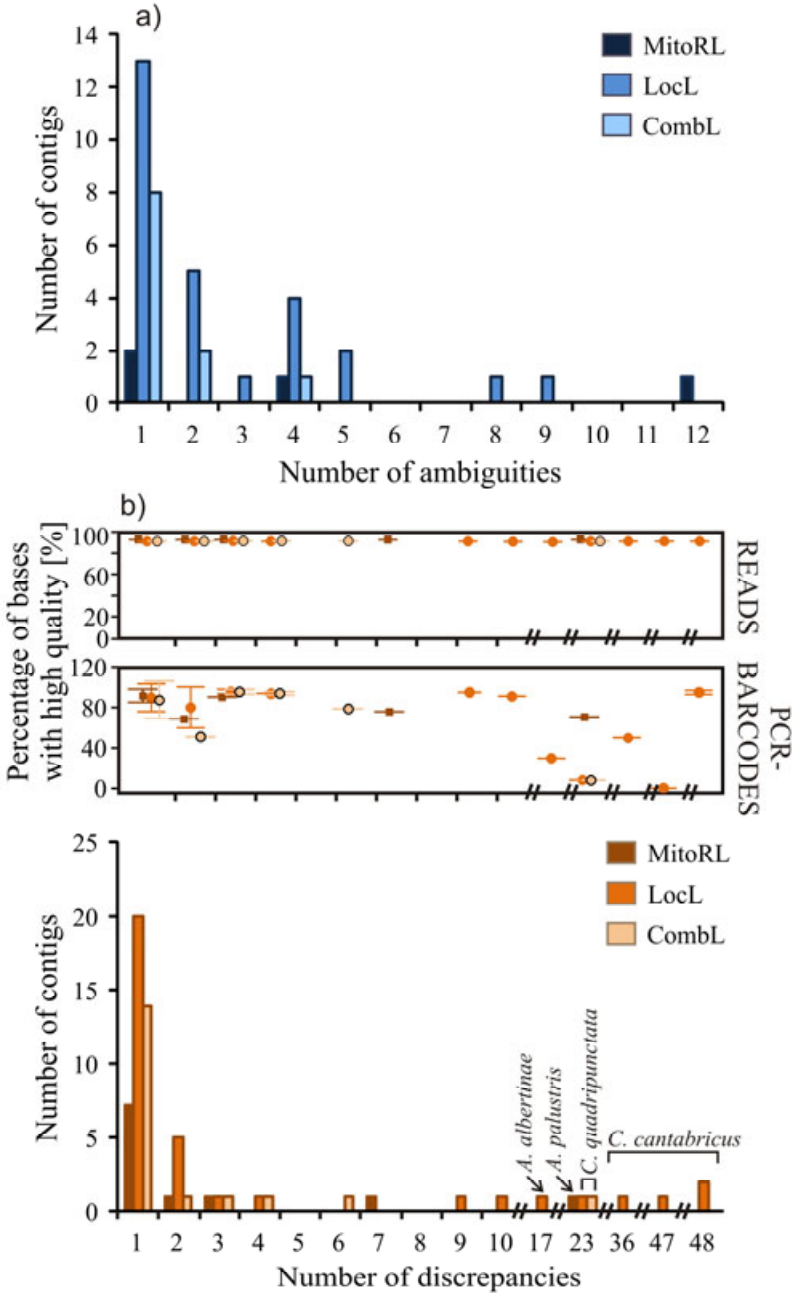
Appendix S6. Climatic drivers of average nucleotide diversity

Multiple regression model of the relationship between average nucleotide diversity and climatic factors.

Table 1. Proportion of contigs with ambiguities introduced by the re-assembly in Geneious. MMG-barcodes with ambiguous sites and average (± standard deviation) number of ambiguous sites in each set of mitochondrial genomes (MitoRL, LocL and CombL). Proportion of MMG-barcodes identical to a PCR-barcode (excluding the ambiguous sites due to re-assembly in Geneious) and number of discrepant bases.

	MitoRL	LocL	CombL
Proportion of ambiguous contigs	3.1%	15.2%	13.5%
#Ambiguous sites (mean ± SD)	4.5 ± 5.2	2.6 ± 2.2	1.4 ± 0.93
Proportion of identical contigs[1]	90.7%	78.8%	76.0%
#Discrepant bases (mean ± SD)	3.8 ± 6.1	7.8 ± 14.4	2.7 ± 5.1

[1]Comparisons could not be performed for 25 contigs because the corresponding PCR-barcodes were missing in the respective dataset (10 in MitoRL, 12 in LocL and 3 in CombL).



[1] PCR-barcodes quality is the mean and SD across barcodes, while quality of shotgun reads is a proportion value for the full set.

Figure 1. Histograms showing for each dataset (a) the number of contigs with a certain number of ambiguities introduced by the re-assembly in Geneious and (b) the number of contigs with a certain number of discrepancies with a PCR-barcode[1]. In cases with more than 10 discrepancies, the morphological species is indicated.

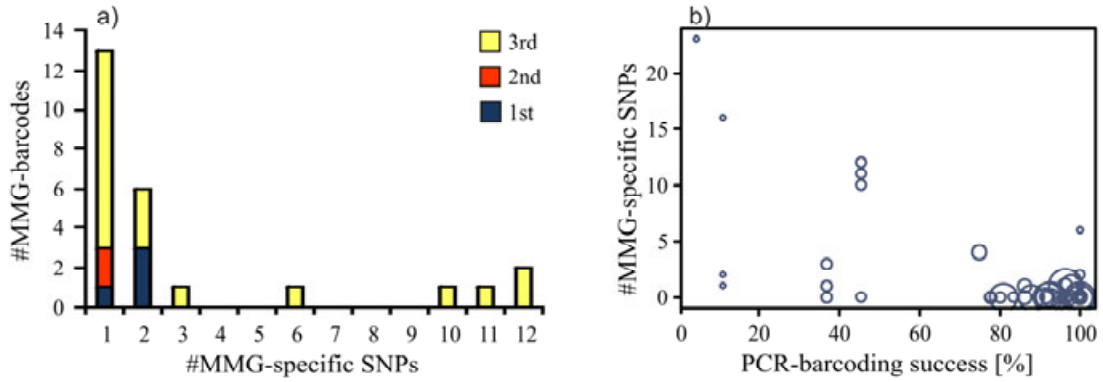


Figure 2. (a) Histogram showing the number of MMG-barcodes that introduce a certain number of MMG-specific SNPs, taking into account if they occurred in first, second or third codon position. Note: The total number of MMG-barcodes with specific SNPs is 21 and the total number of MMG-specific SNPs for the full dataset is 79. (b) Scatterplot of the relationship between the number of MMG-specific SNPs and the success of the Sanger-based barcoding sequencing used for computing the PCR-barcodes SNPs (i.e. number of PCR-barcodes/number of specimens for a given species in the full dataset). The size of the circles corresponds to the total number of sequences available for the species.

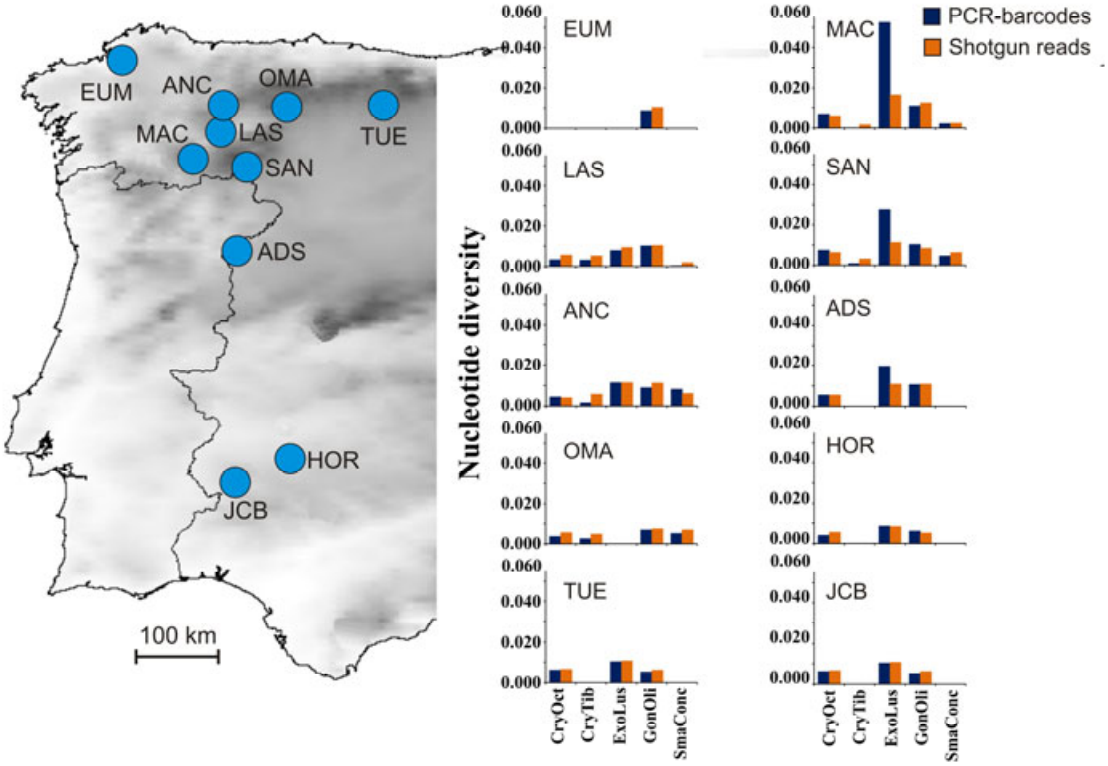


Figure 3. Nucleotide diversity estimated from PCR-barcodes and shotgun reads for the populations of the species present in more than 50% of the localities (CryOct: *Cryptocephalus octoguttatus*, CryTib: *Cryptocephalus tibialis*; ExoLus: *Exosoma lusitanicum*; GonOli: *Gonioctena olivacea*; SmaCon: *Smaragdina concolor*)

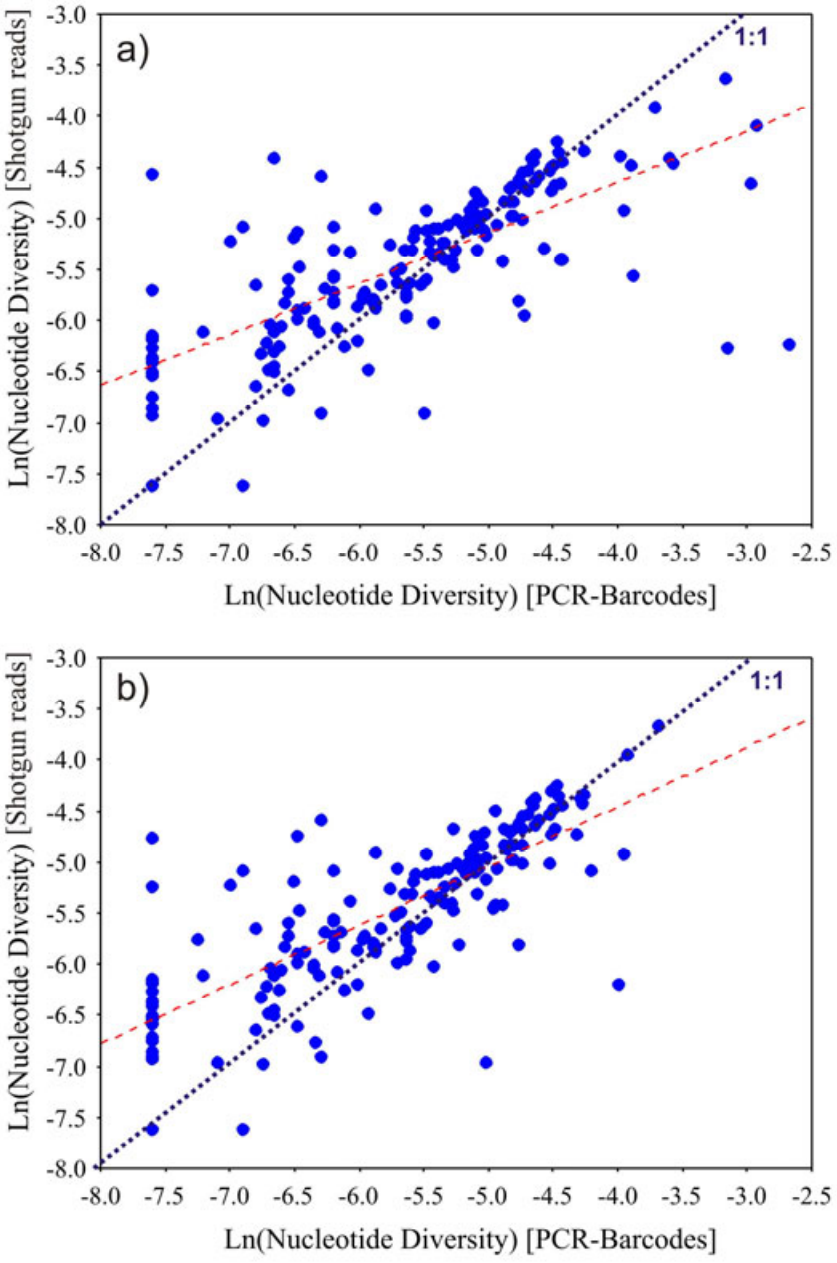


Figure 4. Scatterplot of the relationship between nucleotide diversity estimated for each (a) morphological species or (b) GMYC-group from PCR-barcodes and estimated from shotgun reads. Both the regression line (dashed red) and the 1:1 line (dashed blue) are shown. Values are natural-logarithm transformed.

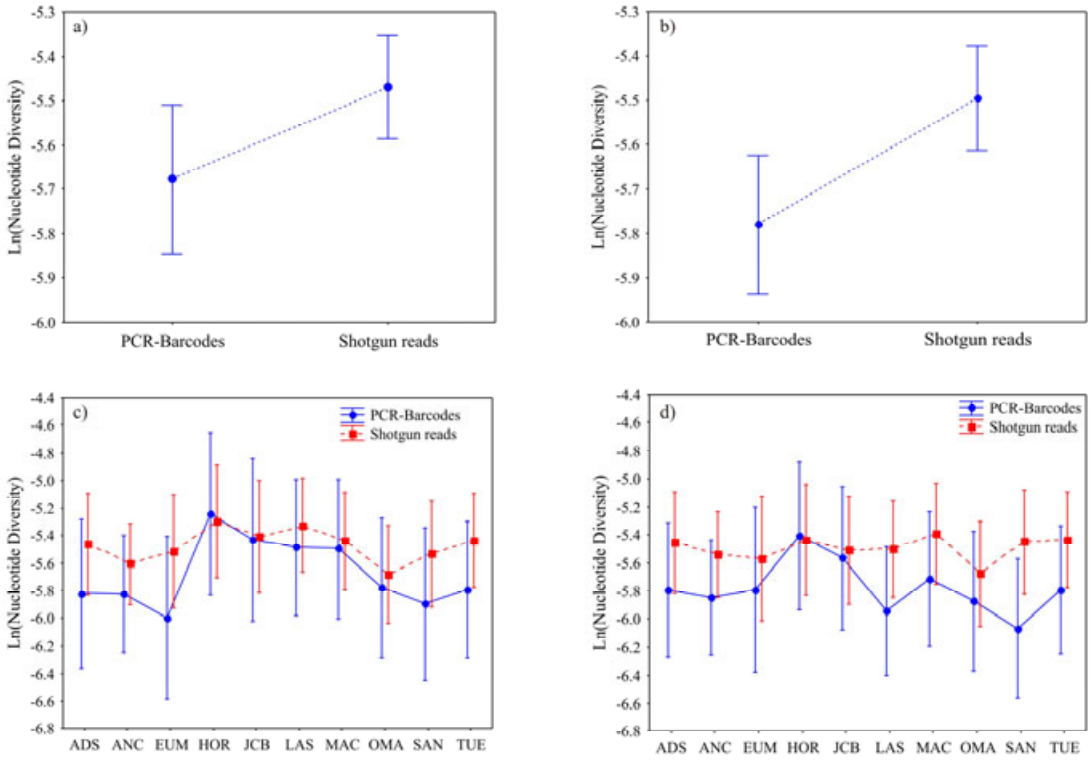


Figure 5. Mean value and 0.95 confidence intervals of nucleotide diversity computed from PCR-barcodes and from shotgun reads using morphological species (a, c) and GMYC-groups (b, d). Results are presented for the full dataset (a, b) and detailed for each locality (c, d).