OXFORD (GIGA)n SCIENCE

GigaScience, 2022, **11**, 1–10
DOI: 10.1093/gigascience/giac006
TECH NOTE

Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes

Romain Feron [1,2] and Robert M. Waterhouse [1,2,*]

[1]Department of Ecology and Evolution, Le Biophore UNIL-Sorge, University of Lausanne, Lausanne 1015, Switzerland
[2]Evolutionary-Functional Genomics Group, L'Amphipole UNIL-Sorge, Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland
*__Correspondence address.__ Robert M. Waterhouse, Department of Ecology and Evolution, Le Biophore UNIL-Sorge, University of Lausanne, Lausanne 1015, Switzerland. E-mail: robert.waterhouse@unil.ch

Abstract

Background: Ambitious initiatives to coordinate genome sequencing of Earth's biodiversity mean that the accumulation of genomic data is growing rapidly. In addition to cataloguing biodiversity, these data provide the basis for understanding biological function and evolution. Accurate and complete genome assemblies offer a comprehensive and reliable foundation upon which to advance our understanding of organismal biology at genetic, species, and ecosystem levels. However, ever-changing sequencing technologies and analysis methods mean that available data are often heterogeneous in quality. To guide forthcoming genome generation efforts and promote efficient prioritization of resources, it is thus essential to define and monitor taxonomic coverage and quality of the data.

Findings: Here we present an automated analysis workflow that surveys genome assemblies from the United States NCBI, assesses their completeness using the relevant BUSCO datasets, and collates the results into an interactively browsable resource. We apply our workflow to produce a community resource of available assemblies from the phylum Arthropoda, the Arthropoda Assembly Assessment Catalogue. Using this resource, we survey current taxonomic coverage and assembly quality at the NCBI, examine how key assembly metrics relate to gene content completeness, and compare results from using different BUSCO lineage datasets.

Conclusions: These results demonstrate how the workflow can be used to build a community resource that enables large-scale assessments to survey species coverage and data quality of available genome assemblies, and to guide prioritizations for ongoing and future sampling, sequencing, and genome generation initiatives.

Keywords: arthropod genomes, biodiversity genomics, BUSCO assessments, genome assembly, genome quality database, reproducible workflow

Introduction

Advances in sequencing technologies are bringing down costs and reducing sample requirements, leading to an accelerating accumulation of new and improved genome assemblies. Ambitious initiatives to coordinate sequencing of all known species are generating representative genomes from across the tree of life that catalogue Earth's genetic biodiversity. In addition to constituting an inventory of biological diversity, the assembled and annotated genomes drive research to understand function and evolution at multiple levels, as well as to benefit human welfare [1, 2]. Investigating such questions using genomic data often requires comprehensive multi-species comparative analyses that benefit from high-quality assemblies [3, 4]. It is therefore essential to be able to define the current taxonomic coverage of high-quality assemblies to guide forthcoming sequencing efforts and promote efficient prioritization of resources globally.

Methods to gauge assembly quality include 2 main families of metrics [5]. One summarizes contiguity using metrics like N50 length, where half the assembly comprises sequences of length N50 or longer, or L50 count, the smallest number of sequences whose lengths sum to 50% of the assembly. Complementary approaches estimate completeness by examining gene or protein content, e.g., the DOmain-based General Measure for transcriptome and proteome quality Assessment (DOGMA) [6, 7] or BUSCO [8, 9]. BUSCO has emerged as a standard and is used by UniProt [10] and the US NCBI [11], as well as by genomics data quality assessment pipelines like MultiQC [12] and BlobToolKit [13]. BUSCO is based on the evolutionary expectation that single-copy orthologues found in nearly all species from a given taxon should be present and single-copy in any newly sequenced species from the same clade. BUSCO datasets are built for multiple taxonomic lineages by identifying near-universal groups of single-copy orthologues from OrthoDB [14, 15]. For assembly evaluations, sequence searches followed by gene predictions and orthology classifications identify complete, duplicated, or fragmented BUSCOs. The proportions recovered indicate the completeness in terms of expected subsets of evolutionarily conserved genes. Extrapolating from these, a high BUSCO completeness score suggests that the sequencing and assembly procedure has successfully reconstructed a reliable representation of the full set of genes.

Using their Complete Proteome Detector algorithm, UniProt classifies proteomes as "standard," "close to standard," or "outlier" and provides BUSCO proteome completeness summaries. For assemblies, the NCBI Assembly database provides summary statistics and metadata for each record. Querying these can provide snapshots of taxonomic coverage and data quality, but researchers currently lack access to comprehensive and standardized assessments of available assemblies. These would allow data producers to compare their assemblies to existing data at the most relevant taxonomic level. They would also provide re-

Received: October 27, 2021. **Revised:** December 12, 2021. **Accepted:** January 13, 2022

searchers with comprehensive overviews of resources for their focal taxa. Such communities would benefit from being able to survey coverage and quality of available genomic resources for selected groups of species from their field of interest. This would (i) aid project design, particularly in the context of comparative genomics analyses; (ii) simplify comparisons of the quality of their own data with that of existing assemblies; and (iii) provide a means to keep up to date with accumulating genomics resources relevant to their ongoing research projects.

To address these needs, we developed an automated analysis workflow that performs BUSCO assessments of assemblies for user-selected taxa from the NCBI, concurrently collating assembly metadata to build a catalogue of metrics in a taxonomically aware framework. To demonstrate the utility of standardized evaluations for a clade, we applied our workflow to the phylum Arthropoda, for which genome data are supporting research on a wide range of topics including their roles as pests and disease vectors [16]. Since sequencing of the fruit fly genome [17], sampling of arthropods has included ants and other Hymenoptera [18, 19], arachnids [20], beetles [21], butterflies and other Lepidoptera [22], flies and other Diptera [23, 24], hemipterans [25], and many others [26, 27]. Through efforts such as the i5k 5000 arthropod genomes initiative [28] and others, the arthropod genomics community has worked to overcome challenges in genome sequencing, assembly, and annotation [29–31]. Despite encompassing only a tiny fraction of all arthropod diversity and showing taxonomic biases in sampling, assemblies are accumulating rapidly and are now publicly available for hundreds of species [32, 33].

Our large-scale assessments allowed us to (i) survey the current taxonomic coverage and assembly quality across Arthropoda, (ii) examine how key assembly metrics relate to gene content completeness, (iii) quantify effects on assessment resolution using different BUSCO lineage datasets, (iv) compare the results of BUSCO v3 with the newer BUSCO v4, and (v) demonstrate how our workflow can be used to build a community resource. We provide the catalogue as an open resource for the arthropod genomics community, and the stand-alone open-source workflow for users to build their own catalogues tailored to the needs of their research communities. Enabling user-customizable, taxonomically aware, standardized, and updatable quality assessments of available genome assemblies will empower genomics data producers and users, as well as helping to prioritize species for genomic sequencing of Earth's biodiversity.

Results and Discussion

An automated workflow for assembly assessments

We developed an automated analysis workflow to build and maintain NCBI genome assembly assessment catalogues for selected taxa. This workflow performs the following steps: (i) query the NCBI GenBank Assembly database [11] to retrieve information about available assemblies and corresponding metadata for a user-defined taxonomic group; (ii) identify all relevant BUSCO lineages based on species taxonomy for each assembly; (iii) run BUSCO on each assembly using each relevant lineage dataset; (iv) generate a summary table that collates all BUSCO results with assembly metrics and metadata; and (v) generate an HTML/JavaScript interactive table containing all data from the summary (Supplementary Fig. S1). Assembly metadata are integrated into a summary file along with 5 metrics obtained from the results of running BUSCO on each assembly with each relevant lineage: the percentages of complete, complete single-copy, complete duplicated, fragmented, and missing BUSCOs. The workflow allows users to systematically assess all assemblies available at the NCBI for a given taxon of interest. Importantly, it is also designed to perform on-demand updates to assess assemblies added to NCBI GenBank since the last run. The final output provides all the information retrieved for each assembly in both JSON and tab-separated formats, and an HTML/JavaScript table is generated to display the data. This output is saved in a summary folder each time the workflow is run. The workflow is implemented using the Snakemake workflow management engine [34, 35], and all software dependencies are managed by the Conda package manager. It is fully automated and can be configured using a YAML file to specify the query to use for the NCBI Assembly database, BUSCO parameters, and the information to display in the output tables. The code and documentation are available from [36].

A survey of arthropod genome assembly resources

Applying the assembly assessment workflow to the phylum Arthropoda on 11 June 2021 resulted in the retrieval of a total of 2,083 assemblies from 1,387 species, providing a snapshot of the taxonomic coverage of available genome resources for arthropods at the NCBI. Of the ∼120 arthropod orders recognized by the NCBI Taxonomy database [37] or the Catalogue of Life [38], 48 are represented by ≥1 genome assembly, with 21 orders represented by ≥5 assemblies (Fig. 1). Currently available genome resources include 1,929 assemblies for 1,262 insect species and a further 154 assemblies for 125 other arthropod species. For Insecta, this is a doubling of the number of species since a November 2020 survey from Hotaling et al. [33]. Species with assemblies represent a ∼0.06% sampling from a total of ∼1 million described arthropod species (792,339 species records and 121 orders in the NCBI Taxonomy database on 10 August 2021; 1,126,288 extant species and 123 orders in the Catalogue of Life 2021–06-10 edition).

This survey highlights the sparsity and taxonomic imbalance of current species sampling, with 79.5% of species (83% of assemblies) belonging to only 3 orders: Lepidoptera—e.g., butterflies, moths (712 species, 1,122 assemblies), Diptera—e.g., flies (216 species, 389 assemblies), and Hymenoptera—e.g., ants, bees, wasps (175 species, 217 assemblies). Similar sampling biases were identified by the November 2020 survey of NCBI resources for Insecta [33], where order-level counts for 601 insect species from 20 orders were 28% Diptera, 20% Lepidoptera, and 27% Hymenoptera. Notably, while roughly one-third of insect orders are represented, only 5–10% of orders from other groups such as crustaceans, myriapods (e.g., centipedes, millipedes), and chelicerates (e.g., spiders, scorpions) have ≥1 assembly. Across Arthropoda, orders with the most sequenced species also show the highest proportions of sequenced versus Catalogue-of-Life–described species despite also being amongst the most species-rich clades: 0.063% sequenced species for Lepidoptera, 0.019% for Diptera, and 0.016% for Hymenoptera. An exception to this observation is Coleoptera—e.g., beetles, weevils, which has the highest number of described species to date with currently available genome assembly resources for only 0.007% of these species.

These uneven distributions likely reflect historical biases in research interests for dipterans, which include the model species *Drosophila melanogaster* and disease vectors like mosquitoes; for lepidopterans, which have been a model to study the genetic basis of complex traits and population genetics; and for hymenopterans, which include many well-studied social insects. While such

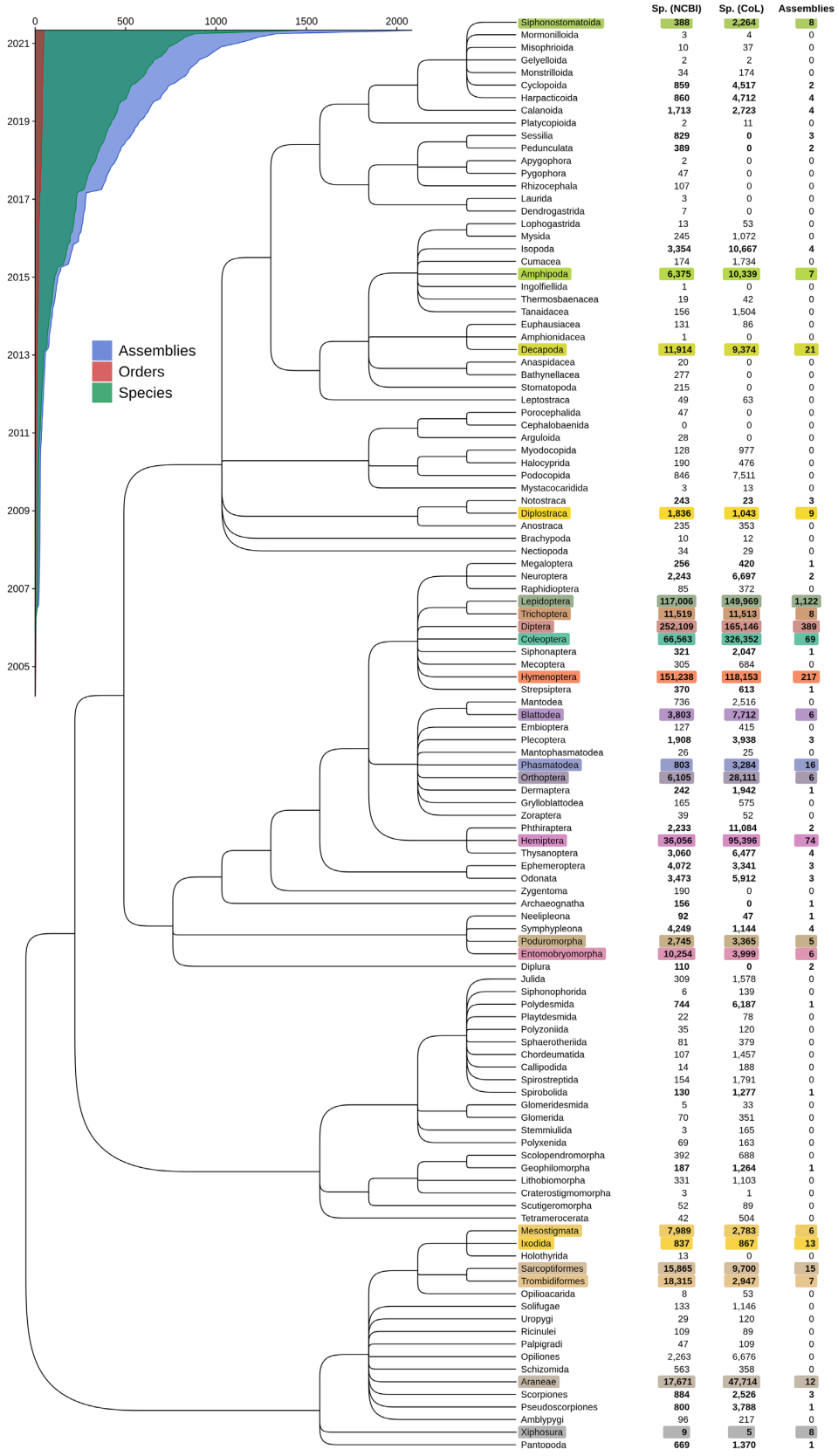

Figure 1: Available genome assembly resources across the arthropod phylogeny. The Arthropoda phylogeny from the US NCBI Taxonomy database shows the evolutionary relationships amongst 114 orders. Counts of described species (Sp.) within each order are shown from the NCBI (v.2021–06-11) and the Catalogue of Life (CoL, v.2021–06-10), alongside numbers of genome assemblies available from the NCBI Assembly database (accessed on 25 August 2021). Of the 114 orders recognized by both the NCBI and the CoL, 48 orders are represented by ≥1 genome assembly. The 21 orders with ≥5 assemblies are highlighted with distinct colours, which are maintained for cross-referencing in Figs 2–4. The inset shows the accumulation of assemblies, species, and orders submitted to the NCBI since 2005 (note that in the case of assembly updates only the latest submission dates are considered).

biases may persist owing to factors such as research priorities and ease of sampling, the balance should improve as the numbers and taxonomic spread of available arthropod genome assemblies continue to grow rapidly (Fig. 1, Inset). Surveying taxonomic representation in this way highlights the increasingly rapid accumulation of new genome assemblies at the NCBI, providing researchers with a comprehensive overview of the species coverage of available genomics resources for their taxa of interest.

Assessing the surveyed species data allows for phylum-wide comparisons of the contiguity and completeness of genome assemblies available at the NCBI. Focusing on the 21 orders with ≥5 assemblies, order representation is notably unbalanced and assembly quality metrics summarized with N50 lengths and BUSCO completeness scores vary greatly among and within orders (Fig. 2). Large differences between assembly and species counts are primarily driven in Lepidoptera by *Heliconius melpomene* (n = 42), *Junonia neildi* (n = 35), *Junonia evarete* (n = 32), and 6 other *Junonia* and *Heliconius* species with >10 assemblies, and in Diptera mainly by *D. melanogaster* (n = 26), *Drosophila simulans* (n = 12), and *Anopheles coluzzii* (n = 10). The 307 species with >1 assembly comprise distinct assembly submissions and not updates that result in new versions of existing submissions (in this case only the latest version is surveyed). Roughly half (142) of these species with multiple assemblies are represented by a chromosome-level assembly. Across all assemblies, those labelled as chromosome-level account for 12.3%, while a further 41.1% are labelled as scaffold-level assemblies, and the remaining 46.6% are contig-level (Supplementary Fig. S2).

Excluding Lepidoptera, which are skewed by a large number of poor-quality assemblies [39], median N50 lengths per order represented by ≥5 assemblies (shown in Fig. 2C) range from 11.6 kb for Sarcoptiformes (mites, 15 assemblies for 12 species) to 96.3 Mb for Xiphosura (horseshoe crabs, 8 assemblies for 4 species). The horseshoe crabs have large genomes of 1.7–2.2 Gb, for which concerted efforts have been successful in producing contiguous assemblies [40–43]. The mite genomes are all much smaller, with a median assembly span (total length) of just 88.5 Mb, where the latest assembly for the parasitic mite, *Sarcoptes scabiei*, provides an example of how long-read technologies are helping to improve available genomic resources [44].

Median BUSCO completeness scores per order represented by ≥5 assemblies for the Arthropoda lineage dataset (Fig. 2D) are less variable than the N50 lengths and, excluding Lepidoptera, range from 72.1% for Sarcoptiformes to >97% for Diplostraca (clam shrimps and waterfleas, 9 assemblies for 7 species), Blattodea (cockroaches and termites, 6 assemblies for 5 species), Diptera, and Hymenoptera. Although within-order distributions can be highly variable, all but 2 of the 21 orders (Sarcoptiformes and Trombidiformes mites) are represented by ≥1 assembly with >90% complete BUSCOs. These contiguity and completeness distributions include all available assemblies, i.e., not filtered by level (contig, scaffold, chromosome) or type (e.g., haploid, principal or alternate pseudohaplotype). The completeness of contig-level assemblies is expectedly lower than that of scaffold- or chromosome-level (Supplementary Fig. S2B) assemblies, and although alternate pseudohaplotype assemblies can achieve high BUSCO completeness scores, they are generally lower than for principal pseudohaplotypes (Supplementary Fig. S2C). Additional partitioning of the datasets by sequencing technologies, assembly algorithms, and so forth is feasible where the metadata labels are applied consistently, or after metadata curation as for previous assessments of insects that contrasted short- and long-read technologies [33]. These phylum-wide comparisons of the qualities of available genome assemblies highlight the unbalanced order-level species representation, as well as the variable levels of contiguity and completeness within and amongst arthropod orders.

Arthropod assembly contiguity, size, and completeness

With 2,083 assemblies exhibiting variable contiguities and sizes, the survey results provide the opportunity to examine expectations of how assembly contiguity and size relate to gene content completeness. Although long-read sequencing technologies are producing improved results [33], large genomes have often been challenging to assemble owing to expanded proportions of repetitive sequences [31]. Even for smaller genomes, repeats can hinder scaffolding of contigs, reducing contiguity and possibly adding undetermined gap regions to the assembly. Less contiguous assemblies are thus expected to have more genes split across scaffolds, or partially or completely missing, resulting in lower completeness scores [45].

The Earth BioGenome Project [2] criteria for a reference-quality assembly include obtaining a complete and single-copy BUSCO score >90% and having the majority of sequences assigned to chromosomes. While 828 of the assessed arthropod assemblies achieve a complete and single-copy BUSCO score >90%, only 229 of these are also labelled as chromosome-level assemblies. Indeed, comparing assembly N50 values with their completeness scores shows that obtaining >90% complete BUSCOs can be achieved across a wide range of contiguities (Fig. 3A). Recovery of >90% complete BUSCOs is observed for assemblies with N50s as low as 3.5 kb (*Tetragonula mellipes*, stingless bee, 92.1% complete) and 3.9 kb (*Chrysomya rufifacies*, blowfly, 97.4% complete). While some with N50s <10 kb are able to achieve >90% (n = 25) or 80–90% (n = 21) completeness, the vast majority of assemblies with such low contiguity levels achieve considerably lower BUSCO completeness scores than contiguous assemblies (i.e., N50 >10 kb). Among the latter, notable anomalies include 24 assemblies with N50s >10 kb that nonetheless all have completeness scores of <50%. One-third of these are labelled as alternate pseudohaplotypes, which offers an explanation for the low completeness levels because they likely represent collections of purged haplotigs. Others include improbably small assembly spans, e.g., *Sertania guttata* (butterfly, 30 Mb span of 628 Mb estimate) and *Dactylopius coccus* (scale insect, 18 Mb span of 386 Mb estimate), or high proportions of undetermined sequence, e.g., the brown recluse spider, *Loxosceles reclusa* (45% gaps). Biological complexity may also offer explanations, such as in the case of the Lord Howe Island stick insect, *Dryococelus australis* (N50 = 17.3 kb, 43.5% complete), a potentially hexaploid genome with an estimated size of 4.2 Gb that achieved an assembly span of 3.4 Gb [46].

The largest assemblies span >5 Gb, with the maximum reported for the Asian longhorned tick, *Haemaphysalis longicornis*, at 7.3 Gb, which shows 92% complete BUSCOs (Fig. 3B). The estimated genome size for this tick however is only 3.4 Gb, and a duplicated BUSCO score of 74.4% suggests that the applied assembly methods failed to collapse the alternative haplotypes. Indeed, an alternative assembly for this tick spans just 2.6 Gb and scores 89.5% complete and 2.1% duplicated BUSCOs. A handful of other large assemblies with high duplicated scores are annotated as being non-collapsed, but others with many duplicated BUSCOs are also likely diploid or partially diploid (Supplementary Fig. S3). The smallest reported genome size for an arthropod to date is that of the tomato russet mite, *Aculops lycopersici* (Trombidiformes), exceptionally streamlined at only 32.5 Mb [47].

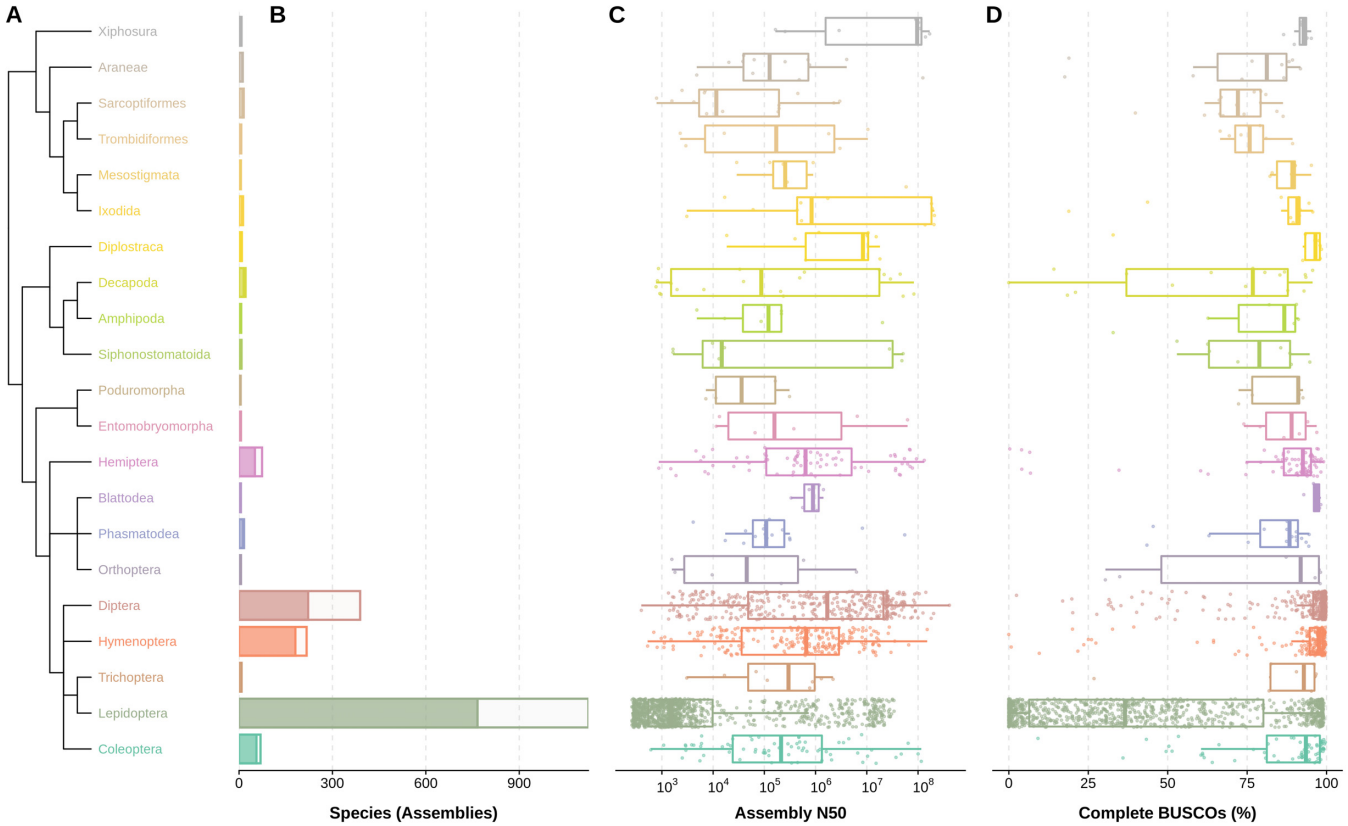

Figure 2: Order-level representation, contiguity, and completeness of 2,024 available assemblies for 1,326 arthropod species from the 21 orders with ≥5 assemblies. Data are presented only for orders with ≥5 assemblies available at the NCBI (2021–06-11). (A) Phylogenetic relationships of the 21 orders as resolved by the NCBI Taxonomy database. (B) Number of assemblies (entire bars) and unique species (dark fractions) retrieved from the NCBI Assembly database for each order. (C) Distribution of assembly NCBI scaffold N50 values (base pairs, log scale) for each order. (D) Distribution of BUSCO completeness (% of 1,013 BUSCOs) for the arthropod lineage dataset (arthropoda_odb10) for each order. Box plots show the median, first and third quartiles, and lower and upper extremes of the distribution (1.5 × IQR), and all values are overlaid as points to show the full distribution.

It achieves a Eukaryota completeness score of 83%, but only 67% Athropoda complete, which could reflect the evolutionary streamlining process but may also be related to challenges during gene prediction in such a gene-dense genome where genes have also experienced large-scale intron losses. The smallest assembly with a >80% Arthropoda completeness score is that of a grasshopper, *Xenocatantops brachycerus* (42 Mb, 92% complete); however, inspecting the metadata reveals this to be a transcriptome rather than a genome assembly [48]. Amongst the smallest true genome assemblies achieving >80% completeness are other Trombidiformes as well as Sarcoptiformes, e.g., the house dust mite *Dermatophagoides farinae* (54 Mb, 84% complete). Although there are fewer large assemblies spanning >1 Gb, across the full range of their sizes most achieve good completeness scores of >90%, indicating that sequencing technologies and assembly methods are able to overcome challenges often associated with large genomes.

Comparing assembly N50s and sizes with BUSCO duplicated scores (Supplementary Fig. S3) identifies several assemblies with high duplication levels. Some of these are labelled as "unresolved-diploid" assemblies, which explains these high duplication levels, but this mechanism to inform users about the non–strictly haploid status of certain assemblies is not widely nor consistently applied. Fragmented BUSCO scores (Supplementary Fig. S4) are expectedly higher for most of the less contiguous assemblies, highlighting those where many genes are likely split across 2 or more scaffolds. The survey results therefore provide the community with a comprehensive overview of genomic dataset qualities and of how contiguity and size relate to gene content completeness across currently available arthropod genome assemblies.

BUSCO dataset lineage and version comparisons

The reference BUSCO lineage datasets are defined at different taxonomic levels that capture sets of near-universal single-copy orthologues from OrthoDB [49] at ancient, intermediate, and younger nodes of the tree of life [8,9]. As duplication and loss events over evolutionary time erode the numbers of identifiable BUSCOs, datasets defined for more ancient lineages are smaller than for the younger ones, e.g., n = 255 for Eukaryota and n = 954 for Metazoa, versus n = 3,285 for Diptera and n = 13,780 for Primates (OrthoDB v10 datasets). An advantage of the smaller older lineage datasets is that compute runtimes are shorter because there are fewer individual genes to search for. The larger younger lineage datasets on the other hand offer greater resolution, meaning that scores are less affected by small differences in counts of complete, fragmented, or missing BUSCOs.

Our results provide the opportunity to compare the scores obtained using different lineage datasets for a large number of arthropod assemblies (Fig. 4). Comparing percentages of complete BUSCOs identified with the Eukaryota (n = 255) and the Arthropoda (n = 1,013) lineage datasets for a total of 1,977 arthropod assemblies shows highly linearly correlated scores, especially for the highest-scoring assemblies (Fig. 4A). For those scoring <80% there is a small but noticeable shift towards Arthropoda producing slightly higher scores than Eukaryota, indicating that proportion-

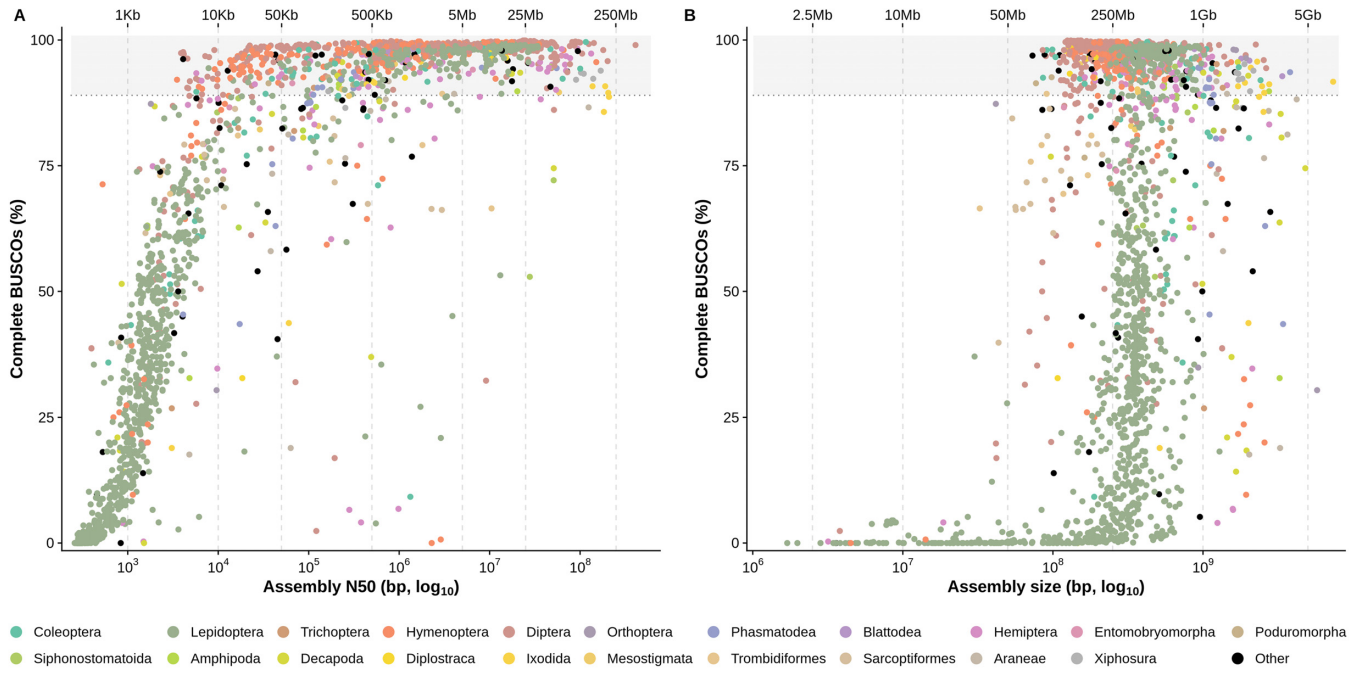

Figure 3: BUSCO completeness compared with assembly contiguity and size. Complete BUSCOs (in % of total BUSCOs for the arthropoda_odb10 dataset) are plotted against assembly N50 in bp (**A**) and assembly size in bp (**B**) for each assessed assembly. Both assembly N50 and assembly size are represented with a log scale. The colour of a point indicates the order of the sequenced species. Dotted lines indicate N50 values of 1, 10, 50, and 500 kb and 5 Mb in **A** and assembly size values of 50 Mb, 250 Mb, 1 Gb, and 5 Gb in **B**. BUSCO completeness scores >90% are highlighted with a grey background.

ately more of the larger set of Arthropoda BUSCOs can be recovered from lower-quality assemblies. Outlier points above the identity (y = x) axis suggest that the lower-resolution Eukaryota lineage dataset occasionally produces overestimates of completeness, where proportionately more of the smaller set of ancient Eukaryota BUSCOs are recovered. Similar trends are observed when comparing the Arthropoda results to the higher-resolution Insecta (n = 1,367) lineage dataset, with highly linearly correlated scores and occasional small overestimates of completeness using the Arthropoda lineage dataset (Supplementary Fig. S5A).

Comparing Arthropoda results to those from 4 insect order-level lineage datasets shows high agreements for the highest-scoring assemblies (Fig. 4B). For lower-scoring assemblies, results from applying the Lepidoptera and Hemiptera lineage datasets tend towards slightly higher scores than for Arthropoda. In contrast, using the Hymenoptera and Diptera lineage datasets generally produces lower completeness scores than for Arthropoda. These shifts could arise from the uneven representations of these orders in the 90-species Arthropoda lineage dataset, which is dominated by 20 hymenopterans and 15 dipterans, with only 9 species each for Lepidoptera and Hemiptera. The same trends are observed when comparing results from the order-level lineage datasets to those from the Insecta dataset (Supplementary Fig. S5B).

In addition to updates to the codebase, BUSCO v4 was released with updated lineage datasets based on orthology data from OrthoDB v10 [49], while BUSCO v3 used data from OrthoDB v9 [50]. Comparing completeness scores using the 2 Arthropoda datasets shows high levels of agreement for the highest-scoring assemblies with a consistent shift towards lower scores reported by BUSCO v4 for lower-quality assemblies (Fig. 4C). A similar pattern is observed when comparing results from the 2 Insecta datasets (Supplementary Fig. S5C). The Diptera comparisons on the other hand reveal some score variations, which nevertheless agree well over the full range of assembly qualities (Fig. 4D), similarly to results from the Hymenoptera datasets (Supplementary Fig. S5D). The different versions therefore produce generally consistent and comparable estimates of completeness, with a tendency for the OrthoDB-v10–based Arthropoda and Insecta datasets to report lower scores, especially for lower-quality assemblies. For objective quantitative comparisons it is thus necessary to assess assemblies using the same BUSCO versions, parameters, and lineage datasets, as presented here for the phylum-wide assessments of available arthropod genome assemblies.

The Arthropoda assembly assessment catalogue: A^{3}Cat

Running the workflow on the selected taxon of Arthropoda (NCBI:txid6656) produced the first version of the Arthropoda Assembly Assessment Catalogue (A^{3}Cat v.2021–06-11), demonstrating how the workflow can be used to build a community resource. The A^{3}Cat is provided as a searchable online table [51] (Arthropoda Assembly Assessment Catalog, RRID:SCR_021864) that makes it possible to browse and download the collated metadata and BUSCO assessment results for arthropod assemblies available from the NCBI (n = 2,083 for A^{3}Cat v.2021–06-11). Through simple text searches and/or applying query filters, users are able to quickly obtain downloadable overviews of the availability and quality of genome assembly resources for their arthropod taxa of interest. Without the computational burden of having to evaluate publicly available resources themselves, users can directly compare the assessments of their own assemblies with the precomputed results available from the A^{3}Cat. In addition, for version and parameter controlled like-for-like comparisons, a user-workflow is provided to compute quality metrics on user-provided

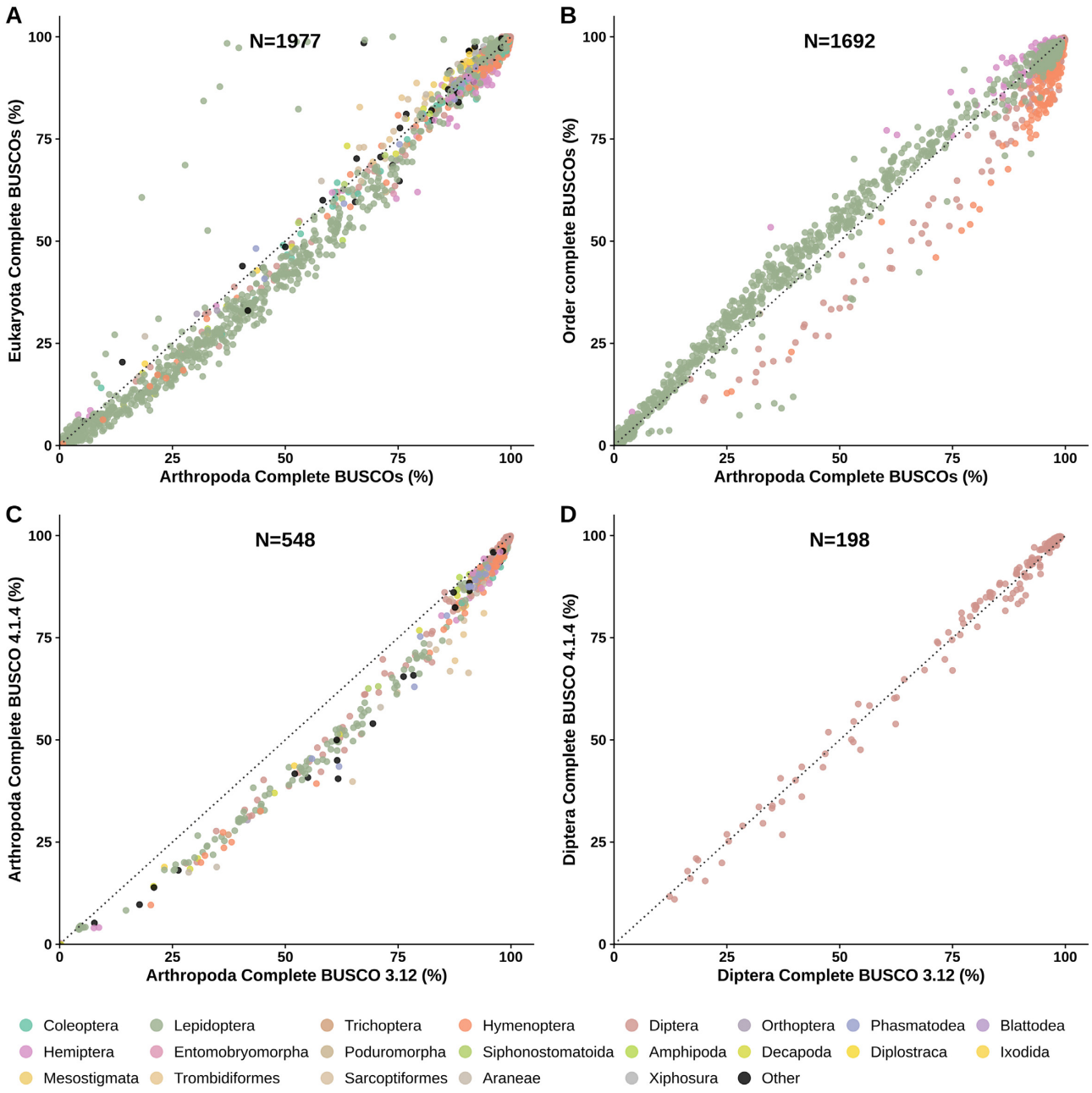

Figure 4: Comparisons of BUSCO lineage datasets and BUSCO versions. Congruence of BUSCO completeness scores is assessed by comparing results from (**A**) the Eukaryota (n = 255) and the Arthropoda (n = 1,013) lineage datasets, (**B**) the Arthropoda and 4 insect order-level lineage datasets (Hemiptera [n = 2,510], Hymenoptera [n = 5,991], Lepidoptera [n = 5,286], Diptera [n = 3,285]), and lineage datasets from BUSCO v4 (OrthoDBv10) and BUSCO v3 (OrthoDBv9) for (**C**) Arthropoda (odb9: n = 1,066) and (**D**) Diptera (odb9: n = 2,799). In each panel, the dotted lines show the identity (y = x).

assemblies and compare them with A^{3}Cat results for species from the same taxonomic clade (code and documentation are available from [52]).

Conclusions

Results from applying the assessment workflow to the phylum Arthropoda demonstrate the utility of building resources that provide a standardized overview of the current taxonomic coverage and quality of genome assembly resources available from the NCBI. The large-scale dataset also offers the opportunity to examine how widely used assembly metrics relate to BUSCO genespace completeness across a heterogeneous collection of genomes. Some anomalies point to errors or inconsistent use of metadata annotations where retractions or revisions would help to avoid misleading users about these resources. Furthermore, comparing results using different BUSCO datasets on large collections of assemblies reveals trends associated with using ancient (lower-resolution) or younger (higher-resolution) lineages, and datasets built for BUSCO v3 or v4. While congruence is high especially for high-scoring assemblies, truly objective comparisons require reporting of the BUSCO versions, parameters, and lineage datasets used. Our data will enable future large-scale comparisons with results from the recently released BUSCO v5, which includes a new

genome assessment strategy that improves efficiency and runtimes [53]. Future workflow developments would aim to capture new metadata attributes made available from the NCBI such as summary information on repeat content, or computed locally, e.g., nucleotide compositions from *k*-mer analyses. The automated analysis workflow to build and maintain NCBI genome assembly assessment catalogues for selected taxa allows users to build updatable community resources, here exemplified with the A^3Cat, which facilitates surveying of species coverage and data quality for available arthropod assemblies and serves to guide ongoing and future genome generation initiatives.

Materials and Methods

Assembly selection and assessment workflow implementation

Accession numbers for all assemblies in the user-specified taxon are retrieved by querying the NCBI datasets API [54] with the ncbi-datasets-pylib library (version 12.3.0 in version 1.0 of a3cat-workflow) (Step 1 in Supplementary Fig. S1). For each assembly, the data package is downloaded to a temporary zip file using the "datasets" command-line utility (version 11.22.0 in version 1.0 of the a3cat-workflow). The nucleotide sequence and metadata are extracted from each data package with the ncbi-datasets-pylib library and stored as fasta and JSON files, respectively (Step 2 in Supplementary Fig. S1). For each assembly, complete taxonomic information is retrieved from the NCBI Taxonomy database [37] using the ete3 python module [55], version 3.1.2 in version 1.0 of the a3cat-workflow) and stored in a JSON file (Step 3 in Supplementary Fig. S1). Taxonomic information is used to determine all BUSCO lineage datasets relevant for each assembly (Step 4 in Supplementary Fig. S1). During this step, assemblies are filtered by size, scaffold N50, and a manual filter list to discard assemblies that are too short and/or fragmented to contain any BUSCOs; this is necessary because BUSCO returns an error if no BUSCOs are found. The completeness of each assembly is assessed using BUSCO in genome mode and all other settings to default (version 4.1.4 in version 1.0 of the a3cat-workflow) for each applicable lineage dataset (Step 5 in Supplementary Fig. S1). The results folder generated by BUSCO is saved as a compressed archive with the exception of the BLAST database (blast_db) and BLAST input sequences (<run_name>/blast_output/sequences). The full results table, missing BUSCO list, and short summary are also retained in the final output for convenience. Metadata retrieved from NCBI and BUSCO scores for all assemblies are aggregated into a JSON file that summarizes all the raw information retrieved and computed by the workflow (Step 6 in Supplementary Fig. S1). This JSON file is converted into a table with formatted headers stored in a tab-separated file where columns represent metadata and BUSCO scores and each line corresponds to an assembly (Step 7 in Supplementary Fig. S1). Finally, an interactive table is generated as an HTML page using the Data Tables JavaScript library [56] (version 1.10.24 in version 1.0 of the a3cat-workflow) (Step 8 in Supplementary Fig. S1). The entire workflow is implemented using the Snakemake workflow management engine [34, 35] and all software dependencies are managed by the Conda package manager; this implementation ensures that the workflow is portable and entirely reproducible. Parameters for each step of the workflow are specified in a YAML file and additional configuration files can be used to customize the table and HTML output. The code and documentation for the workflow are available from [36].

Assessment workflow deployment and data analyses

Results presented in this study were obtained by running version 1.0 of the a3cat-workflow on 11 June 2021. Species estimates were retrieved from the NCBI Taxonomy database using ete3 (version 3.1.2) on 21 August 2021 and from the Catalogue of Life version 2021–06-10. Phylogenetic trees were automatically generated from NCBI taxonomy data with ete3. BUSCO scores for version 4.1.4 were obtained directly from the output of a3cat-workflow, while scores for version 3.12 were obtained with a development release version of the workflow [57]. Figures were generated with ggplot2 version 3.3.5 [58] and ggtree version 3.0.1 [59] in R version 4.1.0 [60]. All data-related figures, numbers, and supplementary material were generated with a Snakemake workflow [35] available from [61] using Snakemake version 6.3.0.

Availability of Supporting Source Code and Requirements

Project name: The Arthropoda Assembly Assessment Catalogue Workflow
Project home page: https://gitlab.com/evogenlab/a3cat-workflow
Operating system: Platform independent
Programming language: Snakemake, Python
Other requirements: Snakemake, Conda
License: GPLv3
RRID:SCR_021864
biotools ID: arthropoda_assembly_assessment_catalogue

Data Availability

The data underlying this article are available in the NCBI Assembly Database at https://www.ncbi.nlm.nih.gov/assembly. An archival copy of the code and supporting data is also available via the *GigaScience* database GigaDB [62].

Additional Files

Supplementary Figure S1. Overview of the automated workflow for assembly assessments. The NCBI GenBank database is queried using the NCBI "datasets" python library **(1)** and assembly packages are downloaded with the "datasets" utility to obtain the genome sequence in a fasta file and metadata in a JSON file **(2)**. The complete taxonomy is retrieved from the NCBI taxonomy database for each assembly using ete3 **(3)** and used to determine relevant BUSCO lineage datasets **(4)**. BUSCO is then run with each lineage dataset on each assembly **(5)**, and BUSCO results are aggregated with taxonomy information and metadata into a single complete JSON summary file **(6)**. Finally, the summary is converted to a tab-separated table **(7)** and an HTML/Javascript searchable table is generated **(8)**.

Supplementary Figure S2. Accumulation over time and BUSCO completeness of contig-level, scaffold-level, or chromosome-level assemblies. (**A**) The cumulative numbers of assemblies labelled as contig-level, scaffold-level, and chromosome-level according to their submission dates at the NCBI Assembly database. (**B**) Distributions of BUSCO completeness scores for assemblies labelled as contig-level, scaffold-level, and chromosome-level at the NCBI Assembly database, and (**C**) those labelled as simply haploid, or distinguishing between the principal and alternate haplotypes. Box plots show the median, first and third quartiles, and lower and

upper extremes of the distribution (1.5 × IQR), and all values are overlaid as points to show the full distribution.

Supplementary Figure S3. Proportion of duplicated BUSCOs compared with assembly contiguity and size. Duplicated BUSCOs (in % of total BUSCOs for the arthropoda_odb10 dataset) are plotted against assembly N50 in bp (**A**) and assembly size in bp (**B**) for each assessed assembly. Both assembly N50 and assembly size are represented with a log scale. The colour of a point indicates the order of the sequenced species. Dotted lines indicate N50 values of 1, 10, 50, and 500 kb and 5 Mb in panel **A** and assembly size values of 50 Mb, 250 Mb, 1 Gb, and 5 Gb in panel **B**.

Supplementary Figure S4. Proportion of fragmented BUSCOs compared with assembly contiguity and size. Fragmented BUSCOs (in % of total BUSCOs for the arthropoda_odb10 dataset) are plotted against assembly N50 in bp (**A**) and assembly size in bp (**B**) for each assessed assembly. Both assembly N50 and assembly size are represented with a log scale. The colour of a point indicates the order of the sequenced species. Dotted lines indicate N50 values of 1, 10, 50, and 500 kb and 5 Mb in panel **A** and assembly size values of 50 Mb, 250 Mb, 1 Gb, and 5 Gb in panel **B**.

Supplementary Figure S5. BUSCO dataset comparisons for Insecta and Hymenoptera. Congruence of BUSCO completeness scores is assessed by comparing results from the Arthropoda (n = 1,013) and Insecta (n = 1,367) lineage datasets (**A**), the Insecta and 4 insect order-level lineage datasets (Hemiptera [n = 2,510], Hymenoptera [n = 5,991], Lepidoptera [n = 5,286], Diptera [n = 3,285]) (**B**), and lineage datasets from BUSCO v4 (OrthoDBv10) and BUSCO v3 (OrthoDBv9) for Insecta (odb9: n = 1,658) (**C**) and Hymenoptera (odb9: n = 4,415) (**D**). Dotted lines represent the identity (y = x).

Abbreviations

API: application programming interface; BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; Gb: gigabase pairs; IQR: interquartile range; JSON: JavaScript Object Notation; kb: kilobase pairs; Mb: megabase pairs; NCBI: National Center for Biotechnology Information.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was supported by Novartis Foundation for medical-biological research grant No. 18B116 and Swiss National Science Foundation grants PP00P3_170664 and PP00P3_202669 to R.M.W.

Authors' Contributions

R.M.W. conceived the study. R.F. developed the workflows and performed the analyses. Both authors wrote the manuscript and read and approved the manuscript.

Acknowledgements

The authors thank Sagane Dind, Giulia Campli, Livio Ruzzante, and Antonin Thiébaut, as well as the reviewers, for providing useful suggestions for improvements and valuable feedback on the workflow and the manuscript.

References

1. Richards, S. It's more than stamp collecting: how genome sequencing can unify biological research. *Trends Genet* 2015;**31**(7):411–21.
2. Lewin, HA, Robinson, GE, Kress, WJ, *et al.* Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci U S A* 2018;**115**(17):4325–33.
3. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* 2020;**587**(7833):240–5.
4. Feng, S, Stiller, J, Deng, Y, *et al.* Dense sampling of bird diversity increases power of comparative genomics. *Nature* 2020;**587**(7833):252–7.
5. Thrash, A, Hoffmann, F, Perkins, A. Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics* 2020;**21**(S4):249.
6. Dohmen, E, Kremer, LPM, Bornberg-Bauer, E, *et al.* DOGMA: domain-based transcriptome and proteome quality assessment. *Bioinformatics* 2016;**32**(17):2577–81.
7. Kemena, C, Dohmen, E, Bornberg-Bauer, E. DOGMA: a web server for proteome and transcriptome quality assessment. *Nucleic Acids Res* 2019;**47**(W1):W507–10.
8. Simão, FA, Waterhouse, RM, Ioannidis, P, *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**(19):3210–2.
9. Waterhouse, RM, Seppey, M, Simão, FA, *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 2018;**35**(3):543–8.
10. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**(D1):D506–15.
11. Sayers, EW, Beck, J, Bolton, EE, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2021;**49**(D1):D10–7.
12. Ewels, P, Magnusson, M, Lundin, S, *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**(19):3047–8.
13. Challis, R, Richards, E, Rajan, J, *et al.* BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)* 2020;**10**(4):1361–74.
14. Waterhouse, RM, Tegenfeldt, F, Li, J, *et al.* OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* 2013;**41**(D1):D358–65.
15. Zdobnov, EM, Kuznetsov, D, Tegenfeldt, F, *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2021;**49**(D1):D389–93.
16. Childers, AK, Geib, SM, Sim, SB, *et al.* The USDA-ARS Ag100Pest Initiative: high-quality genome assemblies for agricultural pest arthropod research. *Insects* 2021;**12**(7):626.
17. Adams, MD. The genome sequence of *Drosophila melanogaster*. *Science* 2000;**287**(5461):2185–95.
18. Favreau, E, Martínez-Ruiz, C, Rodrigues Santiago, L, *et al.* Genes and genomic processes underpinning the social lives of ants. *Curr Opin Insect Sci* 2018;**25**:83–90.
19. Branstetter, MG, Childers, AK, Cox-Foster, D, *et al.* Genomes of the Hymenoptera. *Curr Opin Insect Sci* 2018;**25**:65–75.
20. Garb, JE, Sharma, PP, Ayoub, NA. Recent progress and prospects for advancing arachnid genomics. *Curr Opin Insect Sci* 2018;**25**:51–57.
21. McKenna, DD. Beetle genomes in the 21st century: prospects, progress and priorities. *Curr Opin Insect Sci* 2018;**25**:76–82.

22. Triant, DA, Cinel, SD, Kawahara, AY. Lepidoptera genomes: current knowledge, gaps and future directions. *Curr Opin Insect Sci* 2018;**25**:99–105.
23. Wiegmann, BM, Richards, S. Genomes of Diptera. *Curr Opin Insect Sci* 2018;**25**:116–24.
24. Ruzzante, L, Reijnders, M, Waterhouse, RM. Of genes and genomes: mosquito evolution and diversity. *Trends Parasitol* 2019;**35**(1):32–51.
25. Panfilio, KA, Angelini, DR. By land, air, and sea: hemipteran diversity through the genomic lens. *Curr Opin Insect Sci* 2018;**25**:106–15.
26. González, VL, Devine, AM, Trizna, M, *et al.* Open access genomic resources for terrestrial arthropods. *Curr Opin Insect Sci* 2018;**25**:91–98.
27. Richards, S, Childers, A, Childers, C. Editorial overview: Insect genomics: Arthropod genomic resources for the 21st century: It only counts if it's in the database! *Curr Opin Insect Sci* 2018;**25**:iv–vii.
28. i5K Consortium. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* 2013;**104**(5):595–600.
29. Brown, SJ, Tagu, D. Editorial overview: Insect genomics: How to sequence five thousand insect genomes? *Curr Opin Insect Sci* 2015;**7**:iv–v.
30. Waterhouse, RM. A maturing understanding of the composition of the insect gene repertoire. *Curr Opin Insect Sci* 2015;**7**:15–23.
31. Li, F, Zhao, X, Li, M, *et al.* Insect genomes: progress and challenges. *Insect Mol Biol* 2019;**28**(6):739–58.
32. Hotaling, S, Kelley, JL, Frandsen, PB. Aquatic insects are dramatically underrepresented in genomic research. *Insects* 2020;**11**(9):601.
33. Hotaling, S, Sproul, JS, Heckenhauer, J, *et al.* Long-reads are revolutionizing 20 years of insect genome sequencing. *Genome Biol Evol* 2021;**13**(8):doi:10.1093/gbe/evab138.
34. Köster, J, Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**(19):2520–2.
35. Mölder, F, Jablonski, KP, Letcher, B, *et al.* Sustainable data analysis with Snakemake. *F1000Res* 2021;**10**:33.
36. Feron, R. a3cat-workflow. https://gitlab.com/evogenlab/a3cat-workflow. Accessed 21 October 2021.
37. Schoch, CL, Ciufo, S, Domrachev, M, *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020;**2020**:doi:10.1093/database/baaa062.
38. Roskov, Y, Ower, G, Orrell, T, *et al.* Catalogue of Life - 2019 Annual Checklist. 2020. http://www.catalogueoflife.org/annual-checklist/2019/info/ac. Accessed 13 May 2020.
39. Ellis, EA, Storer, CG, Kawahara, AY. De novo genome assemblies of butterflies. *Gigascience* 2021;**10**(6):doi:10.1093/gigascience/giab041.
40. Zhou, Y, Liang, Y, Yan, Q, *et al.* The draft genome of horseshoe crab *Tachypleus tridentatus* reveals its evolutionary scenario and well-developed innate immunity. *BMC Genomics* 2020;**21**(1):137.
41. Shingate, P, Ravi, V, Prasad, A, *et al.* Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution. *Nat Commun* 2020;**11**(1):2322.
42. Shingate, P, Ravi, V, Prasad, A, *et al.* Chromosome-level genome assembly of the coastal horseshoe crab (*Tachypleus gigas*). *Mol Ecol Resour* 2020;**20**(6):1748–60.
43. Nong, W, Qu, Z, Li, Y, *et al.* Horseshoe crab genomes reveal the evolution of genes and microRNAs after three rounds of whole genome duplication. *Commun Biol* 2021;**4**(1):83.
44. Korhonen, PK, Gasser, RB, Ma, G, *et al.* High-quality nuclear genome for *Sarcoptes scabiei*—A critical resource for a neglected parasite. *PLoS Negl Trop Dis* 2020;**14**(10):e0008720.
45. Waterhouse, RM, Seppey, M, Simão, FA, *et al.* Using BUSCO to assess insect genomic resources. *Methods Mol Biol* 2019;**1858**:59–74.
46. Mikheyev, AS, Zwick, A, Magrath, MJL, *et al.* Museum genomics confirms that the Lord Howe Island stick insect survived extinction. *Curr Biol* 2017;**27**(20):3157–3161.e4.
47. Greenhalgh, R, Dermauw, W, Glas, JJ, *et al.* Genome streamlining in a minute herbivore that manipulates its host plant. *eLife* 2020;**9**:doi:10.7554/eLife.56689.
48. Zhao, L, Zhang, X, Qiu, Z, *et al.* De novo assembly and characterization of the *Xenocatantops brachycerus* transcriptome. *Int J Mol Sci* 2018;**19**(2):520.
49. Kriventseva, EV, Kuznetsov, D, Tegenfeldt, F, *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2019;**47**(D1):D807–11.
50. Zdobnov, EM, Tegenfeldt, F, Kuznetsov, D, *et al.* OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* 2017;**45**(D1):D744–9.
51. Waterhouse, RM. a3cat. https://rmwaterhouse.org/a3cat. Accessed 21 October 2021.
52. Feron, R. a3cat-user-workflow. GitLab. https://gitlab.com/evogenlab/a3cat-user-workflow. Accessed 21 October 2021.
53. Manni, M, Berkeley, MR, Seppey, M, *et al.* BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;**38**(10):4647–54.
54. NCBI Datasets. https://www.ncbi.nlm.nih.gov/datasets. Accessed 21 October 2021.
55. Huerta-Cepas, J, Serra, F, Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 2016;**33**(6):1635–8.
56. DataTables | Table plug-in for jQuery. https://datatables.net. Accessed 21 October 2021.
57. Feron, R. paper-busco-v3 · Waterhouse Lab /a3cat-workflow. GitLab. https://gitlab.com/evogenlab/a3cat-workflow/-/releases/paper-busco-v3. Accessed 21 October 2021.
58. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer; 2009.
59. Yu, G, Smith, DK, Zhu, H, *et al.* ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;**8**(1):28–36.
60. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
61. Feron, R. paper-a3cat. GitLab. https://gitlab.com/evogenlab/paper-a3cat. Accessed 21 October 2021.
62. Feron, R, Waterhouse, R. Supporting data for "Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes." *GigaScience Database* 2022. http://dx.doi.org/10.5524/100974.