



OPEN

Orthoptera-specific target enrichment (OR-TE) probes resolve relationships over broad phylogenetic scales

Seunggwon Shin^{1,2}, Austin J. Baker^{2,3}, Jacob Enk⁴, Duane D. McKenna², Bert Foquet^{5,6}, Amy G. Vandergast⁷, David B. Weissman⁸ & Hojun Song⁶✉

Phylogenomic data are revolutionizing the field of insect phylogenetics. One of the most tenable and cost-effective methods of generating phylogenomic data is target enrichment, which has resulted in novel phylogenetic hypotheses and revealed new insights into insect evolution. Orthoptera is the most diverse insect order within polyneoptera and includes many evolutionarily and ecologically interesting species. Still, the order as a whole has lagged behind other major insect orders in terms of transitioning to phylogenomics. In this study, we developed an Orthoptera-specific target enrichment (OR-TE) probe set from 80 transcriptomes across Orthoptera. The probe set targets 1828 loci from genes exhibiting a wide range of evolutionary rates. The utility of this new probe set was validated by generating phylogenomic data from 36 orthopteran species that had not previously been subjected to phylogenomic studies. The OR-TE probe set captured an average of 1037 loci across the tested taxa, resolving relationships across broad phylogenetic scales. Our detailed documentation of the probe design and bioinformatics process is intended to facilitate the widespread adoption of this tool.

Keywords Phylogenomics, Hybrid capture, Transcriptome, Systematics

With advances in high-throughput sequencing and bioinformatics techniques, we are witnessing a revolution in insect phylogenetics^{1–4}. Novel phylogenetic hypotheses for major insect orders have routinely been proposed based on phylogenomic data^{5–11}, confirming and challenging previous hypotheses based on a small number of genes and/or morphology. One of the most promising methods for phylogenomic data generation is a technique known as target enrichment, which uses hybrid capture probes (or baits) to collect specific genes of interest before sequencing¹². Generally, the targeted orthologs are single-copy and exhibit an appropriate amount of variation for phylogenetic analysis. The targeted genes are enriched after hybridization, greatly increasing the coverage of genes of interest for final sequencing¹³. Due to the enrichment process, this method does not require live specimens or freshly collected samples, and it can even be applied to dried museum samples, thus facilitating a broad sampling of taxa¹⁴. Ultra-conserved elements (UCE)¹⁵ and anchored hybrid enrichment (AHE)¹³ are the two most frequently used approaches for insect phylogenomics using target hybrid enrichment. Though they have different strategies for target selection, both systems capture highly conserved loci in sequencing libraries and use sequence variation within and flanking those targets to infer phylogenetic relationships at various scales.

Genomic resources for identifying target loci are prerequisites for developing target enrichment probes¹⁴. High-quality reference genomes are the gold standard for identifying orthologs that can be used for phylogenomics, but such genomes are not available for many insect orders. While hundreds of insect genomes have been sequenced so far, many are either model organisms or agriculturally or medically important species and are not necessarily representative of the phylogenetic diversity of the group(s) in question. Also, most available insect genomes belong to holometabolous insects, and only a small number of hemimetabolous insect orders

¹School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea. ²Department of Biological Sciences, Center for Biodiversity Research, University of Memphis, Memphis, TN, USA. ³Entomology Department, Natural History Museum of Los Angeles County, Los Angeles, CA, USA. ⁴Daicel Arbor Biosciences, Ann Arbor, MI, USA. ⁵McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. ⁶Department of Entomology, Texas A&M University, College Station, TX, USA. ⁷San Diego Field Station, Western Ecological Research Center, U.S. Geological Survey, San Diego, CA, USA. ⁸Department of Entomology, California Academy of Sciences, Golden Gate Park, San Francisco, CA, USA. ✉email: Hojun.Song@ag.tamu.edu

(except Hemiptera) have been sequenced. Moreover, many polyneopteran insects have larger genome sizes than holometabolous insect orders¹⁶, contributing to the general lack of genome sequencing projects in these groups. Molecular systematists studying those holometabolous orders that contain genetic model organisms, including Coleoptera, Diptera, Hymenoptera, and Lepidoptera, were early adopters of this technique into their toolkits, resulting in ground-breaking phylogenomic studies with extensive taxon sampling^{7,9,17–25}. Now, pre-designed probes are commercially available for these insect orders. For other insect orders, however, the application of phylogenomics has been slower. As such, there has been a discrepancy in applying phylogenomics across different insect orders.

Orthoptera is one of the insect orders that has lagged in transitioning to phylogenomics. It is the most species-rich order among the polyneopteran insect lineages, with more than 29,400 described species worldwide²⁶, and includes some of the most recognisable and familiar insects, such as grasshoppers, locusts, crickets, and katydids, which are economically and culturally significant and evolutionarily fascinating^{10,27,28}. Nevertheless, the number of researchers who study the phylogeny of Orthoptera is small compared to those who study other major insect orders, and most of the molecular phylogenetic studies still rely on Sanger sequencing data or mitochondrial genome sequences. Genomic resources for developing phylogenomic tools have been largely lacking because orthopteran genomes are known to be the largest among insects^{29,30}, and thus, very challenging to sequence and annotate. Although recent efforts to sequence orthopteran genomes have made great strides³¹, initiating a genome sequencing project for any orthopteran species remains challenging. Recently, Song et al.¹⁰ partnered with the 1KITE (1000 Insect Transcriptome Evolution) project to generate nearly 5000 single-copy orthologs from transcriptomes of 50 orthopteran and ten polyneopteran species to resolve the higher-level relationships within the order, which represents one of the first attempts to apply phylogenomics in Orthoptera. However, there has not been an effort to produce target enrichment tools that can be used broadly for Orthoptera.

In this study, we aimed to develop an Orthoptera-specific target enrichment (OR-TE) probe set as a new phylogenomic toolkit. Specifically, we compiled transcriptomes from 80 orthopteran species from across the phylogeny, 30 of which were newly generated, as a new genomic resource to identify phylogenetically informative orthologs. From this initial set of orthologs, we identified both slow-evolving and fast-evolving loci that could resolve relationships at different taxonomic scales to narrow the number of target loci to 1828, which were used to develop target enrichment probes. We designed and manufactured a custom probe set with 39,809 baits and validated the effectiveness of this probe set by generating target enrichment data from 36 orthopteran species across the phylogeny, which were then used to infer the phylogeny of Orthoptera. We carefully documented the probe design and bioinformatics process so that the orthopteran systematics community can widely adopt and use this newly developed tool.

Results

Designing the OR-TE probe set

The development of genomic resources to support identifying orthologs suitable for target enrichment proceeded by sampling whole RNA from 80 orthopteran species, of which 41 were previously collected within the 1KITE project, nine were previously generated by Song et al.¹⁰, and the remaining 30 were newly generated for this study. We also included ten polyneopteran outgroups previously generated for the 1KITE project. The 80 orthopteran species sampled for this study included representatives of 24 orthopteran families belonging to both suborders, Caelifera (58 species) and Ensifera (22 species), covering the majority of higher-level taxonomic diversity within the order (Supplementary Data 1). After assembling transcriptomes, we first explored the phylogenetic information content of the orthologs identified from the transcriptomes to identify target genes to be included in the OR-TE probe set. We narrowed down initial candidate genes to 2378 genes for the probe design, from which we calculated a mean pairwise distance (PD) at the amino acid level for each gene. We assumed that a gene with a low mean PD would be slow-evolving while those with a high mean PD would be fast-evolving. To explore phylogenetic signals in these genes, we created the following four datasets consisting of genes with different PDs: (1) 517 genes with 1–9% mean PD; (2) 990 genes with 10–19% mean PD; (3) 609 genes with 20–29% mean PD; and (4) 262 genes with 30–45% mean PD. After phylogenetic analyses of these datasets, we compared the results across the four resulting trees in terms of topology, nodal support, and branch lengths. All four analyses recovered the monophyly of Orthoptera and each of the two suborders but differed in resulting branch lengths and nodal support values (Fig. 1a–d, Supplementary Fig. S1). The analysis based on 517 genes with 1–9% mean PD (Fig. 1a) resulted in longer internodes for ensiferan taxa but very short internodes with lower support values for Caelifera, while the remaining three analyses (Fig. 1b–d) resulted in better resolution for Caelifera. Compared to the analysis based on slow-evolving genes (Fig. 1a), the placements of Grylloidea and Hagloidea shifted in the analyses with fast-evolving genes (Fig. 1b–d). The placements of Rhaphidophoroidea and Pamphagidae were congruent with the accepted classification only in the analysis with the fastest-evolving genes (Fig. 1d). We found that the genes with mean PD ranging between 10 and 45% could unambiguously resolve relationships across all major lineages within Orthoptera. Therefore, we selected 1853 orthologs with 10–45% PD as input data for our probe design to maximise the potential for phylogenetic resolution.

We collaborated with Daicel Arbor Biosciences (“Arbor”, Ann Arbor, MI, USA) to design and manufacture the OR-TE probe set. We first generated individual nucleotide alignments of the 1853 orthologs to identify regions that could be used as baits. We aimed to find baits that could capture only the regions present in each alignment but remain diverse enough to capture all potential species with the region in each alignment. We achieved this with the following clustering logic and two-stage design. For the first stage, we tiled 120-nucleotide baits every 20 nucleotides across each entry for each ortholog, generating hundreds of thousands of starting bait candidates. To reduce this complexity, we employed a ‘greedy’ clustering technique based on USEARCH³² to generate centroids representing several bait candidates within a given pairwise distance. To increase the chance that centroids were drawn from random selections of individual ortholog entry members, we first shuffled the

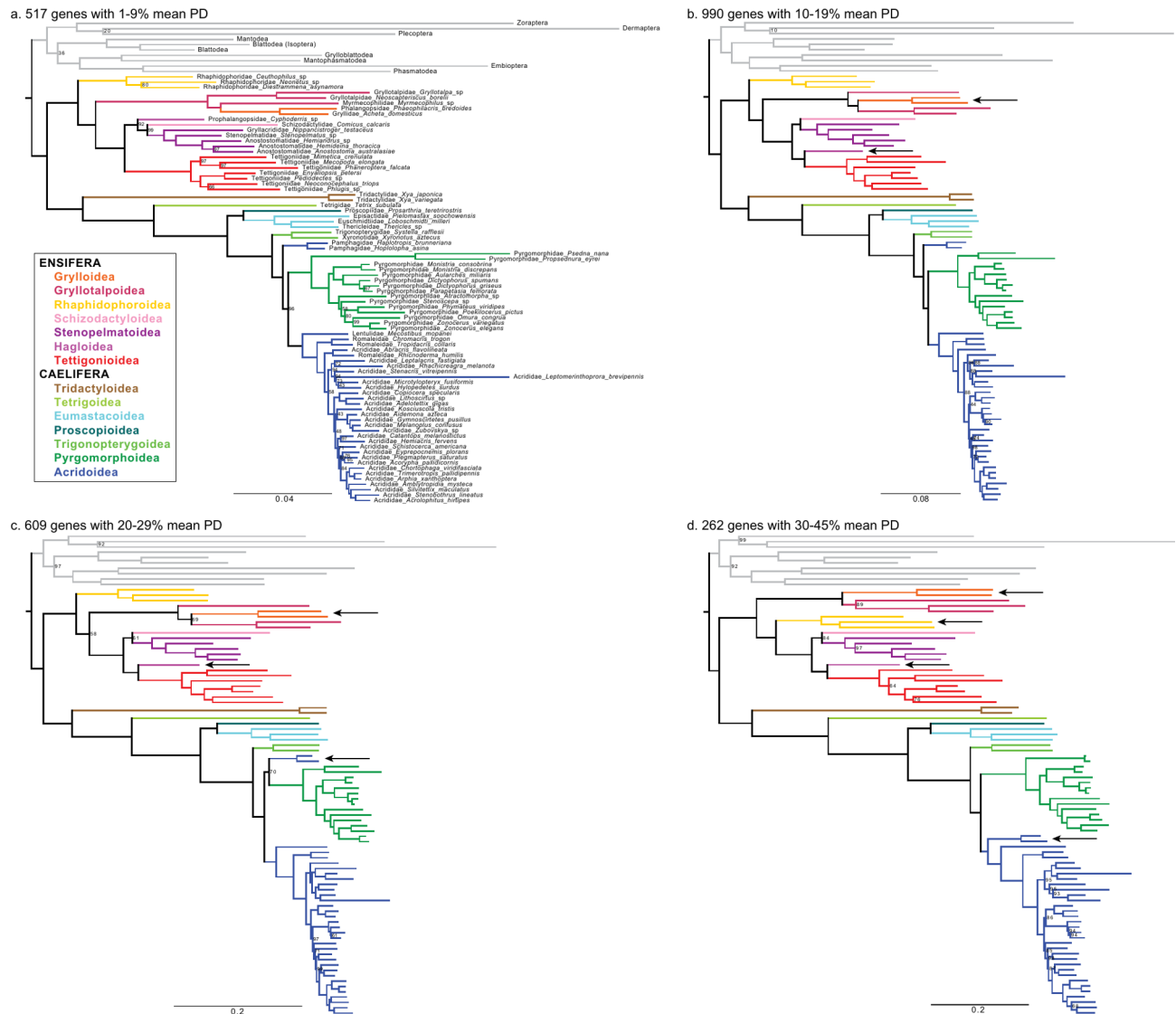


Fig. 1. Comparison of phylogenetic resolution across genes with different evolutionary rates. **(a)** Maximum likelihood (ML) tree based on amino acid sequences of 80 orthopterans and ten polyneopteran outgroups using 517 genes from transcriptomes that show 1–9% mean pairwise distance (PD). **(b)** ML tree based on 990 genes with 10–19% mean PD. **(c)** ML tree based on 609 genes with 20–29% mean PD. **(d)** ML tree based on 262 genes with 30–45% mean PD. For all of the trees, branches are colour-coded by superfamilies. Bootstrap values are shown as nodal support values, and where it is not shown, the value is 100. The arrows in **(b–d)** indicate clades whose phylogenetic placements change compared to the analysis with 517 genes with 1–9% mean PD.

starting bait sequences. Then, we clustered the baits into centroids tolerating up to 85% alignment divergence within a minimum 111 bp alignment overlap. The 25 centroids to which the most bait candidates collapsed were kept for each locus. Each centroid candidate was then searched against the reference transcriptomes using BLAST⁴³, scored for specificity using Arbor's proprietary method, and removed from the probe set if strong potential off-target hits were predicated for the centroid. This filtration resulted in 28,563 bait sequences. To further expand the probe set's capabilities to capture even more divergent sequences, we added back 11,246 bait sequences that diverged 20–30% from the original centroids. This final OR-TE probe set consisted of 39,809 120-nucleotide bait sequences that were designed to capture a total of 1828 loci. These were then manufactured as part of a myBaits[®] target capture kit by Arbor.

Capture efficiency of OR-TE probe set

To test the capture efficiency of the OR-TE probe set, we performed target enrichment and sequencing for 36 orthopteran taxa (18 caeliferan and 18 ensiferan species) representing 27 families and 11 superfamilies from across the order, specifically selected to examine capture efficiency across the order. We included representatives from the following 12 families, which had not been included in any previous phylogenomic study: Mogoplistidae,

Trigonidiidae, Ripterygidae, Cydrachetidae, Chorotypidae, Morabidae, Eumastacidae, Pamphagodidae, Lithidiidae, Pyrgacrididae, Tristiridae, and Ommexechidae. These taxa are rare and not frequently collected, and transcriptomes were not available for any of them at the time of probe design. Therefore, our taxon sampling represented a robust design to test the phylogenetic utility of the OR-TE probe set.

We recovered a total of 312.14 gigabytes (GB) of data from Illumina sequencing. The amount of raw data (unzipped paired fastq files) recovered per taxon ranged from 277 megabytes (MB) to 42.4 GB, with an average of 8.67 GB (Std Dev $\pm$ 7.14) (Supplementary Data File 2, Fig. 2a). Samples were divided between two pools to test the efficacy of multiplexing hybrid enrichment with different numbers of samples, 12-plex and 24-plex. Both pools resulted in statistically similar amounts of data (two-tailed t-test, $p=0.644$), with the 12-plex producing an average of 8.04 ($\pm$ 3.84) GB per sample and the 24-plex producing an average of 8.99 ($\pm$ 8.38) GB per sample (Supplementary Data File 2, Fig. 2b). The amount of data generated did not significantly differ between Caelifera (7.24 ± 4.53 GB) and Ensifera (10.10 ± 8.96 GB) (two-tailed t-test, $p=0.237$) (Fig. 2c).

The OR-TE probe set was designed to capture 1828 orthologs identified from the orthopteran transcriptomes. After filtering, assembling, and conducting an orthology search, the mean capture efficiency was 1037.53 ± 417.83 loci per taxon across the 36 taxa sampled. The capture efficiency (calculated as the number of corrected captured loci divided by 1828) varied from 15.81% (289 genes from a tetrigid *Metrodora reticulata*) to 92.51% (1691 genes from a lentulid *Eremidium sambaba*) (Supplementary Data File 2). The average capture efficiency was statistically higher for Caelifera (1241.50 ± 491.60 genes) than for Ensifera (833.56 ± 172.22 genes) (two-tailed t-test, $p=0.003$) (Fig. 2e). Interestingly, we found that the probes also captured a varying number of loci that were not included in the list of original target loci. We could identify these off-target captures because the Orthograph pipeline we used identified orthologs based on a reference gene set consisting of 5414 single-copy protein-coding genes. The off-target capture rate (calculated as the number of off-target loci divided by the total number of captured loci) ranged from 0.31% (*Hemicharilaus monomorphus*) to 25.25% (*Cyrtoxipha* sp.). When comparing the 12-plex capture reaction and the 24-plex capture reaction, we found that the capture efficiency of the former (1218.58 ± 371.06 genes) was higher than but not statistically different from the latter (947 ± 417.33 genes) (two-tailed t-test, $p=0.058$) (Fig. 2d). In terms of correlation between the amount of raw data and the number of captured loci, there was no strong correlation when all samples were examined (Pearson's correlation coefficient = 0.2287, $p=0.1796$), but we found a lineage-specific pattern in that they were positively correlated in Caelifera (Pearson's correlation coefficient = 0.7794, $p<0.0001$) but not strongly correlated in Ensifera (Pearson's correlation coefficient = 0.1880, $p=0.4548$).

Overlap of captured loci across taxa

Since each taxon recovered a different number of loci, we asked how many captured loci overlapped across the total taxon sampling. Using the 1KITE bioinformatics pipeline used in Song et al.¹⁰, we created gene-specific alignments for all recovered loci. We recovered a total of 2063 alignments that included at least two taxa, which was more than our intended number of target loci. This indicated that even some of the same off-target loci were recovered in multiple taxa. We also found that 730 loci were unaligned because they were unique for a given single taxon. Of these 730 loci, 720 loci were the off-target loci. Only 22 alignments included all 36 taxa, indicating 22 loci were captured for the entire taxon sampling. All remaining alignments had various numbers of missing taxa (Fig. 3). One thousand three hundred four alignments included at least 25% of the taxa sampled.

Phylogenetic utility of OR-TE probe set

We created four matrices to test the performance of the OR-TE probe set. The first two matrices (NT and AA) comprised 46 taxa, with ten polynopteran outgroups and 36 orthopteran ingroup taxa. The sequence data for the outgroups were obtained from transcriptome data, while the sequence data for the ingroup taxa were generated using the probe set. The 46-taxon NT dataset included 1,670,196 aligned nucleotide sequences, and the 46-taxon AA dataset included 556,732 aligned amino acid sequences. These concatenated matrices were constructed to test how well the target capture data would independently resolve higher-level relationships. The third and fourth matrices (NT and AA) comprised 126 taxa, with the same 46 taxa as the first dataset, combined with an additional 80 orthopteran taxa used to develop the OR-TE probe set. The 126-taxon NT dataset included 1,897,806 aligned bp, and the 126-taxon AA dataset included 632,602 aligned AA. These matrices were constructed to test the effect of combining target capture data with transcriptome data regarding the resulting topology. In addition, we also used 1304 individual gene trees that included at least 25% of the taxon sampling to estimate a species tree of the 36 ingroup species under the multi-species coalescent model.

The primary goal of the phylogenetic analyses presented here was to demonstrate the OR-TE probe set's utility and not to test phylogenetic hypotheses exhaustively. Nevertheless, several features of the resulting trees warrant note. All four phylogenetic analyses recovered Orthoptera as monophyletic and the two suborders, Caelifera and Ensifera, with maximal nodal support (bootstrap value of 100 or posterior probability value of 1.00). For the 46-taxon NT and AA matrices, the ML and the Bayesian analyses recovered identical topologies (Supplementary Fig. S2). Maximum likelihood and Bayesian inference also recovered identical topologies in analyses of the AA dataset (Supplementary Fig. S2). However, some differences existed between the relationships recovered from analyses of the NT and AA datasets. Within Ensifera, *Troglophilus neglectus* was placed at the base of the infraorder Tettigoniidea in the NT dataset (Fig. 4b). In contrast, it was placed at the base of the suborder Ensifera in the AA dataset (Fig. 4a). The position of *Glaphyrosoma beretka* also differed between the two datasets, nested within the Stenopelmidae and Anostomatidae in the AA dataset (Fig. 4a) versus grouped with Gryllacrididae in the NT dataset (Fig. 4b). The species tree analysis of the 36 ingroup species resulted in relationships largely congruent with the ingroup relationships of the 46-taxon NT dataset. The only difference was the relationship within the superfamily Acridoidea, especially the placement of *Pyrgacris descampsi* (Pyrgacrididae) and *Ixalidium* sp. (Acrididae) (Supplementary Fig. S3).

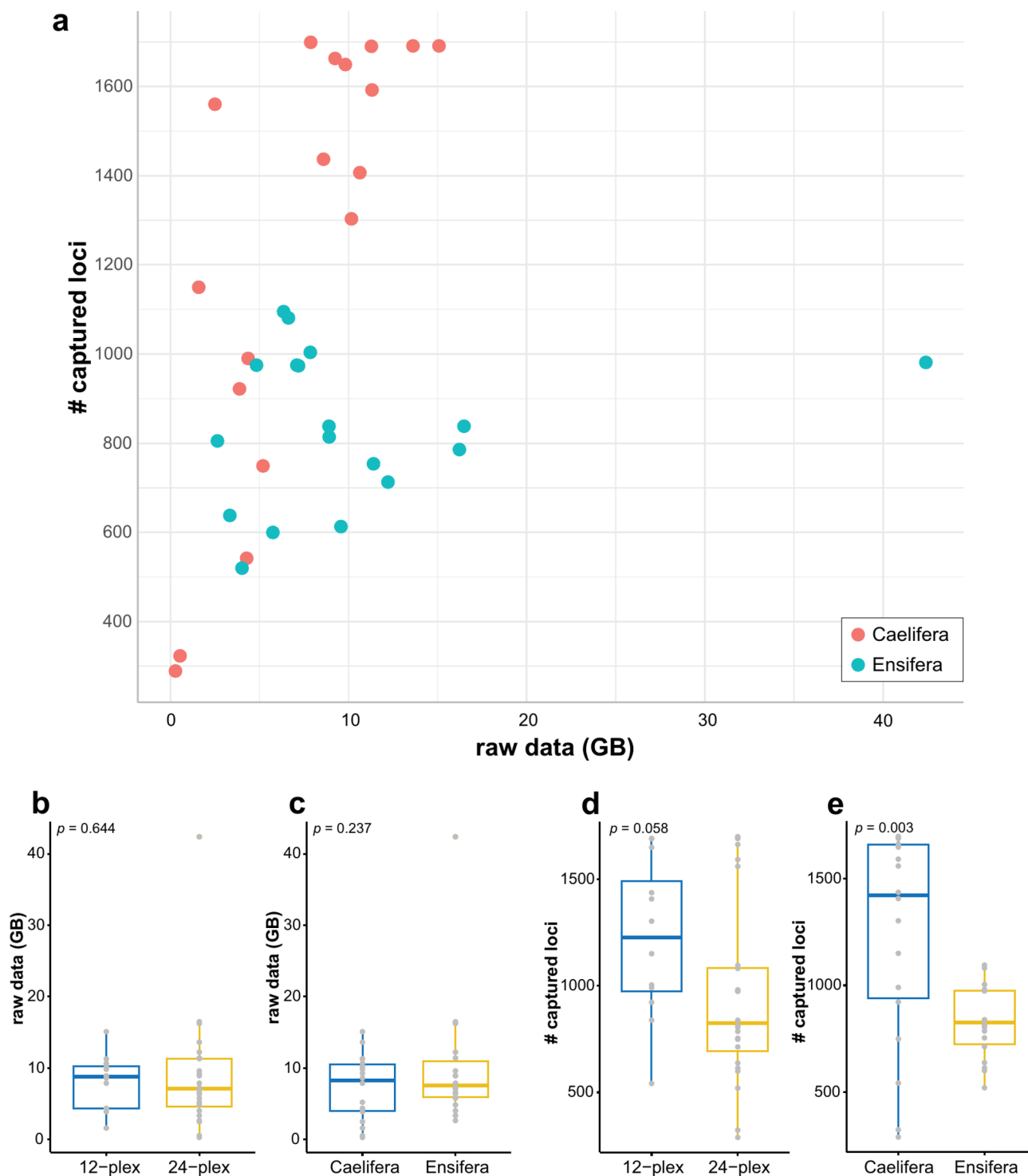


Fig. 2. Capture efficiency of the OR-TE probes. **(a)** Scatterplot showing the relationship between the number of captured loci and the amount of raw data produced from the 36 samples. The sole sample on the right is a cricket *Paroecanthus* sp. which recovered over 10 GB of data. **(b)** A comparison of the amount of raw data produced per sample from the 12-plex and 24-plex capture reactions. **(c)** A comparison of the raw data produced per sample for Caelifera and Ensifera. **(d)** A comparison of the number of loci recovered per sample in the 12-plex and 24-plex capture reactions. **(e)** A comparison of the number of loci recovered per sample for Caelifera and Ensifera.

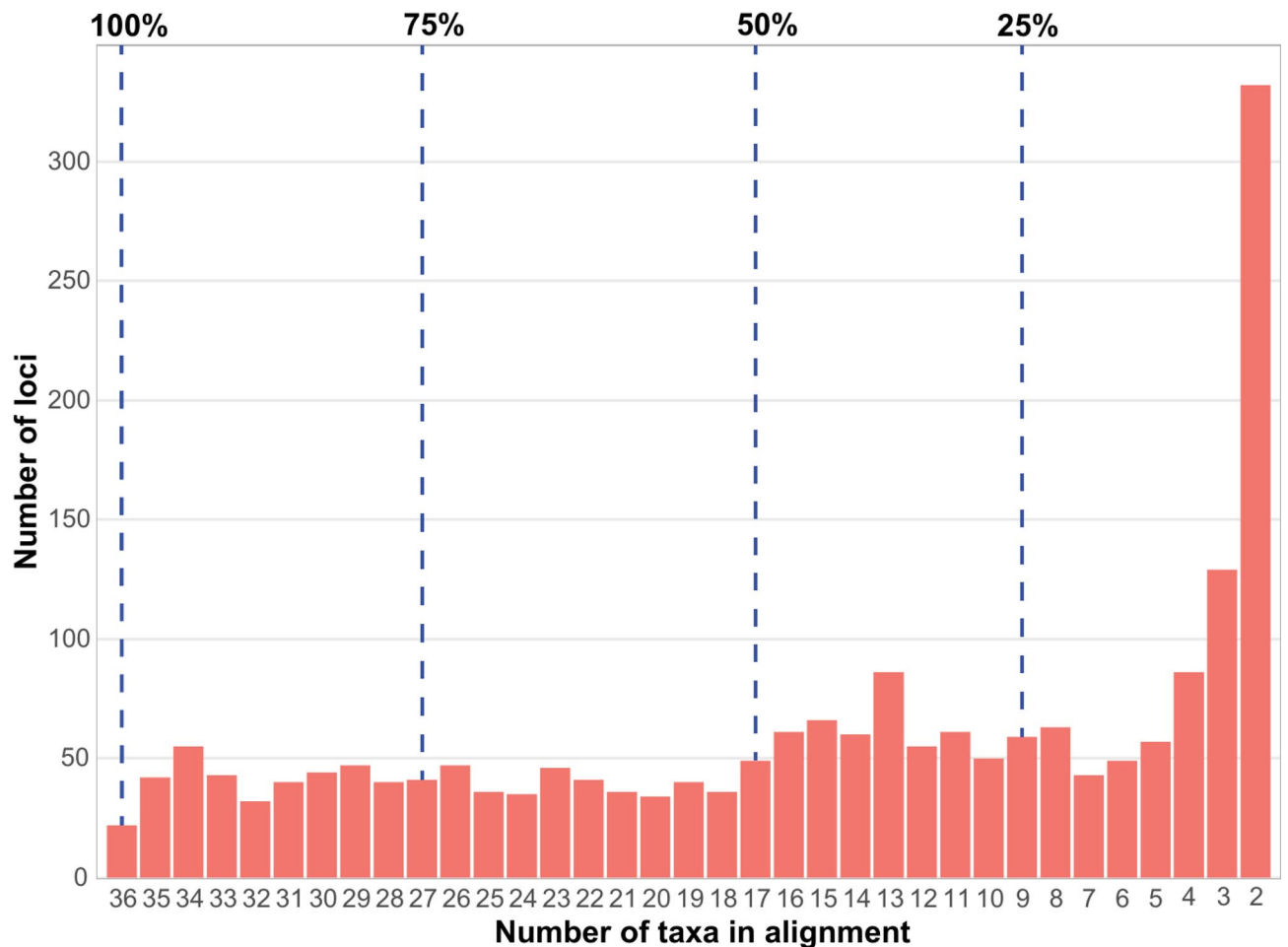


Fig. 3. Overlap of captured loci across taxa. A bar graph showing the distribution of individual alignments that included different numbers of taxon sampling arranged from complete taxon sampling (36 species) to two taxa. Each bar represents the number of captured loci included in the specific number of taxa sampled. The blue dotted lines represent the proportions (100%, 75%, 50%, 25%) of total taxa included in the alignments.

For the 126-taxon matrices, several topological incongruences existed between the NT and AA datasets (Supplementary Fig. S2). In the NT dataset (Fig. 5), the superfamily Rhaphidophoroidea was recovered as the earliest diverging lineage within the infraorder Tettigoniidea, but in the AA dataset, it was recovered as the earliest diverging lineage within Ensifera, both with high support. The placement of the superfamily Schizodactyloidea also differed between the two datasets, sister to the infraorder Tettigoniidea in the NT dataset (Fig. 5) and sister to Stenopelmatoidea (but with lower support) in the AA dataset (Supplementary Fig. S2). Within Caelifera, the placements of *Pyrgacris descampsi* (Pyrgacrididae) and *Rhcnoderma humilis* (Romaleidae) differed between trees generated from analyses of the two datasets. In analyses of both the NT and the AA datasets, the superfamily Acridoidea was recovered as paraphyletic because of the placement of the clade consisting of Pamphagidae and Pamphagodidae.

Regarding the topology, the NT datasets of both the 46-taxon (Fig. 4b) and the 126-taxon matrices (Fig. 5) resulted in higher-level relationships that were largely congruent with the most recent phylogenomic study¹⁰ with some exceptions. The NT datasets recovered two monophyletic ensiferan infraorders, Gryllidea and Tettigoniidea, and two monophyletic caeliferan infraorders, Tridactylidea and Acrididea, and superfamily-level relationships were consistent with the previous finding. Within Caelifera, *Ixalidium* sp. (Acrididae), which was expected to be recovered as sister to *Ommexecha virens*, was recovered in unexpected positions. The 46-taxon NT dataset did not include any member of the family Pyrgomorphidae, but the 126-taxon NT dataset included many members of this family whose data came from transcriptomes. Pyrgomorphidae is the sole member of the superfamily Pyrgomorpoidea, which is currently hypothesised to be sister to Acridoidea. Still, in our analysis, this family is nested within Acridoidea, rendering the latter superfamily paraphyletic. Within Ensifera, Anostomatidae and Stenopelmatoidea were paraphyletic. Analyses of the AA datasets of both matrices recovered the infraorder Tettigoniidea as paraphyletic because Rhaphidophoroidea was at the base of the Ensifera, which conflicts with the currently accepted classification adopted by the Orthoptera Species File²⁶.

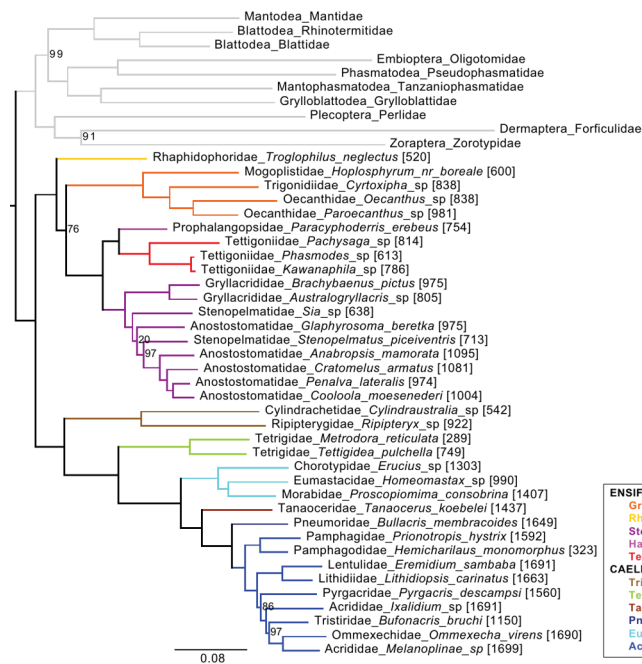
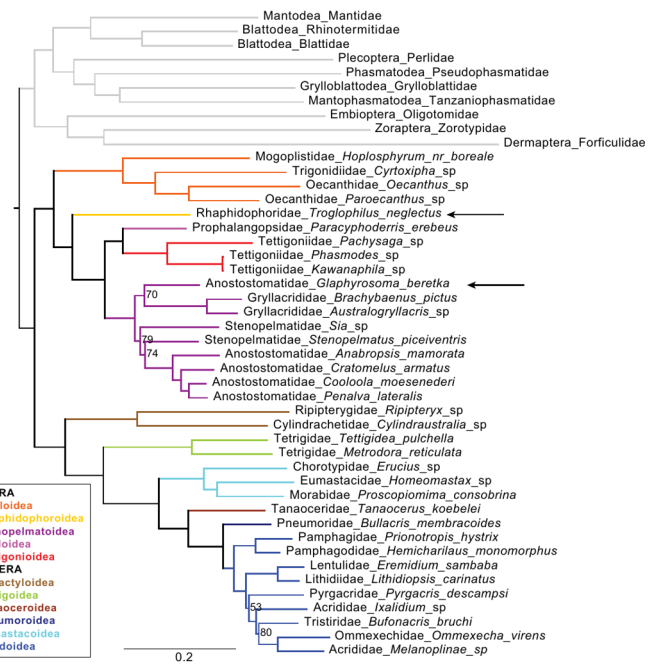
a. 46_taxa AA dataset**b. 46_taxa NT dataset**

Fig. 4. Phylogenetic trees for Orthoptera resulting from analyses of target captured data for 36 species (ingroup only). **(a)** ML tree based on analysis of amino acid sequences. **(b)** ML tree based on analysis of nucleotide sequences. For all of the trees, branches are colour-coded by superfamilies. Bootstrap support values less than 100 are shown. The arrows in b indicate clades whose phylogenetic placement differs from the amino acid tree. Data for the outgroups came from transcriptome data.

Discussion

Our primary goal was to develop Orthoptera-specific target hybrid enrichment probes that could capture hundreds of phylogenetically informative loci from any orthopteran species and resolve relationships across broad phylogenetic scales. We generated such a probe set using the myBaits technology to target 1828 loci, including both fast-evolving and slow-evolving genes, designed from 80 orthopteran transcriptomes. We named our new probe set the OR-TE (Orthoptera Target Enrichment) probe set. We have shown that our OR-TE probe set can reliably capture an average of 1037 loci from diverse orthopteran lineages and resolve expected phylogenetic relationships across broad timescales. Notably, the probe set successfully captured loci from the 13 families that were not included in the probe design, which demonstrates the robustness of our design.

Some unique features of the OR-TE probe set deserve further discussion. While the probe set can capture hundreds of loci across all lineages within Orthoptera, it is more efficient in capturing target loci from Caelifera than Ensifera (Fig. 2e). The transcriptome data that we used for designing the probe set were biased toward Caelifera (58 species vs. 22 ensiferan species) mainly because our available data included many more grasshoppers belonging to two particular families (28 Acrididae and 15 Pyrgomorphidae). This bias in the initial design stage may have contributed to the differences in capture efficiency between the suborders. Nevertheless, the fact that we captured an average of 833 loci from Ensifera demonstrates the utility of our OR-TE probe set.

Another attractive feature of this probe set is that a single capture reaction can multiplex up to 24 libraries, with each library representing one taxon. Although we did find that the 24-plex capture reaction yielded slightly fewer loci than the 12-plex capture reaction (Fig. 2b,d), the difference was not statistically significant. This finding is relevant for reducing the cost of data generation because it demonstrates that one can likely use half the number of capture reactions to generate a comparable amount of data. One caveat is that the quality and quantity of DNA should be sufficiently high enough (1 µg high molecular weight genomic DNA) to reliably multiplex up to 24 samples per capture reaction. If degraded DNA (i.e., from dried museum specimens) is used, it is recommended by the manufacturer (Daicel Arbor Biosciences) to lower the number of samples (e.g., 12 samples per capture reaction) for multiplexing.

We also show a relatively high capture efficiency regardless of the amount of raw data sequenced, except when the raw data are minimal. We found that the two taxa with the smallest number of captured loci (*Hemicharila monomorphus* and *Metrodora reticulata*) started with the least raw data (Supplementary Data File 2). All other taxa that we included started with as little as 1.59 GB raw data (*Bufonacris bruchi*) to as much as 42.40 GB raw data (*Paroecanthus* sp.), and all yielded comparable numbers of captured loci, although the correlation between the amount of raw data and the number of captured loci was positive in Caelifera. We were initially concerned that the large genome sizes of Orthoptera would somehow hinder efficient target enrichment, but our results demonstrate that the high capture efficiency of the probes can yield a sufficient amount of data across the phylogeny of Orthoptera.

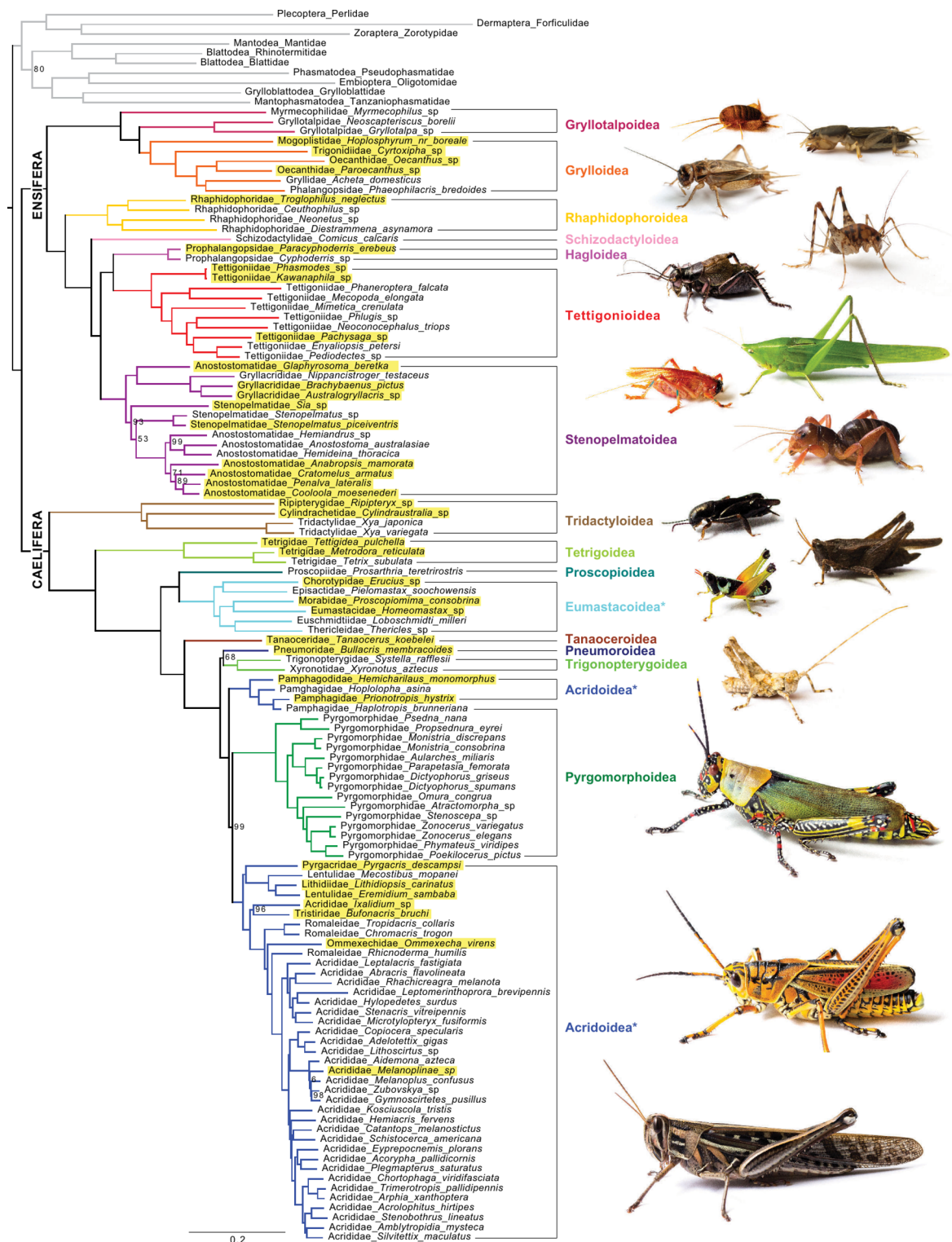


Fig. 5. Phylogenetic tree for Orthoptera resulting from a combined analysis of phylogenomic data obtained via target enrichment and RNASeq. ML tree based on analysis of nucleotide sequences. Branches are colour-coded by superfamilies. Bootstrap support values less than 100 are shown. Terminals highlighted in yellow were included based on target enrichment data. Asterisks indicate paraphyly. Photo credits: Brandon Woo.

Our OR-TE probe set will enable users to resolve higher-level relationships (family, superfamily) and lower-level relationships (genus, species). This study demonstrated the probe set's ability to resolve higher-level relationships. We found a small number of taxa whose recovered phylogenetic positions were unexpected and that the NT and the AA datasets recovered phylogenetic trees with some topological differences. However, this

is not necessarily due to an issue with our probe design. Previously, we successfully used the OR-TE probe set to generate phylogenomic data to resolve species-level relationships in the Jerusalem cricket genus *Stenopelmatus*³⁴ and the lentulid grasshopper genus *Eremidium* (Song, unpublished). We also confirmed the probe set's ability to capture targeted loci from dried museum specimens of various cricket species (Song, unpublished). Furthermore, the data generated using the OR-TE probe set could be combined in the future because the bioinformatics pipeline relies on a pre-defined reference gene set. This means we can continue adding taxa to the existing datasets to achieve greater resolution. The OR-TE probe set is thus a highly versatile tool that is useful at multiple taxonomic levels.

We designed the OR-TE probe set by identifying loci from 80 orthopteran transcriptomes across the phylogeny. Transcriptome data inherently include only the mRNA transcripts after splicing. Genomic DNA, a starting point for target enrichment, consists of both exons, which the baits can hybridise with, and introns, which the baits will not match. Orthoptera are known to have very large genome sizes¹⁶ and long intronic regions between the coding regions³⁵. Because the OR-TE probe set was designed from post-splicing mRNA transcripts, some of these baits might target exons separated by long introns in genomic DNA. In such cases, the baits will not fully hybridise with the target loci using genomic DNA and, therefore, be less efficient in binding to and pulling down molecules. This is a potential limitation of using the OR-TE probe set, although our baits are tiled, which should increase the chances of successful capture. Still, given that we could capture an average of 1009 loci per taxon across the phylogeny, our baits must target many exons uninterrupted by intronic regions.

The genome sizes of Orthoptera are known to be the largest among insects^{16,29}, with a potentially large number of gene duplication and paralogs³⁵. Although the OR-TE probe set was designed based on the known orthologs based on annotated orthologs of five insect species¹⁰, a possibility exists that some of these target genes may have gone through gene duplication events in some orthopteran species. In such a case, it is possible that some baits may hybridise with the duplicated genes with enough sequence similarities instead of the actual target orthologs. Without the availability of well-annotated reference genomes, which is typically the case for most phylogenomic studies, it is difficult to predict how frequently this scenario occurs. This challenging problem affects not only our probe set but all hybridisation-based target enrichment methods.

Nevertheless, the Orthograph pipeline we used is a graph-based orthology prediction algorithm that can accurately search single-copy orthologs while rigorously filtering out paralogs³⁶. Specifically, this pipeline maps transcripts to the globally best-matching orthologous groups (OGs) rather than inferring the orthology of these genes de novo from the transcripts, which circumvents redundant transcript assignments. The pipeline is very efficient in reliably identifying orthologs, detecting known paralogs, and finding known isoforms or alternative transcripts³⁶. Moreover, the Orthograph pipeline has been successfully used in numerous other insect phylogenomic studies utilizing target enrichment methods^{21,23,37}. There are other methods for removing paralogs in target enrichment data³⁸, but often, these are pipeline-specific and difficult to implement across studies. Therefore, it is unlikely that our final datasets included many undetected paralogs that could seriously compromise phylogenetic inferences.

The number of loci captured per taxon differed widely (Fig. 3). Even for the same locus, the lengths of captured regions often differed across taxa. This unequal recovery of loci across samples is a general feature of most hybrid enrichment techniques used for phylogenetically diverse taxa³⁹. These gene and taxon sampling differences would naturally lead to a large amount of missing data in the final concatenated dataset. The negative effect of missing data in phylogenomic analyses is an important issue, which has been investigated for analyses using RADseq data⁴⁰, UCE data^{39,41}, and AHE data⁴². In general, these studies have concluded that uneven missing data can potentially lead to spurious phylogenetic inference, and we agree that the effect of missing data should be explored in depth, especially for large-scale phylogenomic studies.

Interestingly, the unequal number of captured loci seemed to have little impact on phylogeny estimation in our taxon sampling. For instance, we could only capture 289 loci for the pygmy grasshopper *Metrodora reticulata*, but the placement of this taxon within Tettigoniidae was consistent in all our datasets. The total number of recovered nucleotides for this species was 40,905 bp. As a comparison, we recovered 1691 loci for the grasshopper *Eremidium sambaba*, which collectively included 414,858 bp, and the phylogenetic placement of this species was also very consistent. This observation suggests that the relatively low number of captured loci, which still comprise tens of thousands of nucleotides, likely contained sufficient phylogenetic information to place the species within the phylogeny correctly.

Although our taxon sample and limited analysis were intentionally designed to test the phylogenetic utility of the OR-TE probe set rather than to test previous phylogenetic hypotheses, the resulting topology nonetheless revealed a few novel insights regarding the diversification of Orthoptera.

The phylogenetic position of Rhaphidophoridae, the sole member of the superfamily Rhaphidophoroidea (cave crickets and camel crickets), may need a critical re-evaluation. Traditionally, the suborder Ensifera is considered to consist of two infraorders, Gryllidea and Tettigoniidea, based on both morphological and molecular evidence²⁸, and the most recent phylogenomic analysis recovered Rhaphidophoridae as the earliest diverging lineage within Tettigoniidea¹⁰. Morphologically, it is uniquely different from other ensiferans in that it is the only completely apterous family without the ability or the structures to produce sound and hear⁴³. In our dataset, this family's placement within Ensifera changed depending on whether the data were analysed as nucleotides or amino acids. When the NT dataset was used, it was recovered at the base of Tettigoniidea (Fig. 4b). Still, when the AA dataset was used, it was recovered at the base of Ensifera (Fig. 4a).

Interestingly, when comparing the four transcriptome-based datasets differing in PDs used for exploring phylogenetic signals of different loci (Fig. 1a–d, Supplementary Fig. S1), all of which were coded as amino acids, we recovered Rhaphidophoridae at the base of Ensifera in three datasets with the slow-evolving loci (1–9%, 10–19%, 20–29% mean PD), while in the expected position at the base of Tettigoniidea in one dataset with the fastest-evolving loci (30–45% mean PD). These observations collectively suggest that Rhaphidophoridae as a

lineage could have experienced different rates of molecular evolution compared to other ensiferan lineages, which could have affected phylogenetic estimation. However, because these observations were based on a minimal sampling of the family, it is difficult to make a definitive statement about the cause of this discrepancy. How the data will behave when a much more extensive taxon sampling is included in the future remains to be seen, as the phylogenetic position of Rhaphidophoridae within Ensifera is essential for inferring the evolution of sound production and hearing.

The monophyly of the families belonging to the superfamily Stenopelmatoidea needs to be critically tested. Stenopelmatoidea currently includes three families: Stenopelmatidae (Jerusalem crickets), Gryllacrididae (raspy crickets), and Anostomatidae (king crickets, wetas, and Cooloola monsters). This superfamily includes about 1200 described species that are morphologically diverse and ecologically interesting⁴⁴ but remains poorly studied compared to other ensiferan groups, such as crickets and katydids. Previously, Vandergast et al.⁴⁴ conducted a large-scale molecular phylogenetic study of the superfamily based on three loci and found Gryllacrididae to be monophyletic and the remaining three families paraphyletic. In our 46-taxon matrices, we included nine taxa belonging to Stenopelmatoidea and found the superfamily to be monophyletic but Anostomatidae and Stenopelmatidae to be paraphyletic (Fig. 4). Gryllacrididae was recovered as monophyletic as expected, but this clade was recovered as a sister to a Central American anostomatid *Glaphyrosoma beretka*, which did not group with other anostomatids. Anostomatidae shows a classic Gondwanan distribution, and most species within the family, except the New Zealand endemic wetas, are remarkably similar in terms of morphology^{44,45}. This morphological convergence could have contributed to the current state of classification. We included just two representatives of Stenopelmatidae, *Stenopelmatus piceiventris* from Mexico and *Sia* sp. from South Africa. Still, they did not form a monophyletic group, which was also the pattern found in Vandergast et al.⁴⁴. Our taxon sampling is too small to suggest a reclassification of the superfamily—still, our results point to the need to evaluate the current classification of Stenopelmatoidea further.

Finally, the superfamily-level relationships within Caelifera recovered using the OR-TE probe set largely agree with the current phylogenetic understanding. We recovered Tridactyloidea as the earliest diverging lineage, followed by Tetrigoidea, congruent with all previously published molecular phylogenies^{10,28}. The phylogenetic relationships within the superfamily group Acridomorpha were largely consistent with all previous studies. The superfamily Acridoidea was not recovered as monophyletic in several analyses, including those using the most data (Fig. 5), because Pyrgomorpha was placed within Acridoidea. There is strong morphological evidence that Acridoidea is monophyletic, especially based on male internal genitalia⁴⁶, and thus, our results were not congruent with the current classification. Specifically, the phylogenetic position of Pamphagidae seems unstable, which was previously observed in a transcriptome-based phylogenomic study¹⁰. Song et al.¹⁰ used four-cluster likelihood mapping⁴⁷ and permutation tests to check for among-lineage heterogeneity and non-random distribution of missing data and found that the phylogenetic position of Pamphagidae was potentially biased by the confounding signal. The current study used the same data as the previous study¹⁰, and therefore, it is possible that our taxon sampling may suffer from the same bias. A more comprehensive study is required to resolve the phylogenetic position of Pamphagidae fully. The placement of *Ixalidium* sp., a wingless grasshopper from East Africa, is also unexpected. This genus was originally assigned to the subfamily Catantopinae within the Acrididae⁴⁸ based on external and internal morphological characters of doubtful value. A detailed examination of the male genitalia confirmed that this genus was distinctly different from the Acrididae (Hemp et al. submitted). Based on this unexpected phylogenetic placement from our study, a follow-up study was carried out to clarify the taxonomic status of this genus, which eventually led to the description of a new family of grasshoppers soon to be named Ixalidiidae (Hemp et al. submitted). This finding demonstrates the utility of the OR-TE probe set, which revealed the presence of previously unrecognised lineage.

Target enrichment was introduced as a revolutionary new technique for molecular phylogenetics a decade ago^{12,13}, and numerous insect phylogenomic studies have been published using this technique. However, it has remained challenging for researchers familiar with traditional molecular data generation (PCR and Sanger sequencing) to incorporate this new technique into their tool kits. The reasons for this challenge vary but are mainly due to cost and resource availability. Below, we describe our experience developing this tool and the specific costs associated with each step to paint a realistic picture.

The first major cost for developing the OR-TE probe set was the expenses associated with generating transcriptomic data to use as a genomic resource for identifying targeted loci. Of the 80 orthopteran transcriptomes, the 1KITE project had previously generated 41, and nine were otherwise previously published¹⁰, and thus freely available. However, because the available transcriptome data did not cover the phylogenetic diversity within the order, we set out to include additional taxa to achieve diverse taxon sampling, especially by adding different subfamilies of Acrididae and Tettigoniidae, as well as previously unsampled families. Because RNA-grade samples had to be freshly collected from the field and directly preserved in RNAlater, we conducted domestic and international expeditions to collect live specimens. The costs associated with collecting expeditions are often not incorporated into the calculation of data generation, but it is a significant expense that cannot be overlooked. The cost of RNA extraction was estimated to be about \$10 USD per sample, and at the time of data generation, we decided to outsource library preparation to a sequencing core. Library preparation cost was ~\$180 per sample, and three lanes of HiSeq4000 cost \$7200. Thus, the upfront cost of building genomic resources by generating 30 new transcriptomes, excluding the costs associated with sample acquisition, was ~\$12,900. This was a significant but necessary investment because there were not many orthopteran transcriptomes available from public databases at the time of this project. Now, these data are freely available for anyone to use.

The next significant costs were associated with manufacturing the baits and testing the capture efficiency of our probe set. There was no cost for designing the custom baits because we performed the bioinformatics to identify the targeted loci ourselves, and the bait design was performed collaboratively with Arbor. The smallest unit of myBaits custom target capture kit that we could purchase was the 16-capture-reaction kit, which cost

\$3240 at the time of purchase. We estimated the cost of our high molecular weight DNA extraction to be about \$5 per sample. We outsourced library preparation, target capture reaction, and Illumina sequencing to Arbor, and the total cost for generating data from the 36 taxa was ~\$6985, not including the cost of the target capture kit.

The specific dollar amount described here is possibly prohibitive for a typical research lab to generate phylogenomic data just for 36 taxa. We intended to pay the upfront costs of developing a new tool so that the researchers interested in the phylogenomics of Orthoptera do not have to be burdened and duplicate the efforts. Considering the amount of data generated per dollar, the OR-TE probe set is potentially a cost-effective approach for generating phylogenomic data for Orthoptera. If one outsources all the steps of data generation beyond DNA extraction, we estimate the cost per sample to be around \$150–\$200 to generate about two gigabases of data per sample. If one can perform library preparation and target capture in-house, the cost would drop below \$100 per sample. This means the cost per gene would be \$0.15 to \$0.20 if all the steps are outsourced. Target enrichment offers an exceptional value because the cost of data generation by PCR and Sanger sequencing would be about \$5 to \$10 per gene. We anticipate that the cost of sequencing will come down more in the future, making this approach more affordable.

One important resource required for data processing and analyses of target enrichment data is appropriate computational infrastructure with bioinformatics expertise. We were fortunate to have institutional access to high-performance research computing (HPRC) clusters, which allowed us to handle a large amount of molecular data effectively and run pipelines in parallel. The target captured data are raw Illumina reads, which must go through initial quality control, filtering, de novo assembly, and orthology search before undertaking phylogenomic analyses. Because of the sheer data size, all of the bioinformatics pipelines need to be run using the HPRC clusters, and it is nearly impossible to process data on a single desktop computer. The computational time necessary for all downstream analyses, including alignment, data partitioning, and phylogenetic analyses, depends on the project's scope. Still, these analyses do require significant computational resources as well. It is difficult to put monetary values to these computational usages. Still, our data processing and analyses would not have been possible if we did not have access to these resources. Thus, access to HPRC is a potential limiting factor for widely using the OR-TE probe set, especially for those researchers without these computational resources. However, this can be overcome by collaboration with those who have access or by using freely available computational resources, such as the U.S. National Science Foundation-funded ACCESS (<https://access-ci.org/about/>).

In conclusion, we have developed a new phylogenomic tool using target hybrid enrichment specifically for Orthoptera that can resolve relationships over broad phylogenetic scales, thereby advancing the systematics of this important order of insects. With the capacity to reliably capture over 1000 loci from any orthopteran taxa, our approach is the most cost-effective method for generating phylogenomic data within Orthoptera. While we have delved into its utilities and limitations, the analytical challenges associated with uneven missing data across taxa require more rigorous exploration. We envision widespread adoption of this new tool for the future of orthopteran phylogenetic studies, ushering in a new era of discovery and knowledge in this field.

Methods

Transcriptome sequencing and assembly

The newly sequenced 30 samples were originally collected in RNAlater in the field and kept at -20°C until RNA extraction. We followed the same procedures for RNA extraction and de novo transcriptome assembly described in Song et al.¹⁰. Briefly, RNA was extracted using a Trizol-chloroform extraction, followed by a clean-up with an RNeasy mini kit (Qiagen, Valencia, CA). RNA concentrations were measured with a spectrophotometer (DS-11, DeNovix, Wilmington, DE), and RNA integrity was analysed with a Fragment Analyzer (Agilent Technologies, Ankeny, IA). Library preparation, sequencing, and pre-processing were all performed at Texas A&M AgriLife Research Genomics and Bioinformatics Service. Illumina's TruSeq Stranded Total RNA Library Prep Kit was used for library preparation, and paired-end sequencing (150 bp) was performed using three lanes on an Illumina HiSeq4000 (San Diego, CA). Raw reads were imported into a personalised Galaxy environment on a supercomputing cluster of the High-Performance Research Computing group of Texas A&M University (Ada, <https://hprc.tamu.edu>) for trimming and quality check. We transformed reads to Sanger format with FastQ Groover⁴⁹ and filtered them using Trimmomatic⁵⁰. In Trimmomatic, bases were trimmed at both ends if their quality score was lower than 30, whole reads were trimmed with a sliding window of 3 bases and a minimum average quality score of 30, and finally all reads of less than 30 bp were discarded. Subsequently, FastQ Screen⁵¹ was used to filter out reads from bacterial and other contaminating sources (UniVec core (June 6, 2015), PhiX (NC_001422.1), Illumina adapters, *Gregarina niphandrodes* genome (GNI3), *Encephalitozoon romaleae* genome (ASM28003v2), *Escherichia coli* genome (K12), *Methylobacterium* sp., *Bosea* sp., *Bradyrhizobium* sp., *Klebsiella pneumoniae*, *Sphingomonas* sp., *Rhodopseudomonas* sp. and *Propionibacterium acnes*). The filtered reads were used for de novo transcriptome assembly using Trinity v2.2.07⁵². Details on all samples used for this study are provided in Supplementary Table S1 and can be found at the National Center for Biotechnology Information (NCBI) under the respective BioSample numbers.

Selection of target genes

Using the newly assembled transcriptomes as well as the previously generated data¹⁰, we built the reference orthologous gene groups originally identified by Song et al.¹⁰ using the following reference genomes on OrthoDB v7^{53–55}: *Acyrtosiphon pisum* (Hemiptera)⁵⁶, *Nasonia vitripennis* (Hymenoptera)⁵⁷, *Pediculus humanus* (Psocodea)⁵⁸, *Rhodnius prolixus* (Hemiptera)⁵⁹, and *Zootermopsis nevadensis* (Blattodea)⁶⁰. These taxa were chosen based on annotation and sequence quality and because of the lack of availability of reference genomes in Orthoptera or other closely related polyneopteran orders at the beginning of this project. Orthologous gene IDs

were clustered for node Insecta on OrthoDB, and single-copy genes across the five reference taxa were listed. We used the Orthograph pipeline described in Song et al.¹⁰ to identify orthologs from the 90 transcriptomes (80 orthopteran and 10 outgroups). In total, 5414 single-copy protein-coding genes were identified as a reference gene set and used to recover orthologs from the transcriptome data. From the initially identified orthologs, we filtered further to select specific genes that satisfied the following conditions: (1) G.C. content > 40%, (2) length of nucleotide sequences 500–1500 bp (166–500 amino acids), (3) taxon coverage > 60%, and (4) proportion of parsimony informative sites > 30%. From the 2378 genes that met these conditions, we calculated a mean pairwise distance (PD) for each gene using MEGA X⁶¹ to determine its evolutionary rate. To explore phylogenetic signals in these genes, we created four datasets consisting of genes with different PDs. For each dataset, individual genes were aligned as amino acids using MAFFT⁶², concatenated into a single unpartitioned dataset, and subjected to a maximum likelihood (ML) analysis using LG + G as a model of amino acid substitution in RAxML v. 8.2.12⁶³ with 100 standard bootstrap replications. After the phylogenetic analyses, we compared the results across the four resulting trees regarding topology, nodal support, and branch lengths.

Target enrichment

To extract the high molecular weight genomic DNA required for target enrichment and Illumina sequencing (at least 1 µg of DNA), we used Gentra Puregene Tissue Kit (Qiagen) following the manufacturer's guidelines. The quality and concentration of DNA extracts were initially measured using DeNovix Spectrophotometer. The genomic DNA extracts were sent to Arbor to be processed as part of their myReads[®] targeted sequencing services. They performed initial quality control, library preparation, and target enrichment following their manual for myBaits (<https://arborbiosci.com/wp-content/uploads/2018/04/myBaits-Manual-v4.pdf>). After the target enrichment, the samples were sequenced on a single lane of Illumina HiSeq2500 using 125 bp paired-end (PE) sequencing by Novogene (Sacramento, CA).

Bioinformatics pipelines

We trimmed and assembled the raw reads using Trimmomatic⁵⁰ and SOAPdenovo2⁶⁴, respectively, after which contamination was checked for the entire assembly by VecScreen (<http://www.ncbi.nlm.nih.gov/tools/vecscreen/>) and the UniVec database build 7.1 (<http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/>) following Misof et al.² and Peters et al.⁶⁵ Cross-contamination was checked for all taxa by BLAST. Orthograph³⁶ was used for orthology assessment using non-strict reciprocal searches and default parameters, and problematic genes were filtered with 'outlier' custom scripts, which removed outlier sequences for whole gene sets instead of individual outlier sequences to prevent possible phylogenetic rogue regions^{2,22} after MAFFT (v7.130b) alignment using the L-INS-I option⁶². As described above, all 5414 reference genes from the five insect genomes used by Song et al.¹⁰ were used to predict orthology. The filtered sequences were masked with Aliscore v1.2^{66–69} using default sliding window size and following options used by Peters et al.²², which identifies putative alignment ambiguities or randomised multiple sequence alignment (M.S.A) sections in the alignments for each gene. The problematic sequences and positions were removed individually using Alicut (<https://github.com/PatrickKueck/AliCUT>). After generating filtered amino acid (AA) alignments, the nucleotide (NT) data were filtered using the same procedure as the processed AA dataset. The NT alignment was generated using Pal2Nal v14⁷⁰, corresponding to the AA alignments. These alignments were concatenated into supermatrices for downstream analyses.

Phylogenomic analysis

PartitionFinder 2.1.1⁷¹ was used to identify the model-based combination of blocks from each of the four matrices. Parameters were: model_selection = AICc; branch_lengths = linked; search = rcluster, and default for all others; weight options as rate = 1.0, base = 1.0, model = 0.0, alpha = 1.0 with rcluster-percent = 0.1. The models calculated by PartitionFinder were based on the RAxML option, which only selected the best-fitting models RAxML could use. We inferred the phylogenetic relationships in a maximum likelihood framework by using RAxML⁶³. Because RAxML could not handle mixed model partitioned analyses, we used GTR + I + G for NT and LG + G for AA datasets. We ran 1000 standard bootstraps with ten individual ML searches to find the best tree with RAxML. For the 46-taxa matrices, we also inferred relationships in a Bayesian framework, using the best-fitting model for each partition as suggested by PartitionFinder for both NT and AA datasets. We ran 10,000,000 generations, sampling every 1000 using four chains with MrBayes using default priors. Tracer⁷² was used to check the effective sample size (ESS) for each node. We removed 25% as burn-in, which resulted in 7500 trees to summarise. For species tree analyses, we selected 1304 individual nucleotide alignments that included at least 25% of the taxon sampling from the bioinformatics pipeline described above. We estimated individual gene trees in a maximum likelihood framework in RAxML using the GTR + CAT model with 100 rapid bootstraps. The resulting 1034 gene trees were compiled into a single input data file and modified to collapse any node with a bootstrap support value of less than 10. We used ASTRAL ver. 5.7.8⁷³ to estimate the species tree. All resulting trees were examined in FigTree⁷⁴.

Data availability

NCBI BioProject and BioSample accession numbers are provided in online supplementary data. All raw and processed data for developing Orthoptera-specific target enrichment probe set as well as phylogenetic datasets have been deposited to figshare (<https://doi.org/10.6084/m9.figshare.26084731>). The OR-TE probe set is assigned the Arbor Design ID D10583GRHP2.

Received: 1 February 2024; Accepted: 9 September 2024

Published online: 13 September 2024

References

1. Johnson, K. P. Putting the genome in insect phylogenomics. *Curr. Opin. Insect Sci.* **36**, 111–117. <https://doi.org/10.1016/j.cois.2019.08.002> (2019).
2. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
3. Yeates, D. K., Meusemann, K., Trautwein, M., Wiegmann, B. & Zwick, A. Power, resolution and bias: Recent advances in insect phylogeny driven by the genomic revolution. *Curr. Opin. Insect Sci.* **13**, 16–23. <https://doi.org/10.1016/j.cois.2015.10.007> (2016).
4. Chester, D. The phylogeny of insects in the data-driven era. *Syst. Entomol.* **45**, 540–551 (2020).
5. Blaimer, B. B. *et al.* Key innovations and the diversification of Hymenoptera. *Nat. Commun.* **14**, 1212. <https://doi.org/10.1038/s41467-023-36868-4> (2023).
6. Johnson, K. P. *et al.* Phylogenomics and the evolution of hemipteroid insects. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12775–12780. <https://doi.org/10.1073/pnas.1815820115> (2018).
7. Kawahara, A. Y. *et al.* Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22657–22663 (2019).
8. Kutty, S. N., Wong, W. H., Meusemann, K., Meier, R. & Cranston, P. S. A phylogenomic analysis of Culicomorpha (Diptera) resolves the relationships among the eight constituent families. *Syst. Entomol.* **43**, 434–446 (2018).
9. McKenna, D. D. *et al.* The evolution and genomic basis of beetle diversity. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 24729–24737. <https://doi.org/10.1073/pnas.1909655116> (2019).
10. Song, H. *et al.* Phylogenomic analysis sheds light on the evolutionary pathways towards acoustic communication in Orthoptera. *Nat. Commun.* **11**, 4939. <https://doi.org/10.1038/s41467-020-18739-4> (2020).
11. Bybee, S. M. *et al.* Phylogeny and classification of Odonata using targeted genomics. *Mol. Phylogenet. Evol.* **160**, 107115. <https://doi.org/10.1016/j.ympev.2021.107115> (2021).
12. Lemmon, E. M. & Lemmon, A. R. High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **44**, 99–121 (2013).
13. Lemmon, A. R., Emme, S. A. & Lemmon, E. M. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* **61**, 727–744 (2012).
14. Young, A. D. & Gilling, J. P. Phylogenomics—Principles, opportunities and pitfalls of big-data phylogenetics. *Syst. Entomol.* **45**, 225–247 (2020).
15. Faircloth, B. C. *et al.* Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* **61**, 717–726. <https://doi.org/10.1093/sysbio/sys004> (2012).
16. Hanrahan, S. J. & Johnston, J. S. New genome size estimates of 134 species of arthropods. *Chromosome Res.* **19**, 809–823. <https://doi.org/10.1007/s10577-011-9231-6> (2011).
17. Baker, A. J. *et al.* Inverse dispersal patterns in a group of ant parasitoids (Hymenoptera: Eucharitidae: Oraseminae) and their ant hosts. *Syst. Entomol.* **45**, 1–19. <https://doi.org/10.1111/syen.12371> (2020).
18. Breinholt, J. W. *et al.* Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Syst. Biol.* **67**, 78–93 (2018).
19. Faircloth, B. C., Branstetter, M. G., White, N. D. & Brady, S. G. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol. Ecol. Resour.* **15**, 489–501. <https://doi.org/10.1111/1755-0998.12328> (2015).
20. Gillung, J. P. *et al.* Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids. *Mol. Phylogenet. Evol.* **128**, 233–245. <https://doi.org/10.1016/j.ympev.2018.08.007> (2018).
21. Haddad, S. *et al.* Anchored hybrid enrichment provides new insights into the phylogeny and evolution of longhorned beetles (Cerambycidae). *Syst. Entomol.* **43**, 68–89. <https://doi.org/10.1111/syen.12257> (2018).
22. Peters, R. S. *et al.* Evolutionary history of the Hymenoptera. *Curr. Biol.* **27**, 1013–1018. <https://doi.org/10.1016/j.cub.2017.01.027> (2017).
23. Shin, S. *et al.* Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. *Mol. Biol. Evol.* **35**, 823–836 (2018).
24. Young, A. D. *et al.* Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC Evol. Biol.* **16**, 143. <https://doi.org/10.1186/s12862-016-0714-0> (2016).
25. Cruaud, A. *et al.* The Chalcidoidea bush of life: Evolutionary history of a massive radiation of minute wasps. *Cladistics*. <https://doi.org/10.1111/cla.12561> (2023).
26. Cigliano, M. M., Braun, H., Eades, D. C. & Otte, D. Orthoptera species file. Version 5.0/5.0. [1/12/2024]. <http://Orthoptera.SpeciesFile.org>. (2019).
27. Song, H. In *Insect Biodiversity: Science and Society* 1st edn, Vol. 2 (eds Footitt, R. G. & Adler, P. H.) (Wiley, 2018).
28. Song, H. *et al.* 300 million years of diversification: Elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. *Cladistics* **31**, 621–651 (2015).
29. Hawlitschek, O. *et al.* New estimates of genome size in Orthoptera and their evolutionary implications. *PLoS One* **18**, e0275551. <https://doi.org/10.1371/journal.pone.0275551> (2023).
30. Yuan, H. *et al.* The evolutionary patterns of genome size in Ensifera (Insecta: Orthoptera). *Front. Genet.* **12**, 693541. <https://doi.org/10.3389/fgene.2021.693541> (2021).
31. Nakamura, T., Ylla, G. & Extavour, C. G. Genomics and genome editing techniques of crickets, an emerging model insect for biology and food science. *Curr. Opin. Insect Sci.* **50**, 100881. <https://doi.org/10.1016/j.cois.2022.100881> (2022).
32. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461> (2010).
33. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinform.* **10**, 421. <https://doi.org/10.1186/1471-2105-10-421> (2009).
34. Weissman, D. B. *et al.* Generic relationships of New World Jerusalem crickets (Orthoptera: Stenopelmatoidea: Stenopelmatinae), including all known species of *Stenopelmatus*. *Zootaxa* **4917**, 1–122 (2021).
35. Wang, X. *et al.* The locust genome provides insight into swarm formation and long-distance flight. *Nat. Commun.* **5**, 2957 (2014).
36. Petersen, M. *et al.* Orthograph: A versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinform.* **18**, 111. <https://doi.org/10.1186/s12859-017-1529-8> (2017).
37. Frandsen, P. B. *et al.* Phylogenomics recovers multiple origins of portable case making in caddisflies (Insecta: Trichoptera), nature's underwater architects. *Proc. R. Soc. B* **291**, 20240514. <https://doi.org/10.1098/rspb.2024.0514> (2024).
38. Zhou, W., Soghigian, J. & Xiang, Q.-Y. A new pipeline for removing paralogs in target enrichment data. *Syst. Biol.* **71**, 410–425. <https://doi.org/10.1093/sysbio/syab044> (2022).
39. Smith, B. T., Mauck, W. M., Benz, B. W. & Andersen, M. J. Uneven missing data skew phylogenomic relationships within the lories and lorikeets. *Genome Biol. Evol.* **12**, 1131–1147. <https://doi.org/10.1093/gbe/evaa113> (2020).
40. Huang, H. & Knowles, L. L. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Syst. Biol.* **65**, 357–365. <https://doi.org/10.1093/sysbio/syu046> (2016).
41. Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L. & Kimball, R. T. Avoiding missing data biases in phylogenomic inference: An empirical study in the landfowl (Aves: Galliformes). *Mol. Biol. Evol.* **33**, 1110–1125. <https://doi.org/10.1093/molbev/msv347> (2016).

42. Roure, B., Baurain, D. & Philippe, H. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol. Biol. Evol.* **30**, 197–214. <https://doi.org/10.1093/molbev/mss208> (2013).
43. Strauß, J., Stritih, N. & Lakes-Harlan, R. The subgenal organ complex in the cave cricket *Troglophilus neglectus* (Orthoptera: Rhaphidophoridae): Comparative innervation and sensory evolution. *R. Soc. Open Sci.* **1**, 140240 (2014).
44. Vandergast, A. G. *et al.* Tackling an intractable problem: Can greater taxon sampling help resolve relationships within the Stenopelmatoidea (Orthoptera: Ensifera)? *Zootaxa* **4291**, 1–33 (2017).
45. Field, L. H. *The Biology of Wetas, King Crickets and Their Allies* (CABI Publishing, 2001).
46. Song, H. & Mariño-Pérez, R. Re-evaluation of taxonomic utility of male phallic complex in higher-level classification of Acridomorpha (Orthoptera: Caelifera). *Insect Syst. Evol.* **44**, 241–260 (2013).
47. Strimmer, K. & von Haeseler, A. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6815–6819 (1997).
48. Dirsh, V. M. *The African Genera of Acridae* 579 (Anti-Locust Research Centre and Cambridge University Press, 1965).
49. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783–1785 (2010).
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170> (2014).
51. Wingett, S. W. & Andrews, S. FastQ screen: A tool for multi-genome mapping and quality control. *F1000 Res.* **7**, 1338 (2018).
52. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. <https://doi.org/10.1038/nbt.1883> (2011).
53. Kriventseva, E. V., Rahman, N., Espinosa, O. & Zdobnov, E. M. OrthoDB: The hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* **36**, D271–275. <https://doi.org/10.1093/nar/gkm845> (2008).
54. Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M. & Kriventseva, E. V. OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41**, D358–365. <https://doi.org/10.1093/nar/gks116> (2013).
55. Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J. & Kriventseva, E. V. OrthoDB: The hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.* **39**, D283–288. <https://doi.org/10.1093/nar/gkq930> (2011).
56. International Aphid Genomics C. Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol.* **8**, e1000313 (2010).
57. Werren, J. H. *et al.* Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* **327**, 343–348 (2010).
58. Kirkness, E. F. *et al.* Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12168–12173 (2010).
59. Mesquita, R. D. *et al.* Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14936–14941 (2015).
60. Terrapon, N. *et al.* Molecular traces of alternative social organization in a termite genome. *Nat. Commun.* **5**, 3636 (2014).
61. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549. <https://doi.org/10.1093/molbev/msy096> (2018).
62. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
63. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
64. Luo, R. *et al.* SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18. <https://doi.org/10.1186/2047-217X-1-18> (2012).
65. Peters, R. S. *et al.* Transcriptome sequence-based phylogeny of chalcidoid wasps (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and evolutionary success. *Mol. Phylogenet. Evol.* **120**, 286–296. <https://doi.org/10.1016/j.ympev.2017.12.005> (2018).
66. Meusemann, K. *et al.* A phylogenomic approach to resolve the arthropod tree of life. *Mol. Biol. Evol.* **27**, 2451–2464. <https://doi.org/10.1093/molbev/msq130> (2010).
67. Li, B., Lopes, J. S., Foster, P. G., Embley, T. M. & Cox, C. J. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. *Mol. Biol. Evol.* **31**, 1697–1709. <https://doi.org/10.1093/molbev/msu105> (2014).
68. Kuck, P. *et al.* Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front. Zool.* **7**, 10. <https://doi.org/10.1186/1742-9994-7-10> (2010).
69. Misof, B. & Misof, K. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst. Biol.* **58**, 21–34. <https://doi.org/10.1093/sysbio/syp006> (2009).
70. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612. <https://doi.org/10.1093/nar/gkl315> (2006).
71. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773. <https://doi.org/10.1093/molbev/msw260> (2017).
72. Tracer: MCMC Trace Analysis Tool Version v1.5.0 (2003–2009).
73. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153. <https://doi.org/10.1186/s12859-018-2129-y> (2018).
74. FigTree: Tree Figure Drawing Tool Version 1.3.1 (2006–2009).

Acknowledgements

We thank numerous collaborators who provided valuable specimens used in this study: the late Christiane Amédégnato, Corinna Bazelet, Sven Bradler, Maria Marta Cigliano, Antoine Foucart, David Gray, Claudia Hemp, Paul Lenhart, Kelly Miller, Joey Mugleston, Daniel Otte, Nik Tatarnic, Precious Tshililo, and Michael Whiting. We also thank several colleagues who provided logistic support and expertise during our field expeditions to Australia, Costa Rica, Dominican Republic, Mexico, Mozambique, South Africa, and the U.S.: Adrian Armstrong, Greg Cowper, Paolo Fontana, Eugenio Gonzalez, Brigido Hierro, Piotr Naskrecki, Kurt Nguyen, Ricardo Mariño-Pérez, Joey Mugleston, Oscar Salomon Sanabria-Urban, Ryan Selking, Nik Tatarnic, and Derek Woller. We are grateful to the 1KITE team for allowing us to use the transcriptome data for the probe design, especially Sabrina Simon, Karen Meusemann, and Bernie Misof. We thank Charlie Johnson at Texas A&M AgriLife Research Genomics and Bioinformatics Service for the NGS data generation and data processing. We thank the Texas A&M High-Performance Research Computing facility for enabling data analyses. We also thank Brandon Woo for generously providing macro photographs of various orthopterans used in Fig. 5. Fieldwork in the Dominican Republic was conducted under the Ministerio de Medio Ambiente y Recursos Naturales, authorization number 1424, and export permit number 692. Fieldwork in Western Australia was conducted under license number SF007010. Fieldwork in Mozambique was conducted under permit number PNG/DSC/C34/2016. Fieldwork in Costa Rica was conducted under Comision Nacional para la Gestion de la Biodiversidad

(CONAGEBIO) permit R-050-2018-OT-CONAGEBIO. Fieldwork in South Africa was conducted under ordinary permit number OP 4344/2018. This work was supported by the National Science Foundation (Grant Numbers DEB-1064082, IOS-1253493, DEB-1655202, and DEB-1937815 to H.S.) and the United States Department of Agriculture (Hatch Grant TEX0-2-6584 to H.S.). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author contributions

S.S. and H.S. conceived the study and designed experiments. B.F. generated new transcriptomic data. A.G.V. and D.B.W. provided critical samples and contributed to the design of taxon sampling. S.S., J.E., and H.S. designed the OR-TE probes. D.D.M. provided computational expertise and resources for data analysis. S.S., A.J.B., and H.S. analysed the data. S.S., A.J.B., and H.S. produced the original draft. All authors contributed to the writing and revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-72622-6>.

Correspondence and requests for materials should be addressed to H.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024