Metazoa-level USCOs as markers in species delimitation and classification

Lars Dietz[1], Christoph Mayer[1], Eckart Stolle[1], Jonas Eberle[2], Bernhard Misof[1], Lars Podsiadlowski[1], Oliver Niehuis[3], and Dirk Ahrens[1]

[1]Zoologisches Forschungsinstitut und Museum Alexander Koenig
[2]Paris Lodron University of Salzburg
[3]University of Freiburg

October 16, 2023

Abstract

Metazoa-level Universal Single-Copy Orthologs (mzl-USCOs) are universally applicable markers for DNA taxonomy in animals which can replace or supplement single-gene barcodes. While previously mzl-USCOs from target enrichment data were shown to reliably distinguish species, here we tested whether USCOs are an evenly distributed, representative sample of a given metazoan genome and therefore able to cope with past hybridization events and incomplete lineage sorting. This is relevant for coalescent-based species delimitation approaches, which critically depend on the assumption that the investigated loci do not exhibit autocorrelation due to physical linkage. Based on 239 assessed chromosome-level assembled genomes, we confirmed that mzl-USCOs are genetically unlinked for practical purposes and a representative sample of a genome in terms of reciprocal distances between USCOs on a chromosome and of distribution across chromosomes. We tested the suitability of mzl-USCOs extracted from genomes for species delimitation and phylogeny in four case studies: Anopheles mosquitos, Drosophila fruit flies, Heliconius butterflies, and Darwin's finches. In almost all instances, USCOs allowed delineating species and yielded phylogenies that correspond to those generated from whole genome data. Our phylogenetic analyses demonstrate that USCOs may complement single-gene DNA barcodes and provide more accurate taxonomic inferences. Combining USCOs from sources that used different versions of ortholog reference libraries to infer marker orthology may be challenging and at times impact taxonomic conclusions. However, we expect this problem to become less severe as the rapidly growing number of reference genomes provides a better representation of the number and diversity of organismic lineages.

Metazoa-level USCOs as markers in species delimitation and classification

Lars Dietz[1], Christoph Mayer[1], Eckart Stolle[1], Jonas Eberle[1,2], Bernhard Misof[1,3], Lars Podsiadlowski[1], Oliver Niehuis[4], Dirk Ahrens*[1]

[1] Museum A. Koenig, Leibniz Institute for the Analysis of Biodiversity Change, Bonn, Germany

[2] Paris-Lodron-University, Salzburg, Austria.

[3] Rheinische Friedrich-Wilhelms-Universität Bonn, Germany.

[4] Abt. Evolutionsbiologie und Ökologie, Institut für Biologie I, Albert-Ludwigs-Universität Freiburg, Germany.

*Corresponding author; Email: ahrens.dirk_col@gmx.de.

Abstract

Metazoa-level Universal Single-Copy Orthologs (mzl-USCOs) are universally applicable markers for DNA taxonomy in animals which can replace or supplement single-gene barcodes. While previously mzl-USCOs from target enrichment data were shown to reliably distinguish species, here we tested whether USCOs are an evenly distributed, representative sample of a given metazoan genome and therefore able to cope with past hybridization events and incomplete lineage sorting. This is relevant for coalescent-based species delimitation approaches, which critically depend on the assumption that the investigated loci do not exhibit autocorrelation due to physical linkage. Based on 239 assessed chromosome-level assembled genomes, we confirmed that mzl-USCOs are genetically unlinked for practical purposes and a representative sample of a genome in terms of reciprocal distances between USCOs on a chromosome and of distribution across chromosomes. We tested the suitability of mzl-USCOs extracted from genomes for species delimitation and phylogeny in four case studies: *Anopheles* mosquitos,*Drosophila* fruit flies, *Heliconius* butterflies, and Darwin's finches. In almost all instances, USCOs allowed delineating species and yielded phylogenies that correspond to those generated from whole genome data. Our phylogenetic analyses demonstrate that USCOs may complement single-gene DNA barcodes and provide more accurate taxonomic inferences. Combining USCOs from sources that used different versions of ortholog reference libraries to infer marker orthology may be challenging and at times impact taxonomic conclusions. However, we expect this problem to become less severe as the rapidly growing number of reference genomes provides a better representation of the number and diversity of organismic lineages.

Keywords

WGS, USCOs, Metazoa, genome assembly, systematics, DNA taxonomy.

Running title:

Metazoa-level USCOs for systematics

Introduction

During the past two decades, DNA-based approaches have increased the quality and reproducibility of species delimitation and identification (Ahrens, 2023). Standardized and automated species recognition using DNA has made it easy to link taxonomic information with diverse biological questions and applied research aspects (e. g., rapid assessment of biodiversity; Yu et al., 2012). Species delimitation and identification of animals are often based on information from a single mitochondrial gene, cytochrome oxidase I (COI) (Hebert et al., 2003; Fontaneto et al., 2015). Such single-marker reliance can lead to errors due to extrachromosomal inheritance, incomplete lineage sorting, sex-biased dispersal, asymmetrical introgression, and *Wolbachia* - mediated genetic sweeps of the marker (Funk & Omland, 2003; Ballard & Whitlock, 2004). At the same time, species delimitation approaches using nuclear-encoded markers have considerably improved in accuracy, allowing to complement the currently established single-gene barcoding approach (Dowton et al., 2014; Eberle et al., 2020; Gueuning et al., 2020; Prebus, 2021; Erikson et al., 2021; Dietz et al., 2023).

Besides mitochondrial genes, a variety of conserved nuclear markers have been used for species delimitation in different phylogenetic groups of Metazoa, such as nuclear ribosomal RNA genes (Lebonah et al., 2014; Chen et al., 2017; Krehenwinkel et al., 2019) and various housekeeping genes (Joshi et al., 2022). Furthermore, restriction site-associated DNA sequences (RADseq) (Baird et al., 2008; Pante et al., 2015; Herrera & Shank, 2016) and ultra-conserved elements (UCE) linked to more rapidly evolving flanking regions (Faircloth et al., 2012; Bejerano et al., 2004; Ješovnik et al., 2017; Zarza et al., 2018; Gueuning et al., 2020; Prebus, 2021) were used. However, these nuclear marker systems can hardly be applied universally across animals, either because they insufficiently capture intraspecific variation or because they do not provide orthologous loci across distantly related taxa (Pierce, 2019; Eberle et al., 2020).

Recently, Metazoa-level Universal Single Copy Orthologs (USCOs) have been proposed as a universal marker set for species-level DNA taxonomy of animals as an extension and improvement of conventional DNA barcoding (Eberle et al., 2020). USCOs are defined as protein-coding genes that are present and single-copy in at least 90% of the species within the available genomes of a given taxonomic group. They have originally

been developed to benchmark the quality of genome assemblies ("BUSCO", Simão et al., 2015). However, they also proved to be highly informative for addressing phylogenomic questions (Waterhouse et al., 2018; Fernández et al., 2018; Zhang et al., 2019; Stolle et al., 2022). This insight has led to the development of a recently published automated software pipeline that extracts USCOs from genome assemblies and generates phylogenies from the extracted sequence data (Sahbou et al., 2022). Finally, Metazoa-level USCOs (mzl-USCOs) have been shown to allow distinguishing highly similar morphospecies (even when *COI* was unable to do so) and reliably estimating their phylogenetic relationships in several clades of arthropods and vertebrates (Dietz et al., 2023).

What has remained unclear is whether mzl-USCOs can be considered a genetically unlinked representative sample of a species' genome, which is a prerequisite for USCOs being reliable and useful in coalescent-based phylogenetic analyses and applications. Knowledge of the spatial distribution and physical linkage of mzl-USCOs is hence fundamental to assess whether these markers are indeed as suitable for delimiting species with coalescent-based approaches as currently assumed. We here study the two parameters "spatial distribution" and "physical linkage" by extracting USCOs from published whole genomes assembled to chromosome-level (WG) of various species of Metazoa and analyzing the physical distances between USCOs and their distribution across chromosomes. Furthermore, using unassembled reads from whole genome sequencing (WGS) datasets of four metazoan lineages (i.e,*Anopheles* mosquitos, *Drosophila* fruit flies,*Heliconius* butterflies, and Darwin's finches), we assess to what extent phylogenetic analysis of the extracted mzl-USCOs provides results consistent with those of previous studies that used more extensive sets of markers from the same genomes.

Materials and methods

Distribution and linkage patterns of mzl-USCOs

All available (as of July 2021) metazoan genomes assembled to chromosome level were downloaded from NCBI RefSeq (O'Leary et al. 2016; see Table S1). Contigs not assembled to chromosome level were excluded with a custom Perl script (Supplementary Material). The genomic nucleotide sequences were then searched for mzl-USCOs with the program BUSCO v. 4.0.6 (Manni et al., 2021) using the program's default parameters for genomic data and the metazoa_odb10 dataset from the BUSCO website. Mzl-USCOs were first sorted according to a) the chromosome on which they were located and b) their start position on a given chromosome. Next, we calculated the distances between start positions of consecutive mzl-USCOs on the same chromosome as predicted by BUSCO. Distances were recorded as absolute distances (nucleotides) and as normalized distances, with the latter being calculated by dividing the absolute distance by the genome size of the respective species. In a second step, both absolute and normalized distance values (d) were binned into ten categories based on log(d) values. For normalized distances, these were log(d) $<$ -6, -6 [?] log(d) $<$ -5.5, ..., -2.5 [?] log (d) $<$ -2, log(d) [?] -2. For absolute distances, 9 was added to the logarithmic range of each category, as the average size of analyzed genomes was about 10^9. For each taxon, the proportion of distances in each bin was calculated with a custom Perl script (Supplementary Material). Based on the proportions across all taxa, we conducted principal component analyses (PCA) for absolute and normalized distances, respectively, in PAST v. 4.03 (Hammer et al., 2001). The resulting scores of all taxa for the first and the second PC axes were mapped on the phylogenetic tree of the taxa (see below) with MESQUITE v. 3.51 (Maddison & Maddison, 2018). Furthermore, we analyzed the distribution of distances between start positions of adjacent mzl-USCOs to assess whether mzl-USCOs tend to cluster spatially more than a randomly chosen identical number of protein-coding genes would do. To achieve this, we downloaded the official gene set of coding sequences (CDS) for each genome and, using a custom Perl script (Supplementary Material), randomly selected the same number of protein-coding genes as the number of mzl-USCOs found in the respective taxon. This random drawing was repeated 10,000 times for each genome, and for each replicate, the median distance (absolute and normalized, separately) between neighboring genes was calculated. To infer whether mzl-USCOs cluster significantly more than randomly chosen genes, we counted for each taxon the number of replicates in which the median distance between neighboring protein-coding genes was lower than the median distance between neighboring mzl-USCOs.

We used a custom Perl script (Supplementary Material) to estimate the adjusted evenness of the distribution of mzl-USCOs between chromosomes in each taxon according to the formula $e^{(H/S)}$, where H is the Shannon-Wiener entropy (Heip et al. 1998) of the distribution and S the number of chromosomes. While in ecology species which are not present in a sample are not considered in the calculation of evenness (Heip et al. 1998), here, S includes all chromosomes, even those with no mzl-USCOs, representing thus an "adjusted evenness". For comparison, we also calculated the adjusted evenness of the number of all protein-coding genes on the chromosomes, as well as that of the length of the chromosomes in base pairs. Additionally, we used another Perl script (Supplementary Material) to (i) conduct a chi-square test in search of significant deviations of the distribution of mzl-USCOs between chromosomes from a distribution proportional to chromosome length, and (ii) for significant deviations from the distribution of protein-coding genes in general. To assess the degree by which chromosomal linkage between mzl-USCOs is phylogenetically conserved across taxa, we calculated with the aid of a custom Perl script (Supplementary Material) the proportion of taxa in which a given pair of mzl-USCOs was found to be co-located on a chromosome.

Phylogenetic analysis of metazoan genomes

We performed phylogenetic analyses with the Metazoa-level USCO nucleotide sequences to assess their reliability in recovering phylogenies and classifications. To this end, we analyzed all orthologous nucleotide sequences of each mzl-USCO gene from all genome assemblies in which more than half of the loci were recovered as being complete and single-copy. Nucleotide and amino acid sequences of USCOs were taken from the output of the BUSCO software. Amino acid sequences were aligned with MAFFT v. 7.305b (Katoh & Standley, 2013) using the L-INS-I algorithm. Poorly aligned regions were identified and removed from the amino acid alignments with ALISCORE v. 2.0 (Misof & Misof, 2009; Kuck et al., 2010) and ALICUT v. 2.31 (available from:*https://github*.com/PatrickKueck/AliCUT), and outlier sequences were identified and removed with OliInSeq v. 0.9.3 (*https://github*.com/cmayer/OliInSeq). Multiple nucleotide sequence alignments based on the amino-acid alignments were inferred with pal2nal v. 14.1 (Suyama et al., 2006), and all third codon positions were excluded with a custom Perl script (Supplementary Material). Maximum-likelihood analyses were performed with IQ-TREE v. 2.1.2 (Minh et al., 2020) using multiple sequence alignments of individual genes and concatenated multiple sequence alignments of all genes, respectively, and analyzing amino-acid sequence data or nucleotide sequence data with third codon positions removed. For both the concatenated nucleotide dataset and the concatenated amino-acid dataset, the best fitting substitution model and partitioning scheme were inferred with ModelFinder (Chernomor et al., 2016; Kalyaanamoorthy et al., 2017) and PartitionFinder (Lanfear et al. 2014) as implemented in IQ-TREE using the full list of models and the IQ-TREE option -m MFP+MERGE. Data blocks in the partition merging steps were the USCO genes. For analyzing the nucleotide dataset, we applied the inferred substitution model and partitioning scheme and performed 50 replicate maximum likelihood tree searches from random starting trees. We performed a single maximum likelihood tree search when analyzing the amino-acid dataset, as performing replicates would have been computationally unreasonably expensive with respect to the expected benefit. Branch support was estimated from 1,000 ultrafast bootstrap replicates (UFBoot, Hoang et al., 2018) as well as approximate likelihood ratio tests (aLRT) using nearest neighbor interchange (NNI) as tree rearrangement method. The tree with the highest likelihood was then chosen among all replicates. The individual gene trees were further used for a coalescent-based tree analysis with ASTRAL v. 5.6.1 (Zhang et al., 2018) applying the program's default settings.

Sequence overlap in multiple sequence alignments was examined using the concatenated alignment containing all taxa. We calculated with a custom script (Supplementary Material) the overlap for each pair of individuals, defined as the number of alignment positions with data in both individuals, divided by the number of alignment positions with data in at least one of the two individuals.

Case studies using mzl-USCOs from whole genome sequences: data extraction

To investigate the usefulness of mzl-USCOs to resolve species boundaries in recent radiations and to assess the practicability of the data extraction and assembly pipelines that we developed and applied, we analyzed mzl-USCOs obtained from raw reads of WGS data sets of species of four well-studied radiations: *Heliconius*

butterflies, Darwin's finches, *Anopheles* mosquitoes, and *Drosophila* fruit flies (Table 1; Table S2). Each of these four case studies included multiple specimens of each involved species. The WGS raw reads were downloaded from NCBI. To assemble genomic raw reads to individual USCOs, we extracted mzl-USCOs (Eberle et al., 2020; Dietz et al., 2023) from one selected fully assembled and annotated genome per study group (Table 1) and then used each gene to map the raw reads of each individual onto it (see below).

One rationale for prioritizing USCOs over other genomic nuclear markers (Eberle et al., 2020) is that they allow us to build a comprehensive database in which USCO data referring to different taxonomic groups are stored. This data can be obtained at different times (i.e., with different ortholog sets) and with different data extraction approaches (e.g., DNA target enrichment, WGS; Eberle et al., 2020; Dietz et al., 2023). To evaluate the data yield and ability to resolve species-level relationships with different extraction approaches and genome reference systems (Zdobnov et al., 2017; Kriventseva et al., 2019), mzl-USCO nucleotide sequences were extracted from the reference genomes of the four case studies with three different methods. In the first approach, exonic nucleotide sequences of USCOs were extracted from the assembled genomes with the BUSCO program v. 4.0.6 (Simao et al., 2015; Manni et al., 2021) using the genome mode and the Metazoa dataset from OrthoDB v. 10 (Kriventseva et al., 2019), in the following text referred to as BUSCO data set. In the second approach, Orthograph v. 0.7.1 (Petersen et al., 2017) was used with HMMs from OrthoDB v. 9 (Zdobnov et al., 2017), in the following text referred to as OrthoDB v. 9 data set. For this, we downloaded the official gene sets (OGS) of all species included in the Metazoa OrthoDB v. 9 dataset from the OrthoDB site and the HMMs and information files for that dataset from the BUSCO website (*https://busco*-archive.ezlab.org/v3/). We used these to create an SQLite database with Orthograph, which was used together with the HMMs from BUSCO to extract the respective USCO nucleotide sequences from the coding sequences (CDS) of each taxon's OGS using Orthograph with its default setting. Our methodology was thus identical to the one used in approach A2 by Dietz et al. (2023) to assemble USCO raw reads retrieved via DNA target enrichment. The third approach was identical to the second with the one exception that we used OrthoDB v. 10 (*https://busco*.ezlab.org/busco_v4_data.html) instead of OrthoDB v. 9, in the following text referred to as OrthoDB v. 10 data set.

In all three approaches, nucleotide sequences of single-copy USCOs extracted from the respective genome were used as a reference against which raw reads were mapped with bwa v. 2.1 (Li & Durbin, 2009) using the software's default setting, except that the minimum seed length was set to 30. Diploid consensus sequences, in which heterozygous sites were represented by an IUPAC ambiguity code, were generated with samtools v. 1.10 (Li et al., 2009) and bcftools v. 1.10.2 (*https://github*.com/samtools/bcftools). As the nucleotide sequences were aligned to the reference sequence by bwa, no further alignment was necessary. Phylogenetic analyses were done with IQ-TREE v. 2.1.2 (Minh et al., 2020) using a supermatrix of the concatenated nucleotide sequences (positions with missing data or gaps were not removed at this point). The substitution model and partitioning schemes were chosen as described above, and 50 replicate analyses were performed for each dataset. With the same method, we performed phylogenetic analyses based on the nucleotide sequence alignment of each individual USCO and used the resulting trees as input for a multispecies coalescent analysis with ASTRAL v. 5.6.1 (Zhang et al., 2018). All trees were rooted with the outgroup taxa used in the respective original studies from which the data were taken (Table 1).

Case studies: analysis of USCO nucleotide sequence variation

To infer nucleotide sequence variation and to perform phenetic analyses on the extracted USCO data, such as Bayesian clustering or non-metric multidimensional scaling (NMDS), SNPs were extracted from the USCO nucleotide alignments of the four case studies obtained as described in the previous section. SNP sites were extracted from the multiple nucleotide sequence alignments of diploid consensus sequences of each USCO with the software SNP-sites (Page et al., 2016), excluding low-quality sites masked by the software bcftools with lowercase letters. For this purpose, all outgroup taxa were excluded from the multiple nucleotide sequence alignments. In the dataset of Darwin's finches, we additionally excluded the divergent ingroup genus *Certhidea* .

SNPs were filtered in three steps using custom Perl scripts as done by Dietz et al. (2023). First, all non-

informative SNP sites (i.e., those in which all individuals except one had the same allele) were removed (removegaps_snp_inf_d.pl); second, SNP sites with missing data or with gaps in more than 50% of individuals were removed (removegaps_d.pl); and third, only SNP sites in a given gene present in the largest number of individuals in the respective gene's nucleotide sequence alignment were kept (removegaps_snp_d.pl).

Based on the extracted SNP sites, we conducted a population structure analysis by clustering SNPs with the software STRUCTURE v. 2.3.4 (Pritchard et al., 2000) using an MCMC chain length of 50,000 (burn-in of 20,000) and a range of values for the number of ancestral populations (K) from 1 to 10. For each K value, the analysis was repeated ten times and the result with the highest likelihood was chosen. Additionally, we performed NMDS in two dimensions with the software PAST v. 4.03 (Hammer et al., 2001) using only biallelic SNPs. With a custom script (snp-pca_d.pl), genotypes homozygous for the majority allele were coded as 0, those homozygous for the rarer allele as 2 and heterozygous genotypes as 1. NMDS was then performed based on Euclidean distances. We repeated the analysis at least ten times to reduce the risk of reporting results with only locally optimal parameters and chose the results with the lowest stress value (i.e., those with the best fit to the data).

To assess whether different data extraction methods can be combined, we created multiple sequence alignments including the mzl-USCOs extracted with the BUSCO software and those extracted with Orthograph using OrthoDB v. 9 and v. 10. We combined the sequences of each gene obtained from the three approaches in a single dataset for each of the four case studies, using only those 580 genes classified as mzl-USCOs in both versions of OrthoDB. In this dataset, every specimen was consequently represented three times. We aligned the amino acid sequences from the three approaches with MAFFT v. 7.543 (Katoh & Standley, 2013) and inferred the corresponding nucleotide sequence alignments with pal2nal v. 14.1 (Suyama et al., 2006). However, we only used the nucleotide sequence alignments for phylogenetic inference, as they provide more phylogenetic signal for closely related taxa than the amino acid sequence alignments. The nucleotide sequence alignments were concatenated and the resulting supermatrix was phylogenetically analyzed with IQ-TREE as described above, except that only one analysis was conducted per dataset. Trees were rooted at the midpoint. We tested the effect of removing alignment positions with missing data and gaps on the phylogenetic results by removing alignment sites that have missing data or gaps in at least one individual and performing a phylogenetic analysis on the reduced dataset as described above. Analogously, using both data including or excluding positions with missing data or gaps, we performed phylogenetic analyses on multiple nucleotide sequence alignments of each gene and inferred a coalescent-based tree with ASTRAL as described above. Furthermore, we visually inspected the entire multiple nucleotide sequence alignment of each gene in all four case studies and removed alignment regions which contained strongly divergent sequences between extraction approaches and manually corrected obvious misalignments in AliView (Larsson, 2014). These corrected data were then used for coalescent-based and concatenation-based phylogenetic analyses as described above. Additionally, we generated versions of these corrected multiple nucleotide sequence alignments in which sites containing missing data and gaps were removed as described above.

Case studies: species delimitation

To test the performance of species delimitation algorithms applied on different USCO datasets, we delineated species with tr2 (Fujisawa et al., 2016) and SODA v. 1.0.2 (Rabiee et al., 2020) based on the results of the three different extraction approaches in all four case studies, using the programs' default parameters and not providing a guide tree. Tr2 conducts species delimitation based on the topological variation between rooted gene trees using the distribution of triplets (i.e., topologies with three individuals). As input, we used the gene trees generated by IQ-TREE, with the gene trees being re-rooted with nw_reroot (part of Newick Utilities 1.6.0; Junier & Zdobnov, 2010) on the outgroup of the respective study case. We included only trees that contained all specimens, as gene trees lacking specimens cannot be handled by the program. SODA uses the distribution of topologies of quartets of individuals in unrooted gene trees to infer species boundaries (Rabiee et al., 2020). We used all gene trees as input for SODA.

Results

Spatial distribution and potential linkage patterns of mzl-USCOs in genomes

We extracted mzl-USCOs from chromosome-level assembled genomes of 239 species of Metazoa, covering almost all major lineages of Protostomia and Deuterostomia. As expected, we found that the large majority of the mzl-USCOs were consistently present in most investigated species, and pairwise aligned nucleotide or amino acid sequences of mzl-USCOs from different species were found to overlap in the multiple sequence alignment of each gene to a high degree (Figure S1). The median distance between neighboring mzl-USCOs on a chromosome was on average 742,876 bp (+/- 607,054 bp SD). Considering all possible pairs of mzl-USCOs, we found that in the vast majority of genomes the two mzl-USCOs in a pair were located on different chromosomes (Fig. 1). Specifically, we found only 1.3% of the analyzed pairs of mzl-USCOs to be located on the same chromosome in more than 50% of the analyzed species. Only 0.2% of all analyzed pairs of mzl-USCOs were found on the same chromosome in more than 75% of the analyzed species. Looking at these latter pairs in more detail, we found the two mzl-USCOs in each pair to be spatially separated on average by a mean distance over all taxa of 11.6 Mbp (+/- 6.1 Mbp SD) on a given chromosome, with the spatial separation differing widely between taxa (average standard deviation 18.2 Mbp, +/- 8.8 Mbp SD).

While these data imply that mzl-USCOs can be regarded as genetically largely unlinked in practical applications, mzl-USCOs show a slight tendency to cluster compared to randomly chosen protein-coding genes. Specifically, we found physical distances between neighboring mzl-USCOs normalized by genome size to be consistently slightly lower than expected by chance when compared with distances from the same number of randomly chosen protein-coding genes. In all but three taxa, the median distance, both absolute and normalized by genome size, was lower in the USCO data than the median in the randomly chosen protein-coding genes (inferred from 10,000 simulations in each taxon; Fig. 2). In 195 taxa (82% of all investigated taxa), the difference was statistically significant ($p < 0.05$). On average, the median absolute distance was lower by 106,062 +/− 91,451 bp in the real data, the normalized distance by $9.91*10^{-5}$ +/− $6.57*10^{-5}$ of genome size (15.77 +/− 9.4 %). The extent to which mzl-USCOs cluster more than randomly chosen genes tends to be larger in arthropods than in vertebrates (Table S1).

We found the distribution of absolute distances (in nucleotides) between neighboring mzl-USCOs on chromosomes to be highly correlated with the taxon's genome size (correlation of median distance with genome size: $r = 0.9714$, $p < 0.001$). When binning absolute distances in eleven categories and using a PCA to visualize the degree of similarity between taxa in their distance values (plot not shown), separation of taxa along the first axis (which explained 71% of the total variance) strongly correlated with the logarithm of the taxon's genome size ($r = -0.9818$, $p < 0.001$). We focused in the present investigation on the conspicuous patterns found in normalized distances (nucleotides divided by genome size), as this metric was less confounded by the organism's genome size: correlation of median normalized distance with genome size was -0.17201 ($p = 0.008$). When binning normalized distances between neighboring mzl-USCOs on chromosomes in eleven categories and using a PCA to visualize the degree of similarity (Fig 3b), we found the clustering of taxa in some instances to correspond noticeably with high systematic units, such as Insecta (red triangles), teleost fishes (gray dots), birds (black squares), and mammals (black triangles; Fig 3).

The adjusted evenness of the distribution of mzl-USCOs between chromosomes ranged between 0.58 and 0.99 (mean 0.87 +/− 0.09). It tends to be especially low in birds and especially high in teleost fish (Table S1). It is highly correlated with both the evenness of chromosome length ($r = 0.83$, $p = 6.26 * 10^{-61}$) and especially that of the distribution of all protein-coding genes ($r = 0.94$, $p = 1.98 * 10^{-110}$).

In many taxa, our chi-square test showed significant deviations of USCO distribution from the distribution of chromosome lengths (Table S1). In 215 taxa (90% of all investigated taxa), the chi-square test showed a statistically significant ($p < 0.05$) deviation without correction for multiple test, and in 153 of the taxa (64%), the test result remained significant after Bonferroni correction for multiple tests. The deviation tended to be particularly high in birds and particularly low in teleost fish. The chi-square test showed that the deviation from the distribution of all protein-coding genes was significant in 170 taxa (71%), but in only 43 of these taxa (18%) it remained so after Bonferroni correction. A correlation with phylogenetic placement of the taxa was less obvious than in the comparison with chromosome length.

To assess whether the phylogenetic signal contained in mzl-USCOs is sufficient to infer the phylogenetic relationships of the investigated taxa, we used the extracted mzl-USCOs of the 239 species of Metazoa for phylogenetic analyses. The inferred phylogenetic trees based on a supermatrix of amino acid sequences (Fig. S2) were largely consistent with the respective current state of the art phylogenetic hypotheses (e.g., Laumer et al., 2019; Irisarri et al., 2017; Esselstyn et al., 2017). Discrepancies occurred in a few rapid radiations. For example, in the USCO-derived phylogenies of Neoaves we found hummingbirds to be more closely related to passerines than to falcons and parrots, contradicting results from phylogenomic studies of Jarvis et al. (2014) and Prum et al. (2015). Such discrepancies were also found in multi-species coalescent-based trees obtained from analyzing amino acid data (Fig. S4), which had overall low support values, however. Both supermatrix- and coalescent-based phylogenetic inferences based on nucleotide sequence data using codon positions 1 and 2 (Fig. S3, S5) resulted in some highly questionable phylogenetic estimates, such as a non-monophyly of Arthropoda.

Systematics with mzl-USCOs from whole genomes: data recovery of different extraction methods

Reference genomes of each of the four study groups contained at least 90% of the mzl-USCOs with exactly one copy. We found no consistent differences in the number of detected mzl-USCOs across the analyzed individuals (Figure S8) irrespective of what software we used to identify mzl-USCOs and their copy numbers. In all four study groups, all mzl-USCOs present in the reference genomes were recovered in at least some target individuals, and in all specimens, except some Darwin's finches, the majority of mzl-USCOs was recovered (Figure S8).

The concatenated multiple nucleotide sequence alignments of mzl-USCOs extracted with the BUSCO software were more than a million sites long; the corresponding supermatrices of USCO nucleotide sequences extracted with Orthograph were on average about 30% shorter (Table 2). The Orthograph/bwa-based approach was found to consistently miss some mzl-USCOs in some specimens: the number of mzl-USCOs recovered across all specimens proved to be consistently lower when using Orthograph for target gene identification than when using BUSCO (Figure S8). Total alignment completeness at the nucleotide level exceeded 90% in all study groups, except in Darwin's finches with a completeness of 45–52%. Alignment completeness of Orthograph-based datasets was slightly lower than of BUSCO-based datasets (Figure S9). The number of SNP sites was higher than 5,000 in all studied taxonomic groups, except in Darwin's finches. The number was generally much smaller in the Orthograph-derived datasets than in the BUSCO-derived ones (Table 2).

Case studies: phylogeny and nucleotide sequence variation of mzl-USCOs

In all four case studies, most interspecific but also many intraspecific nodes of the inferred phylogenetic trees had high (i.e., > 90) branch support and showed few topological differences between datasets obtained by different USCO extraction methods (Fig. 4; Figures S10–13). We detected few topological differences between trees inferred from concatenation supermatrices and trees inferred by using a multispecies coalescent approach on gene trees.

In the *Anopheles gambiae* complex, the topology of interspecific nodes in all USCO-based trees (Figure S10) was identical to the published one inferred by applying the maximum likelihood optimality criterion on aligned WGS data (Fontaine et al., 2015), except that we found neither *A. gambiae* nor *A. coluzzii* to be monophyletic. However, the topology of Fontaine et al. (2015) differed from the species tree inferred by the same authors from the X chromosome data only. According to Fontaine et al. (2015), the X chromosome-derived tree more likely represents the true phylogeny of the group, because the remainder of the genome exhibits extensive signatures of introgression. The USCO-derived topology suggested the monophyly of all species except*A. gambiae* and *A. coluzzii* . Monophyly of the latter was also not found in the study by Fontaine et al. (2015) when analyzing SNPs extracted from WGS data applying the neighbor-joining tree inference method. Only in the tree obtained from concatenated data containing mzl-USCOs extracted with Orthograph/OrthoDB v. 9, both species were found to be reciprocally monophyletic. NMDS plots that visualized the similarity in SNPs showed nearly all species as clearly distinct clusters irrespective of the applied USCO extraction method (Figure S14). The only exceptions were *A. gambiae* and *A. coluzzii* ,

forming together a single cluster. Our model-based clustering analyses using STRUCTURE also showed all species with the exception of*A. coluzzii* and *A. gambiae* as separate clusters with some levels of admixture (Figure S11; Supplementary Text).

In the *Drosophila nasuta* complex, our analyses inferred most species to be monophyletic (Figure 4; S11). These findings are largely consistent with those reported by Mai et al. (2019). (Sub-)species that had not been inferred as monophyletic in our phylogenetic analyses were also not resolved when applying NMDS or STRUCTURE (Figure 4). Otherwise, all (sub)species were clearly distinguishable from each other (Supplementary Text).

Regarding *Heliconius* butterflies, our phylogenies inferred from analyzing mzl-USCOs largely agreed with the phylogeny published by Martin et al. (2013) (Figure S12). We found only few topological differences between analyses that were based on different data extraction approaches and/or phylogenetic reconstruction methods (see Supplementary Text for details). STRUCTURE (Pritchard et al., 2000) and NMDS revealed clusters that were largely consistent with the topology of the phylogenetic trees, with few exceptions described in the Supplementary Text. Analyses based on the datasets from the three USCO extraction approaches gave very similar results (Fig. 4; Figures S14, 15). Even when allowing STRUCTURE to find more clusters than known (sub)species in the analyzed sample by specifying a K value higher than 5, the clustering never supported more than five clusters, and individuals were always assigned to clusters with a probability of more than 90%. A small amount of admixture was detected between sympatric populations (e.g., those of *Heliconius melpomene* and *H. timareta* in Peru).

In Darwin's finches, the alignment completeness of extracted mzl-USCOs was very low (Table 2). The incompleteness of the Darwin's finches' datasets was likely caused by a low sequence coverage ($<$ 10x) and in consequence a poor assembly quality. Therefore, for the analysis of sequence variation we included not only SNPs present in all individuals (as in the other case studies), but also SNPs absent in less than five. Possibly due to the large amount of missing data in the alignments, the inferred phylogenetic trees differed in many details from each other and from the original maximum-likelihood tree based on WGS data (Lamichhaney et al., 2015). Consequently, also NMDS plots of SNP similarity did not provide results that allowed to visually distinguish between different species within the genus*Camarhynchus* and between most of the species within*Geospiza* , except for the species *G. difficilis* and*G. septentrionalis* . However, differentiation between genera was clearly visible. SNP clustering with STRUCTURE also did not allow us to distinguish species of *Camarhynchus* from each other and to distinguish some species of *Geospiza* from each other (Supplementary Text).

Case studies: combination of different USCO extraction methods

Species delimitation methods are highly sensitive to intraspecific nucleotide sequence variation, which affects branch lengths, the topology of single gene trees, and in the end the outcome of the delimitation of populations and species. We used a multiple nucleotide sequence alignment combining data from all three USCO extraction approaches to assess the comparability and combinability of data from different approaches (Figures S16–23). In the ideal case, nucleotide sequences obtained by the three different USCO extraction methods that belong to a given specimen would form a monophyletic group, with no or at least only little nucleotide sequence divergence. In practice, we observed that the presence of nucleotide sequence alignment positions with missing data had an enormous impact on the tree topology, species monophyly, and on the clustering of sequences belonging to the same specimen (Table 4). A particular impact in this regard was caused by discrepancies, both in data yield and in the actual extracted nucleotide sequences, between the BUSCO-based data extraction and the Orthograph-based data extraction.

Visual inspection of the alignments in all four case studies revealed discrepancies between the results of the three USCO extraction methods in 29 to 79 (5–14% of all) of the multiple nucleotide sequence alignments of individual USCO genes. The discrepancies manifested in a clustering of nucleotide sequences that reflected the extraction method rather than individual specimens (Figures S16–23). This pattern was observed in all four case studies, sometimes affecting only a few, sometimes all taxa, and it was more prevalent when

phylogenetically analyzing the extracted data as supermatrix rather than using a summary multispecies coalescent approach that depends on gene trees as input. In a minority of gene loci, the discrepancies could be explained by incorrect alignment of nucleotides across gaps and positions with missing data or at one of the ends of the nucleotide sequence. In the majority of instances, the extraction methods had extracted partially different sequences from the WGS libraries. Such differences were almost always found at the ends of the nucleotide sequences obtained with BUSCO and Orthograph, indicating that different coding nucleotide sequence fragments were evaluated as being part of the gene and were joined together. Editing the datasets by excluding positions with gaps and/or missing data reduced the erroneous inference of non-monophyly of individual samples which is the ultimate test scenario for an error-free species delimitation procedure (Table 4; Figures S16–23).

Case studies: species delimitation

In *Anopheles* and in particular in *Drosophila* , the number of species-level entities recognized was much higher than the currently recognized number of morphospecies. All or most morphospecies were split into multiple species (Fig. 5; Figure S24). SODA exhibited a higher tendency to split individual morphospecies into multiple species than tr2. In Darwin's finches, over-splitting was also visible, especially in the earlier-diverging species. However, we also found many morphospecies in the genera *Geospiza* and *Camarhynchus* to be lumped into a single species. This problem was prevalent when using tr2, but it also occurred with SODA. In *Heliconius* , over-splitting was less of a problem, but it occurred in some cases. Species-level entities mostly corresponded to established (sub)species in which, however, some entities were lumped (*Heliconius melpomene aglaope / H. m. amaryllis* ; *Heliconius melpomene melpomene* [from Panama] */H. m. rosina*). Since over-splitting (or lumping in the case of Darwin's finches) was the overwhelming problem in almost all analyses, we refrained from performing additional analyses with the full WGS data (such analyses were not performed in the studies that generated the WGS data either).

Discussion

Distribution and potential linkage patterns of mzl-USCOs

This study is the first comparative analysis of the physical distribution of mzl-USCOs in the genomes of a wide range of animal taxa. We did not find mzl-USCOs to exhibit a noteworthy tendency of physical linkage when compared to randomly chosen protein-coding genes. Physical distances between USCO genes were found to be in general much larger than the average distances across which loci can be assumed to be linked in evolutionary timescales (<1000 bp; Springer & Gatesy, 2016). The resulting average extent of linkage of loci located on the same chromosome is thus likely negligible and cannot be *a priori*assumed to violate assumptions of multispecies coalescent analyses, irrespective of whether the method is used for phylogenetic reconstruction or species delimitation. Although there was considerable variation across taxa, we found neighboring pairs of mzl-USCOs to be on average spatially located somewhat more closely together than pairs obtained by randomly choosing the same number of annotated protein-coding genes. A possible explanation for this result could be that mzl-USCOs have a small tendency to cluster in genomic regions that are under selection to remain in single-copy.

Mzl-USCOs were found to be rather evenly distributed over the chromosomes and do not cluster on particular chromosomes, indicated by high values of adjusted evenness of the USCO distribution. However, taxa with chromosomes of unequal length tended to have an unequal distribution of mzl-USCOs. This was demonstrated by the positive and significant correlation of the evenness of chromosome length and protein-coding gene distribution with that of the USCO distribution. As expected, longer chromosomes, and especially chromosomes with relatively more protein-coding genes than others, also contain more mzl-USCOs. However, chi-square tests showed that this correlation is not necessarily linear. In nematodes, for example, the correlation of the number of mzl-USCOs with that of protein-coding genes was negative, although this was based on few chromosomes of rather similar length. In particular the deviation of USCO number from chromosome length tended to be higher in birds which also have highly unequal chromosome sizes within their genomes. This deviation is probably due to the fact that gene density is high in short chromosomes (mi-

crochromosomes; e.g., International Chicken Genome Sequencing Consortium, 2004), which are particularly common in birds but are also found in some other vertebrates (Waters et al., 2021). Significant deviations from the distribution of protein-coding genes in general are probably caused by taxon-specific groupings of mzl-USCOs on certain chromosomes. However, such deviations do not seem to be conserved across major lineages, a pattern that is consistent with our observation that groupings of mzl-USCOs on the same chromosome are in most cases not phylogenetically conserved according to the current sampling of taxa. However, as some lineages were poorly covered by these analyses, it is difficult to make accurate statements about this for metazoans in general.

Intra-locus recombination is known to bias coalescent-based phylogenomic analyses (Gatesy & Springer, 2014; Edwards et al., 2016; Springer & Gatesy, 2018). Among eukaryotes, the genome-wide recombination rate is known to vary over at least one order of magnitude (Stapley et al., 2017). Intraspecific recombination rates are also known to vary between the sexes and across the genome, with recombination hot spots in which most crossovers occur (Jeffreys et al., 2001; Kauppi et al., 2004; Niehuis et al., 2010). Recombination hot spots have been studied in a variety of species, including fruit flies (Chan et al., 2012), crickets (Blankers et al., 2018), birds (Kawakami et al., 2017), and mammals (Jeffreys et al., 2001; Kauppi et al., 2004; Arnheim et al., 2007; Penalba & Wolf, 2020). In humans, recombination hot spots are regions of 1 to 2 kbp that are spatially separated from each other by larger regions (50–100 kb) with lower recombination activity (Myers et al., 2005; Baudat et al., 2010). Simulation studies have shown that species tree estimation is robust to recombination even if the amount of recombination exceeds that found in extant organisms (Lanier & Knowles, 2012; Zhu et al., 2022). However, these studies used a model of constant recombination rates across the genome (instead of a model of recombination hot spots), which might not reflect the situation in a given genome properly. We therefore expect that data partitioning and its implementation within models of species inference using the multispecies coalescent will remain a hot topic in the future, as will be some other parameters in species delimitation approaches, e.g., effective population size, whose fluctuation is known to impact species delimitation analyses (Ahrens et al., 2016).

The distribution of distances between USCO genes reported by us exhibited lineage-specific patterns (Fig. 3; Figure S6, S7). Some of these lineages showed an extraordinary variation. This lineage-specific variation likely reflects peculiarities in the genomic architecture of different higher taxa, but a closer investigation of these phenomena is beyond the scope of this study.

Case studies: efficiency of USCO extraction

Our results confirmed that mzl-USCOs can easily be retrieved and extracted from published WGS datasets in sufficient number to be useful to infer reliable phylogenies on all systematic levels, from subspecies to phylum. We found the quantity of recovered mzl-USCOs as well as the number of retrieved alignment positions to be similar to those obtained with approaches in which USCOs were obtained by DNA target enrichment (Table 2; see also Dietz et al., 2023). Similar to the results when assembling raw reads from hybrid enrichment libraries (Dietz et al., 2023), results of gaining USCOs from WGS data were heavily influenced by the data extraction method used and the quality of the genomic data (see also Dietz et al. 2023). These factors might pose limits to the possibilities of how genomic USCO sequence data can be generated and combined in future systematic research (see below).

Independently from how Metazoa-level USCO markers were obtained, either through target enrichment (Dietz et al., 2023) or by extraction from WGS raw reads (this study), it is crucial for understanding the robustness and ability to standardize species inference with Metazoa-level USCO markers to clarify if and how differently generated mzl-USCOs can be simultaneously analyzed. This knowledge is important to evaluate the sustainability of the marker system, particularly for the case that a marker-specific database is to be created. In this context, orthology assessment is crucial, as it is used to identify and define the fundamental entities that subsequent tools rely on. Due to the increasing quality and number of available published genome assemblies, the data available from successive versions of OrthoDB is continuously evolving towards a more comprehensive taxon coverage. Consequently, the set of genes defined as mzl-USCOs (i.e., those present in single-copy in at least 90% of known genomes of a given group; in our case, Metazoa) is

expected to change, but also to converge over time. This change over time is reflected by the number of single-copy genes present in at least 90% of the genomes of Metazoa in different versions of OrthoDB: it has changed from 978 in OrthoDB v. 9, to 954 in OrthoDB v. 10, to 1,268 in OrthoDB v. 11 (Kuznetsov et al., 2023). This change is likely driven by the steady addition of genomes considered for orthology prediction in OrthoDB (330 in v. 9, 448 in v. 10; 812 in v. 11). While the turnover in which genes are considered mzl-USCOs is currently still high, we expect changes regarding the set of genes flagged as mzl-USCOs becoming smaller with increased knowledge of the evolution of gene families.

Although genes identified as mzl-USCOs were not entirely identical between the different versions of OrthoDB used here, we observed that these changes had minimal impact on the overall size of USCO nucleotide alignments, on alignment completeness, on the topology of inferred phylogenetic trees, on nucleotide sequence variation (SNP) clustering, and on species delimitation results, as long as a sufficiently large number of orthologs are involved and if incomplete alignment sections are removed.

Comparing the results of BUSCO-based to the Orthograph-based extraction of USCOs, we find that the former yields generally longer nucleotide sequences per gene. This is probably because, besides the conserved region of a gene identified by the HMMs, BUSCO also includes flanking regions with a length of 5-20 kpb (Simao et al. 2015). Within these regions, the full gene is then determined by using the gene prediction of the Augustus pipeline (Stanke & Waack, 2003; Stanke et al., 2006) to extend the gene model beyond the conserved region specified by the HMM. In contrast, Orthograph searches only for the conserved region for which the HMM was created. This effect is particularly important since the mzl-USCO HMMs contain only the gene region conserved across all Metazoa. Full length coding regions could also be extracted with Orthograph by using HMMs created specific tor the individual groups, although these would then be not necessarily homologous across all groups anymore.

The additional nucleotide sequence information of data based on BUSCO software seems to impact the phylogenetic signal neither positively nor negatively. Consequently, the results of phylogenetic tree reconstruction, SNP clustering, and species delimitation analysis hardly differ between BUSCO- and Orthograph-based extraction approaches. However, our results show that, even with extensive filtering and manual curation, BUSCO- and Orthograph-derived USCO data should not be analyzed together. The resulting phylogenies obtained with mixed data may be severely misleading. Data pruning improves the results, but large discrepancies may remain, especially when using approaches that analyze concatenated alignments rather than coalescent-based approaches that rely on gene trees for phylogenetic inference. Therefore, we recommend extracting mzl-USCOs in one consistent way. The underlying problem is that alignments of systematically different sequences pose a problem to currently used alignment programs. The ongoing change in sets of genes being classified as USCOs could become a problem in terms of data overlap for the sustainability of the marker system, as datasets generated with older and newer OrthoDB versions might not be fully comparable. Here the overlap between Metazoa-level orthologs from OrthoDB v. 9 and OrthoDB v. 10 was relatively moderate (only 580 out of 978 resp. 954 genes; 59% resp. 61%). This has an impact especially on the possibility to combine future USCO data with already available mzl-USCOs generated with DNA target enrichment using baits which are based on earlier and/or different OrthoDB versions. However, we expect that future versions of OrthoDB will have an increasing amount of overlap of mzl-USCOs between versions, since genes that are present and single copy in 90% of the genomes should converge as soon as all taxonomic groups are covered evenly in OrthoDB.

Case studies: phylogenetic trees and species clustering using extracted mzl-USCOs

Mzl-USCOs extracted from WGS were confirmed to separate closely related species in a wide systematic context from each other, as previously shown with USCO data generated with DNA target enrichment (Dietz et al., 2023). Our results demonstrate that the majority of phylogenetic topologies obtained with mzl-USCOs is consistent with the relationships and species entities inferred previously with WGS datasets exemplified in the four case studies of vertebrate and arthropod taxa. The few exceptions of deviating topologies include cases of closely related species that are still frequently hybridizing and for which phylogenies based on single or few genes may give unreliable results. For example, in Darwin's finches and *Drosophila* , the monophyly

of some species was not confirmed. For some of these species this was also the case in the original analyses with WGS data (Lamichhaney et al., 2015; Mai et al., 2019). These results are likely caused by extensive hybridization between species (e.g. in *Anopheles* and *Heliconius*) but possibly also due to large amounts of missing data (particularly in the Darwin's finches). Introgression was reported to occur in*Heliconius* butterflies by Martin et al. (2013) and Edelman et al. (2019) and is also known in other groups studied by us including*Drosophila* (Suvorov et al. 2022). Introgression has been identified in all major organism groups, such as fungi, vertebrates, insects, and angiosperms (Suvorov et al. 2022), indicating that hybridization across species barriers is not uncommon.

In several cases, we found discrepancies between the concatenation data-based and the coalescent analysis-based trees (Table 3). In most of these cases, the ASTRAL trees agreed better with previously published WGS phylogenies (e.g., monophyly of *Heliconius melpomene*and interspecific phylogeny of Darwin's finches) than the concatenation data-based trees. This confirms that coalescent-based approaches are more reliable for inferring the phylogeny of closely related species still under introgression than concatenation data-based phylogenies which are based on the often-incorrect assumption that all loci share the same phylogenetic history (e.g., Solis-Lemus et al., 2016; Bryant & Hahn, 2020; Stolle et al., 2022). However, concatenation data-based approaches seem to give better results if data completeness is highly heterogeneous across individuals. Samples for which information is missing to a high degree are often placed closer to the root in the ASTRAL trees, as seen for example in *Drosophila* . The underlying cause of this may be mapping reference bias and low coverage of some samples (Stolle et al. 2022).

Analyses of nucleotide sequence variation based on SNPs extracted from mzl-USCOs confirmed the results of the phylogenetic analyses regarding the circumscription of species entities. NMDS and STRUCTURE plots allowed us to visually distinguish generally recognized species in most case studies, as was the case in studies that analyzed more extensive WGS data. However, in Darwin's finches, several closely related species were indistinguishable from each other. This result is probably a consequence of a high degree of admixture between the species. It could have alternatively or additionally been caused by the fact that the analyzed dataset suffered from a high degree of missing data. Finally, it is possible that the separation of some species requires the analysis to include more than two dimensions due to the complex distribution of variation. NMDS may also be unreliable if more data are missing in some specimens than in others, as was the case with some *Drosophila*individuals which were placed far apart from others of the same species. Clustering of SNPs with STRUCTURE did not exhibit this problem due to the simpler nature of this analysis as a group reassignment test.

Case studies: Species delimitation

Similar to observations reported in previous studies involving large multi-gene datasets, we found species delimitation algorithms to exhibit a tendency for over-splitting (Sukumaran & Knowles, 2017; Chambers & Hillis, 2020; Dietz et al., 2023). This tendency is probably caused by intraspecific population structure being mistaken for divergence between species — an effect that is expected to positively correlate with dataset size (Leache et al., 2019). Over-splitting happened more frequently when using the species delimitation software SODA than when using the software tr2 — a trend previously already observed by Joshi et al. (2023). This bias in the tendency to over-split species may be caused in part by the larger number of trees used by SODA in comparison to tr2: SODA considers all user-provided gene trees, while tr2 can use only those gene trees that consistently contain all samples. The significantly lower amount of over-splitting in *Heliconius* in comparison to the results obtained from studying the other three taxonomic groups is probably caused by the lower number of individuals in the analysis, which may result in fewer intraspecific clusters that the algorithms could mistake for species. The subspecies of*Heliconius* lumped here (*Heliconius melpomene aglaope* /*H. m. amaryllis* ; *Heliconius melpomene melpomene* [from Panama] / *H. m. rosina*) had the lowest Fst values of all involved taxon pairs (Martin et al. 2013), while they represent highly distinctive morphological forms. These taxa were found to be differentiated almost exclusively at loci related to wing coloration, while the rest of the genome showed very little differentiation (Martin et al. 2013). The lumping in some parts of the Darwin's finch datasets seems to be caused by a large amount of discordance among

gene trees, which may be explained in part by the relative incompleteness of the data, but also by frequent interspecific hybridization (Lamichhaney et al., 2015; Grant & Grant, 2016).

Conclusions

We demonstrated the usefulness of mzl-USCOs as markers for reliably inferring phylogenies on all systematic levels (species group to phylum) irrespective of the specific taxonomic group under consideration. Our analysis of four different recent radiations using WGS datasets showed that USCO data allow distinguishing between species in almost all tested cases and in most cases allow drawing the same conclusions as corresponding studies that analyzed a more comprehensive amount of genomic data. Mzl-USCOs have been proven to be a useful marker system for DNA taxonomy in diverse animal groups, integrating the overarching issues of marker standardization, data production, data repository, and reuse (Miralles et al., 2020). Mzl-USCOs, like any other marker system, may lack the resolving power of whole genome data, especially when speciation is triggered by a single locus or by a few loci (speciation genes; e.g., Orr et al., 2004; Nosil & Schluter, 2011). However, they have the advantage of being universally applicable and comparable across animals. While their distribution in the genome is not fully random, they are widely spread across the genome rather than being clustered in certain genomic regions or on some specific chromosomes. Mzl-USCOs hence constitute a representative sample of the genome for purposes of phylogeny and taxonomy.

Author contributions

D.A., L.D., O.N., B.M., and C.M. designed the study, D.A., O.N., C.M., L.P., and B.M. acquired funding; D.A., L.D., E.S., and C.M. conceptualized and supervised data collection; L.D. did the USCO data extraction, assembly, and analysis; L.D., C.M., and D.A. wrote the original draft of the manuscript; L.D., J.E., C.M., L.P., B.M., E.S., O.N., and D.A. reviewed and edited the manuscript.

AcknowledgementsThis work benefited from the collaborative expertise shared within the DFG priority program SPP 1991 Taxon-Omics. The study was funded by the following grants: German Research Foundation (DFG) grant AH175/6-2 (D.A.), MA 3684/5-2 (C.M.), MI 649/18-2 (B.M.), NI 1387/6-1, 6-2 (O.N.), PO 765/12-2, which are all part of the DFG priority program 1991 (TaxonOmics), German Research Foundation grants STO 1240/3-1 (433110898), STO 1240/4-1 (445756277) and STO 1240/9-1 (503360601, part of DFG priority program GEVOL, SPP 2349) (E. S.).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

Data availability statement

No new sequence data were generated during this study. Data for the four case studies are available at NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject) under the following accession numbers: Darwin's finches: PRJNA263122 (Lamichhaney et al. 2015), *Anopheles* : PRJNA67511 (Fontaine et al. 2015), *Drosophila* : PRJNA554139 (Mai et al. 2019), *Heliconius* : PRJEB1749 (Martin et al. 2013). Metazoan genomes are available at NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/; O'Leary et al. 2016); for individual accession numbers see Table S1. Multiple sequence alignments, phylogenetic trees, and other results of this study are available at DataDryad (DOI: 10.5061/dryad.kprr4xhb3).

Orcid

Lars Dietz https://orcid.org/0000-0001-6469-381X

Christoph Mayer https://orcid.org/0000-0001-5104-6621

Eckart Stolle https://orcid.org/0000-0001-7638-4061

Jonas Eberle https://orcid.org/0000-0003-2519-0640

Bernhard Misof https://orcid.org/0000-0003-4175-6798

Lars Podsiadlowski https://orcid.org/0000-0001-7786-8930

Oliver Niehuis https://orcid.org/0000-0003-4253-1849

Dirk Ahrens https://orcid.org/0000-0003-3524-7153

References

Ahrens, D. (2023) Species diagnosis and DNA taxonomy. In: Desalle, R. (ed) DNA Barcoding: Methods and protocols. Preprint DOI:10.5281/zenodo.8079017

Ahrens, D., Fujisawa, T., Krammer, H.-J., Eberle, J., Fabrizi, S., & Vogler, A.P. (2016). Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology* , 65 (3), 478-494.

Ahrens, D., Ahyong, S. T., Ballerio, A., Barclay, M. V. L., Eberle, J., Espeland, M., Huber, B.A., Mengual, X., Pacheco, T.L., Peters, R. S., Rulik, B., Vaz-de-Mello, F., Wesener, T., & Krell, F.-T. (2021). Is it time to describe new species without diagnoses? – A comment on Sharkey et al. (2021). *Zootaxa* , 5027 (2), 151–159.

Arnheim, N., Calabrese, P., & Tiemann-Boege, I. (2007). Mammalian meiotic recombination hot spots. *Annual Review of Genetics* , 41, 369–399.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* , 3, e337.

Ballard, J. W., & Whitlock M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecolology* , 13, 729–744.

Baudat, F., Buard, J., Grey C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop G., & de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* , 327(5967), 836–840.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* , 304, 1321–1325.

Blankers, T., Oh, K.P., Bombarely, A., & Shaw, K.L. (2018). The genomic architecture of a rapid island radiation: Recombination rate variation, chromosome structure, and genome assembly of the Hawaiian cricket*Laupala* . *Genetics* , 209(4), 1329–1344.

Bryant, D., & Hahn, M. W. (2020). The concatenation question. In: Scornavacca C., Delsuc F., Galtier N. Phylogenetics in the genomic era. No commercial publisher | Authors open access book, pp.3.4:1–3.4:23. Hal-02535651

Chambers, E.A., & Hillis, D. M. (2020). The multispecies coalescent over-splits species in the case of geographically widespread taxa.*Systematic Biology* , 69, 184-193.

Chan, A. H., Jenkins, P.A., & Song Y. S. (2012). Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster* .*PloS Genetics* , 8(12), e1003090.

Chen, C.-S., Huang, C.-T., & Hseu, R.-S. (2017). Evidence for two types of nrDNA existing in Chinese medicinal fungus *Ophiocordyceps sinensis* . *AIMS Genetics* , 4, 192–201.

Dietz, L., Eberle, J., Mayer, C., Kukowka, S., Bohacz, C., Baur, H., Espeland, M., Huber, B. A., Hutter, C., Mengual, X., Peters, R. S., Vences, M., Wesener, T., Willmott, K., Misof, B., Niehuis, O., & Ahrens, D. (2023) Standardized nuclear markers improve and homogenize species delimitation in Metazoa. *Methods in Ecology and Evolution* , 14, 543-555.

Eberle, J., Ahrens, D., Mayer, C., Niehuis, O., & Misof, B. (2020). A plea for standardized nuclear markers in metazoan DNA taxonomy. *Trends in Ecology and Evolution* , 35, 336–345.

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., Garcia-Accinelli, G., Belleghem, S. M. V., Patterson, N., Neafsey, D. E., Challis, R., Kumar, S., Moreira, G. R. P., Salazar, C., Chouteau, M., Counterman, B. A., Papa, R., Blaxter, M., Reed, R. D., Dasmahapatra, K. K., Kronforst, M., Joron, M., Jiggins, C.D., McMillan, W. O., Palma, F. D., Blumberg, A. J., Wakeley, J., Jaffe, D., Mallet, J., 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* , 366, 594–599.*https://doi.org/10.1126/science.aaw2090*

Edwards, D. L., & Knowles, L. L. (2014). Species detection and individual assignment in species delimitation: can integrative data increase efficacy? *Proceedings of the Royal Society B: Biological Sciences* , 281, 20132765.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leache, A. D., Liu, L., & Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics.*Molecular Phylogenetics and Evolution* , 94, 447–462.

Erickson, K. L., Pentico, A., Quattrini, A. M., & McFadden, C. S. (2021). New approaches to species delimitation and population structure of anthozoans: Two case studies of octocorals using ultraconserved elements and exons. *Molecular Ecology Resources* , 21, 78–92.*https://doi.*org/10.1111/1755-0998.13241

Esselstyn, J. A., Oliveros, C. H., Swanson, M. T., & Faircloth, B. C. (2017). Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements.*Genome Biology and Evolution* , 9, 2308-2321.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield ,R. T., Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales.*Systematic Biology* , 61, 717–726.

Fernandez, R., Kallal, R.J., Dimitrov, D., Ballesteros, J.A., Arnedo, M.A., Giribet, G., & Hormiga, G. (2018). Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Current Biology* , 28, 1489-1497.e5.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y. C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W., & Besansky, N. J. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* , 347, 1258524.

Fontaneto, D., Flot, J. F., & Tang, C. Q. (2015). Guidelines for DNA taxonomy, with a focus on the meiofauna. *Marine Biodiversity* , 45, 433–451.*https://doi.*org/10.1007/s12526-015-0319-7

Fujisawa T., Aswad A., & Barraclough T.G. (2016). A rapid and scalable method for multilocus species delimitation using Bayesian model comparison and rooted triplets. *Systematic Biology* , 65, 759-71.

Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics* , 34, 397–423.

Gatesy, J., & Springer, M. S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution* , 80, 231–266.

Grant, P. R., & Grant, B. R. (2016). Introgressive hybridization and natural selection in Darwin's finches. *Biological Journal of the Linnean Society* , 117, 812-822.

Gueuning, M., Frey, J. E., & Praz, C. (2020). Ultraconserved yet informative for species delimitation: Ultraconserved elements resolve long-standing systematic enigma in Central European bees.*Molecular Ecology* , 29, 4203–4220.

Hammer, O., Harper, D. A. T., & Ryan, P. D. (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica* , 4, 1–9.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* , 270, 313–321.

Heip, C. H. R., Herman, P. M. J., & Soetaert, K. (1998). Indices of diversity and evenness. *Oceanis* , 24(4), 61–87.

Herrera, S., & Shank, T. M. (2016). RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Molecular Phylogenetics and Evolution* , 100, 70–79.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation.*Molecular Biology and Evolution* , 35, 518–522.

International Chicken Genome Sequencing Consortium. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* , 432, 695–716.

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.-Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., & Philippe, H. (2017). Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature Ecology & Evolution* , 1, 1370–1378.

Jeffreys, A. J., Kauppi, L., & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* , 29(2), 217–222.

Ješovnik, A., Sosa-Calvo, J., Lloyd, M. W., Branstetter, M. G., Fernández, F., & Schultz, T. R. (2017). Phylogenomic species delimitation and host-symbiont coevolution in the fungus-farming ant genus *Sericomyrmex* Mayr (Hymenoptera: Formicidae): ultraconserved elements (UCEs) resolve a recent a recent radiation. *Systematic Entomology* , 42, 523–542.

Joshi, C. J., Ke, W., Drangowska-Way, A., O'Rourke, E. J., & Lewis, N. E. (2022). What are housekeeping genes? *PloS Computational Biology* , 18(7), e1010295. Doi: 10.1371/journal.pcbi.1010295.

Joshi, M., Espeland, M., Huemer, P., deWaard, J., Mutanen, M. (2023). Species delimitation under allopatry: genomic divergences within and across continents in Lepidoptera. bioRxiv 2023.03.06.531242; doi: https://doi.org/10.1101/2023.03.06.531242

Junier, T., & Zdobnov, E.M. (2010). The Newick Utilities: High-throughput phylogenetic tree processing in the UNIX shell.*Bioinformatics* , 26, 1669-1670.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., & Jermiin, L.S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* , 14, 587–589.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability.*Molecular Biology and Evolution* , 30, 772–780.

Kauppi, L., Jeffreys, A. J., & Keeney, S. (2004). Where the crossovers are: recombination distributions in mammals. *Nature Reviews Genetics* , 5(6), 413–424.

Kawakami, T., Mugal, C. F., Suh, A., Nater, A., Burri, R., Smeds, L., & Ellegren, H. (2017). Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Molecular Ecology* , 26(16), 4158–4172.

Krehenwinkel, H., Pomerantz, A., Henderson, J. B., Kennedy, S. R., Lim, J. Y., Swamy, V., Shoobridge, J. D., Graham, N., Patel, N. H., Gillespie, R. G., & Prost, S. (2019). Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience* , 8, 1–16.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* , 47(D1), D807–D811.

Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B. M., Wägele, J. W., & Misof, B. (2010). Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology* , 7, 10.

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., & Zdobnov, E. M. (2023). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* , 51(D1), D445-D451.

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C. J., Wang, C., Zamani, N., Grant, B. R., Grant, P. R, Webster, M. T., & Andersson, L. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* , 518, 371-375.

Lamichhaney, S., Han, F., Berglund, J., Wang, C., Almén, M. S., Webster, M. T., Grant, B. R., Grant, P. R., & Andersson, L. (2016). A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science* , 352, 470–474.

Lamichhaney, S., Han, F., Webster, M. T., Andersson, L., Grant, B. R., & Grant, P. R. (2018). Rapid hybrid speciation in Darwin's finches.*Science* , 359, 224–228.

Lanier, H. C., & Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Systematic Biology* , 61(4), 691–701.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics* , 30(22), 3276–3278.

Laumer, C. E., Fernández, R., Lemer, S., Combosch, D., Kocot, K. M., Riesgo, A., Andrade, S. C. S., Sterrer, W., Sørensen, M. V., & Giribet, G. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B: Biological Sciences* , 286, 20190831.

Lebonah, D. E., Dileep, A., Chandrasekhar, K., Sreevani, S., Sreedevi, B., & Pramoda Kumari, J. (2014). DNA barcoding on bacteria: A review.*Advances in Biology* , 2014, 541787. https://doi.org/10.1155/2014/541787

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). 1000 genome project data processing subgroup. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* , 25, 2078–2079.

Maddison, W. P., & Maddison, D.R. (2018). Mesquite: a modular system for evolutionary analysis. Version 3.51*http://www.*mesquiteproject.org

Mai, D., Nalley, M.J., & Bachtrog, D. (2019). Patterns of genomic differentiation in the *Drosophila nasuta* species complex.*Molecular Biology and Evolution* , 37, 208–220.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* , 38(10), 4647–4654.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., & Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in*Heliconius* butterflies. *Genome Reserach* , 23, 1817–1828.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era.*Molecular Biology and Evolution* , 37, 1530–1534.

Miralles, A., Bruy, T., Wolcott, K., Scherz, M.D., Begerow, D., Beszteri, B., Bonkowski, M., Felden, J., Gemeinholzer, B., Glaw, F., Glöckner, F. O., Hawlitschek, O., Kostadinov, I., Nattkemper, T. W., Printzen, C., Renz, J., Rybalka, N., Stadler, M., Weibulat, T., Wilke, T., Renner, S. S., & Vences, M. (2020). Repositories for taxonomic data: where we are and what is missing. *Systematic Biology* , 69(6),1231–1253

Misof, B., & Misof, K. (2009). A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology* , 2009: 58.

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* , 310, 321–324.

Niehuis, O., Gibson, J. D., Rosenberg, M. S., Pannebakker, B. A., Koevoets, T., Judson, A. K., Desjardins, C. A., Kennedy, K., Duggan, D., Beukeboom, L. W., van de Zande, L., Shuker, D. M., Werren, J. H., & Gadau, J. (2010). Recombination and its impact on the genome of the haplodiploid parasitoid wasp *Nasonia* . *PloS One* , 5, e8597.

Nosil, P., & Schluter, D. (2011). The genes underlying the process of speciation. *Trends in Ecology & Evolution* , 26(4), 160–7.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* , 44(D1), D733-745.

Orr, H. A., Masly, J. P., & Presgraves, D. C. (2004). Speciation genes.*Current Opinion in Genetics & Development* , 14(6), 675-9.

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., & Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* , 2(4), e000056.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S. C., Boisselier, M. C., & Samadi, S. (2015). Use of RAD sequencing for delimiting species. *Heredity* , 114, 450–459.

Penalba, J. V., & Wolf, J. B. W. (2020). From molecules to populations: appreciating and estimating recombination rate variation. *Nature Reviews Genetics* , 21(8), 476–492.

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R. S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., & Niehuis, O. (2017). Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics* , 18, 111.

Pierce, M. P. (2019). Filling in the gaps: adopting ultraconserved elements alongside COI to strengthen metabarcoding studies.*Frontiers in Ecology and Evolution* , 7(469), 1–6.

Prebus, M. M. (2021). Phylogenomic species delimitation in the ants of the *Temnothorax salvini* group (Hymenoptera: Formicidae): an integrative approach. *Systematic Entomology* , 46, 307–326.

Pritchard, J. K., Stephens, M., & Donnelly, P. J. (2000). Inference of population structure using multilocus genotype data. *Genetics* , 155, 945–959.

Rabiee M., & Mirarab S. (2020). SODA: Multi-locus species delimitation using quartet frequencies. *Bioinformatics* , 36, 5623–5631.

Sahbou, A.-E., Iraqi, D., Mentag, R., & Khayi, S. (2022). BuscoPhylo: A webserver for Busco-based phylogenomic analysis for non-specialists.*Scientific Reports* , 12, 17352

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* , 31, 3210–2.

Springer, M. S., & Gatesy, J. (2016). The gene tree delusion.*Molecular Phylogenetics and Evolution* , 94A, 1–33.

Springer, M. S., & Gatesy, J. (2018). Delimiting coalescence genes (C-Genes) in phylogenomic data sets. *Genes* , 9(3), 123.

Solís-Lemus, C., Yang, M., Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology* , 65(5), 843–851, https://doi.org/10.1093/sysbio/syw030

Stanke, M., & Waack S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* , 19 (suppl 2), ii215–ii225, https://doi.org/10.1093/bioinformatics/btg1080

Stanke, M., Tzvetkova, A., & Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology* , 7 (Suppl 1), S11 (2006). https://doi.org/10.1186/gb-2006-7-s1-s11

Stolle, E., Pracana, R., López-Osorio, F. et al. (2022). Recurring adaptive introgression of a supergene variant that determines social organization. *Nature Communications* , 13, 1180*https://doi.*org/10.1038/s41467-022-28806-7

Sukumaran, J., & Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences* , 114, 1607–1612.

Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'Agostino, E. R. R., Price, D. K., Waddell, P. J., Lang, M., Courtier-Orgogozo, V., David, J. R., Petrov, D., Matute, D. R., Schrider, D. R., & Comeault, A. A. (2022). Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Current Biology* , 32(1), 111-123.e5,*https://doi.org/10.1016/j.cub.2021.10.052.*

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.*Nucleic Acids Research* , 34, W609-W612.

Thawornwattana, Y., Dalquen, D., & Yang, Z. (2018). Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Molecular Biology and Evolution* , 35, 2512–2527.

Vicente, J. L., Clarkson, C. S., Caputo, B., Gomes, B., Pombi, M., Sousa, C. A., Antao, T., Dinis, J., Bottà, G., Mancini, E., Petrarca, V., Mead, D., Drury, E., Stalker, J., Miles, A., Kwiatkowski, D. P., Donnelly, M. J., Rodrigues, A., della Torre, A., Weetman, D., & Pinto, J. (2017). Massive introgression drives species radiation at the range limit of *Anopheles gambiae* . *Scientific Reports* , 7, 46451.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* , 35, 543–548.

Waters, P. D., Patel, H. R., Ruiz-Herrera, A., & Marshall Graves, J. A. (2021). Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proceedings of the National Academy of Sciences* , 118(45), e2112494118.

Wu, C.-I. (2001). The genic view of the process of speciation.*Journal of Evolutionary Biology* , 14: 851–865.

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

Zarza, E., Connors, E. M., Maley, J. M., Tsai, W. L. E., Heimes, P., Kaplan, M., & McCormack, J. E. (2018). Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (*Sarcohyla* ; Hylidae). *PeerJ* , 6, e6045.

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., Seppey, M., Loetscher, A., Kriventseva, E. V. (2017). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.*Nucleic Acids Research* , 45(D1), D744–D749.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* , 19, 153.

Zhang, F., Ding, Y., Zhu, C.D., Zhou, X., Orr, M.C., Scheu, S., & Luan, Y.X. (2019). Phylogenomics from low-coverage whole-genome sequencing.*Methods in Ecology and Evolution* , 10, 507–517.

Zhu, T., Flouri, T., & Yang, Z. (2022). A simulation study to examine the impact of recombination on phylogenomic inferences under the multispecies coalescent model. *Molecular Ecology* , 31, 2814–2829.

Zink, R. M., & Vázquez-Miranda, H. (2019). Species limits and phylogenomic relationships of Darwin's finches remain unresolved: potential consequences of a volatile ecological setting.*Systematic Biology* , 68, 347–357.

Supporting Information

The online version contains supplementary material available at Dryad (DOI: 10.5061/dryad.kprr4xhb3).

Table 1. Information on taxa that served as a reference in our four taxonomic study groups and for which we extracted mzl-USCOs from published whole genomes: number of species (N_{sp}) excluding outgroup taxa (number of subspecies in parentheses), number of individuals (N_{ind}), reference genome used for mapping raw sequence reads, and data source (individual samples listed in Table S2).

Group	**N_{sp}**	**N_{ind}**	**Reference genome**	**Reference**
Anopheles gambiae **complex**	6	74	*Anopheles arabiensis* (AaraD3)	Fontaine et al. (2015
Drosophila nasuta **complex**	9 (12)	68	*Drosophila albomicans* (drosAlbom15112-1751.03v1)	Mai et al. (2019)
Heliconius	3 (6)	30	*Heliconius melpomene* (Hmel1)	Martin et al. (2013)
Darwin's finches	18	120	*Camarhynchus parvulus* (STF_HiC)	Lamichhaney et al. (

Table 2. Results of extracting mzl-USCOs from published genomes in the four taxonomic groups when using different extraction approaches. All mzl-USCOs found in the reference were also found in the genomic reads of at least some specimens.

	Number of mzl-USCOs found	**Number of mzl-USCOs found in all specime**
BUSCO / OrthoDB v. 10		
Anopheles	936	911
Drosophila	906	798
Heliconius	900	884
Darwin's finches	914	423
Orthograph / OrthoDB v. 9		
Anopheles	961	800
Drosophila	952	589
Heliconius	953	864
Darwin's finches	964	208
Orthograph / OrthoDB v. 10		
Anopheles	900	761
Drosophila	908	578

	Number of mzl-USCOs found	Number of mzl-USCOs found in all specime
Heliconius	896	828
Darwin's finches	932	217

Table 3. Number of monophyletic species (*or subspecies) in the trees of the original studies (WGS; see Table 1) and in the corresponding USCO alignment-based trees in our study. We analyzed a supermatrix of concatenated (C) USCO alignments and determined species trees with the multispecies coalescent approach implemented in ASTRAL (A).

Case study	**Species***	**WGS**	**BUSCO**	**BUSCO**	**Orthograph/OrthoDB v. 9**	**Orthograph/OrthoDB v. 9**	**Orthograph/OrthoDB v. 10**	**Orthog Orthol v. 10**
			C	**A**	**C**	**A**	**C**	**A**
Anopheles	6	5	4	5	6	5	4	5
Drosophila	12	10	8	8	8	8	8	8
Heliconius	6	4	3	3	3	3	3	3
Darwin's finches	18	18	17	15	14	13	16	10

Table 4 . Compatibility of USCO extraction approaches: number of individuals within each case study in which nucleotide sequences from different data extraction approaches formed a monophyletic group when analyzed together. Numbers before slashes refer to individuals whose extracted sequences from all three approaches grouped together, numbers behind the slash refer to individuals whose extracted sequences at least from the two Orthograph-based approaches formed a monophyletic group when analyzed together (see also Figures S16–23). m.d. = missing data.

	Anopheles	Drosophila	Heliconius	**Darwin's finches**
concatenated				
full dataset	38 / 74	8 / 65	0 / 5	0 / 120
m.d./gaps excluded	74 / 74	68 / 68	4 / 26	6 / 117
corrected	74 / 74	68 / 68	30 / 30	120 /120
corrected + m.d./gaps excluded	74 / 74	67 / 67	30 / 30	27 / 113
coalescent				
full dataset	38 / 62	53 / 60	17 / 30	111 / 120
m.d./gaps excluded	74 / 74	66 / 68	30 / 30	119 / 120
corrected	60 / 74	56 / 60	30 / 30	120 / 120
corrected + m.d./gaps excluded	74 / 74	67 / 68	30 / 30	118 / 119
$\mathbf{N_{ind}}$	**74**	**68**	**30**	**120**

Figure captions:

Fig. 1. Histogram showing the number of mzl-USCO gene pairs analyzed in this study which occur on the same chromosome in a given proportion of the examined taxa. The histogram shows that the proportion of genomes in which a gene pair occurs on the same chromosome is typically rather small.

Fig. 2. Distribution of median distances between neighboring mzl-USCO genes, in nucleotides divided by genome size. Left: based on real USCO data across all taxa, right: based on a random selection of protein-coding genes for each taxon. Lines connect dots belonging to the same taxon.

Fig. 3. Phylogenetic signal and systematic correlation of distances between neighboring USCOs with major metazoan lineages. A: PC axes 1 (left tree) and 2 (right tree) from a PCA on frequencies of size classes of Metazoa-level USCO distances, mapped onto the Metazoa phylogeny based on concatenated amino acid sequences. B: Plot of axes 1 and 2 from the same PCA, showing a clustering of major metazoan lineages (Protostomia and Deuterostomia with unfilled and filled color symbols, respectively).

Fig. 4. Data yield and results of analyses on mzl-USCOs extracted from *Drosophila* WGS reads when applying three different USCO extraction methods: A: Number of mzl-USCOs recovered per number of specimens; B: ASTRAL trees based on generated USCO datasets; C: Outcome of SNP clustering analyses with STRUCTURE; D: NMDS plots of SNP similarity.

Fig. 5. Species delimitation of the four case studies based on the programs tr2 and SODA on each data set from the three different extraction methods. Colored boxes indicate that inferred species entities match with currently recognized morphospecies.

Figure S1. Proportion of pairwise sequence overlap in the concatenated alignment of USCO loci between pairs of chromosome-level annotated metazoan genomes.

Figure S2. Maximum likelihood phylogenetic tree based on concatenated amino acid USCO sequences of all analyzed chromosome-level annotated genomes of Metazoa. Numbers above branches are support values from approximate likelihood ratio tests and ultrafast bootstrapping.

Figure S3. Maximum likelihood phylogenetic tree based on concatenated nucleotide USCO sequences (codon positions 1 and 2) of all analyzed chromosome-level annotated genomes of Metazoa. Numbers above branches are support values from approximate likelihood ratio tests and ultrafast bootstrapping.

Figure S4. Multispecies coalescent-based phylogenetic tree based on gene trees of amino acid USCO sequences of all analyzed chromosome-level annotated genomes of Metazoa. Numbers above branches are local posterior probabilities.

Figure S5. Multispecies coalescent-based phylogenetic tree based on gene trees of nucleotide USCO sequences (codon positions 1 and 2) of all analyzed chromosome-level annotated genomes of Metazoa. Numbers above branches are local posterior probabilities.

Figure S6 . Quotient of median distance between neighboring mzl-USCOs to the median distance between neighboring randomly selected annotated protein-coding genes, mapped onto the Metazoa phylogeny based on concatenated amino acid sequences.

Figure S7. Axes 1 and 2 of a PCA on frequencies of size classes of distances between neighboring Metazoa-level USCOs mapped onto the Metazoa phylogeny based on concatenated amino acid sequences (detailed version with taxon names of analyzed chromosome-level genomes).

Figure S8. Number of mzl-USCOs recovered per number of specimens when applying different USCO extraction methods.

Figure S9. Proportion of pairwise sequence overlap in the concatenated alignment of USCO loci between pairs of specimens within each case study (*Anopheles* , *Drosophila* , *Heliconius* , Darwin's finches) analyzed in the present investigation, sorted by extraction method (BUSCO, Orthograph + OrthoDB v. 9, Orthograph + OrthoDB v. 10).

Figure S10. Phylogenetic trees of *Anopheles* species inferred with concatenated USCO nucleotide sequences (above) and with the multispecies coalescent (below) generated with different USCO extraction methods.

Figure S11. Phylogenetic trees of *Drosophila* species inferred with concatenated USCO nucleotide sequences (above) and with the multispecies coalescent (below) generated with different USCO extraction methods.

Figure S12. Phylogenetic trees of *Heliconius* species inferred with concatenated USCO nucleotide sequences (above) and with the multispecies coalescent (below) generated with different USCO extraction methods.

Figure S13. Phylogenetic trees of Darwin's finches inferred with concatenated USCO nucleotide sequences (above) and with the multispecies coalescent (below) generated with different USCO extraction methods.

Figure S14. NMDS plots showing similarities between specimens inferred with SNP data of mzl-USCOs for the four study groups based on datasets generated with different data extraction methods.

Figure S15. Diagrams of STRUCTURE clustering results inferred with SNP data of mzl-USCOs for the four study groups based on datasets generated with different data extraction methods.

Figure S16. ML trees of concatenated multiple nucleotide sequence alignments of 580 genes classified as mzl-USCOs in both OrthoDB versions v.9 and v.10 and extracted with three methods from *Anopheles* genomic data. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S17. Coalescent-based trees inferred in the *Anopheles* case study with data from the three USCO extraction approaches aligned in a single dataset using only those 580 genes classified as mzl-USCOs in both OrthoDB v.9 and v.10. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S18. ML trees of concatenated multiple nucleotide sequence alignments of 580 genes classified as mzl-USCOs in both OrthoDB v.9 and v.10 and extracted with three methods from *Drosophila* genomic data. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S19. Coalescent-based trees inferred in the *Drosophila* case study with data from the three USCO extraction approaches aligned in a single dataset using only those 580 genes classified as mzl-USCOs in both OrthoDB v.9 and v.10. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S20. ML trees of concatenated multiple nucleotide sequence alignments of 580 genes classified as mzl-USCOs in both OrthoDB v.9 and v.10 and extracted with three methods from *Heliconius* genomic data. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S21. Coalescent-based trees inferred in the *Heliconius* case study with data from the three USCO extraction approaches aligned in a single dataset using only those 580 genes classified as mzl-USCOs in both OrthoDB v.9 and v.10. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S22. ML trees of concatenated multiple nucleotide sequence alignments of 580 genes classified as mzl-USCOs in both OrthoDB v.9 and v.10 and extracted with three methods from genomic data of Darwin's finches. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually

corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S23. Coalescent-based trees inferred in the Darwin's finches case study with data from the three USCO extraction approaches aligned in a single dataset using only those 580 genes classified as mzl-USCOs in both OrthoDB v.9 and v.10. Trees, from left to right, are based on: 1) all data, 2) data after excluding alignment positions with missing data and gaps (gaps excluded), 3) a manually corrected alignment (corrected), and 4) a manually corrected alignment with additional exclusion of alignment positions with missing data and gaps (corrected + gaps excluded).

Figure S24. Results of species delimitation using tr2 and SODA in each case study and applying each of the three data extraction approaches.

Table S1. Metazoan genomes assembled to chromosome level included in this study, with numbers of single-copy mzl-USCO genes found in these genomes with the BUSCO software, number of chromosomes, genome size, median distance between neighboring USCOs, median distance between neighboring randomly chosen annotated protein coding genes, logarithms of those two distances, the distances divided by genome size, the quotient between these distances, p-value based on 10,000 replicates for the mzl-USCO distance being smaller, adjusted evenness values for chromosome length, number of coding genes, number of mzl-USCOs, chi-square values for distribution of mzl-USCOs compared to chromosome length and to number of coding genes and p-values derived from the chi-square tests.

Table S2. NCBI accession numbers of the raw reads from individuals analyzed in the four taxonomic case studies.