G3 Genes | Genomes | Genetics

INVESTIGATION

BlobToolKit – Interactive Quality Assessment of Genome Assemblies

Richard Challis,*,†,1 Edward Richards,‡ Jeena Rajan,‡ Guy Cochrane,‡ and Mark Blaxter*,†

*Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK, †Wellcome Sanger Institute, Cambridge CB10 1SA, UK, and ‡European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK

ORCID IDs: 0000-0002-3502-1122 (R.C.); 0000-0001-5975-6003 (E.R.); 0000-0002-3613-0013 (J.R.); 0000-0001-7954-7057 (G.C.); 0000-0003-2861-949X (M.B.)

ABSTRACT Reconstruction of target genomes from sequence data produced by instruments that are agnostic as to the species-of-origin may be confounded by contaminant DNA. Whether introduced during sample processing or through co-extraction alongside the target DNA, if insufficient care is taken during the assembly process, the final assembled genome may be a mixture of data from several species. Such assemblies can confound sequence-based biological inference and, when deposited in public databases, may be included in downstream analyses by users unaware of underlying problems. We present BlobToolKit, a software suite to aid researchers in identifying and isolating non-target data in draft and publicly available genome assemblies. BlobToolKit can be used to process assembly, read and analysis files for fully reproducible interactive exploration in the browser-based Viewer. BlobToolKit can be used during assembly to filter non-target DNA, helping researchers produce assemblies with high biological credibility. We have been running an automated BlobToolKit pipeline on eukaryotic assemblies publicly available in the International Nucleotide Sequence Data Collaboration and are making the results available through a public instance of the Viewer at https://blobtoolkit.genomehubs.org/view. We aim to complete analysis of all publicly available genomes and then maintain currency with the flow of new genomes. We have worked to embed these views into the presentation of genome assemblies at the European Nucleotide Archive, providing an indication of assembly quality alongside the public record with links out to allow full exploration in the Viewer.

KEYWORDS
Bioinformatics
visualisation
web-tool
genome
assembly
quality control

Genome sequences are part of the basic data economy of modern bioscience. Using assembled genomes, it is possible to identify loci underpinning key traits of interest, discover the regulatory logic of gene expression, investigate disease processes, and explore the evolutionary histories of genes and species. These research programs rely implicitly on the correctness of the genome sequences. Errors in genome sequences risk distracting or even derailing their effective use.

doi: https://doi.org/10.1534/g3.119.400908
Manuscript received November 14, 2019; accepted for publication February 15, 2020; published Early Online February 18, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.10303865.
[1]Corresponding author: Wellcome Sanger Institute, Cambridge CB10 1SA, UK. Email: rc28@sanger.ac.uk

Assembly of true genome sequences from reads shorter than the length of a replicon remains a difficult task (Ekblom and Wolf 2014). This task is made more complex when isolation of the original samples or the processing of DNA to generate the raw sequence data cannot avoid contamination of the target genome with DNA from non-target sources (Salzberg *et al.* 2005; Salter *et al.* 2014). Sequencing instruments are agnostic as to species-of-origin of the fragments they are tasked with processing, and thus a contaminated sample will result in a contaminated raw dataset. If insufficient care is taken during the assembly process, this can mean that the final assembled genome is a mixture of data from several species, and cannot be used as a good representation of the target species (Merchant *et al.* 2014). Downstream, this can result in erroneous attribution of biochemical or genetic properties to the target species that are actually derived from the contaminants' genomes (Artamonova *et al.* 2015; Arakawa 2016).

However, not all "contaminants" are uninteresting. Many eukaryotic species live in close biological association with symbionts, and

many bacteria exist in, and can only be grown as, consortia of interacting species (López-García *et al.* 2017). In these systems genome sequencing aims to reconstruct the genomes of all the independent species and strains involved (Kumar and Blaxter 2011).

We are developing BlobToolKit, a software suite that will aid researchers in identifying contamination before it is erroneously blessed as being part of a target genome and to separate sequences that belong to different members of biological consortia. BlobToolKit is based on BlobTools written by Dominik Laetsch (Laetsch and Blaxter 2017) which was in turn based on the original Blobology pipeline from Sujai Kumar (Kumar *et al.* 2013). We present a toolkit that has been rewritten in its entirety to make use of advanced web frameworks and visualization. Like its progenitors, BlobToolKit uses GC proportion and coverage as two major axes on which contigs or scaffolds from an assembly can be displayed and overlays a taxonomic assignment to help separate the signal from noise when using these measures to identify sequence from different sources.

Genomes from different taxa frequently have different mean GC proportion, with values ranging from 34–50% among monocot species (Šmarda *et al.* 2014), 44–58% among mammals (Romiguier *et al.* 2010) and 16–75% among bacteria (Lightfield *et al.* 2011). Intra-genomic GC proportion typically shows a unimodal distribution around a mean value, particularly in bacteria (Bohlin *et al.* 2010), however, regions of differing composition may yield bi- (or multi-) modal distributions (Salinas *et al.* 1988). These intragenomic differences result from biological processes, such as biased gene conversion (Galtier *et al.* 2001), and may be as large or larger than differences between species (*e.g.*, 35–60% in the human genome (Lander *et al.* 2001)). Primary separation of sequences from different sources may be possible on this axis, provided the organisms have distinct mean values with relatively little deviation.

The read coverage of each contig or scaffold in an assembly is an estimate of the relative stoichiometry of the replicon from which it derives. All the contigs or scaffolds from one species should have similar coverage, however, as with GC content, there are a number of confounding factors to consider. Organelles will usually have high coverage relative to the nuclear genome (Bakker *et al.* 2016) and collapsed segmental duplications in the assembly will also have elevated coverage (Bailey *et al.* 2002). Sex chromosomes are expected to have 50% coverage in the heterogametic sex (Tomaszkiewicz *et al.* 2017) and uncollapsed haploid segments of autosomes will also have 50% coverage. Contaminant or cobiont genomes are likely to have different, internally consistent stoichiometry, in which case they may be distinguished on this axis.

To provide initial identification of sets of contigs or scaffolds from distinct taxa, BlobToolKit also decorates each scaffold or contig with a taxonomic attribution. This is assessed using the sequence similarity search tools BLAST (Altschul 1997) and Diamond (Buchfink *et al.* 2015) to perform local alignments against public sequence databases to identify the most closely matching sequences. Two tools are used as BLAST is most efficient at identifying closely related sequences in nucleotide databases, while Diamond is more efficient when searching against protein databases (Buchfink *et al.* 2015). This taxonomic attribution is tentative due to the presence of mis-annotated records in the public databases. In conjunction with GC proportion and coverage measures this serves to highlight clusters (or blobs) of contigs that share distinct properties and coherent taxonomic source.

This richly marked-up annotation of the assembly makes it possible to assess whether it derives from single or multiple source organisms. The BlobToolKit data can be used to separate contigs and scaffolds (and the reads that generated them) into separate bins for subsequent reanalyses. BlobToolKit can be used as part of the process of genome assembly, playing a role both in separating raw input data for assembly of distinct components and in quality assurance of the final product. For genome assemblies released publicly, BlobToolKit can be used to provide quality assurance and to identify issues that should be taken into consideration in downstream reuse of the data.

Here we present the latest version of BlobToolKit, show how it can be used to probe the integrity of genome assemblies, describe the visualizations available and present snapshots of our ongoing BlobToolKit analyses of all eukaryotic genome assemblies available in the European Nucleotide Archive (ENA) (Amid *et al.* 2019).

MATERIALS AND METHODS

BlobToolKit is comprised of four distinct components: BlobTools2 (command line tools to create and filter datasets), Specification (a formal specification and validator for the JSON-based data format), Viewer (interactive dataset visualization), and INSDC-pipeline (a Snakemake pipeline to run the BlobToolKit workflow on publicly available datasets). A Docker image (BlobToolKit-Docker) containing each of these components and their dependencies is also available.

BlobTools2

BlobTools2 is a command line program to import a genome assembly together with BLAST, Diamond, read mapping and BUSCO analysis output files to generate a dataset that can be filtered using the command line and/or explored interactively in a web browser using the BlobToolKit Viewer.

BlobDir format

BlobTools2 is a re-implementation of BlobTools (Laetsch and Blaxter 2017), written in python3 and based around a *BlobDir* directory of JSON format files. This data structure has been chosen as it can be easily validated using JSON-schema and is highly extensible. Separate JSON files contain distinct attributes of the assembly, with one entry per contig or scaffold. The attributes include GC proportion, length, coverage from a single sequencing library, taxonomic inference based on BLAST hits. Because the attributes are treated as generic datatypes (identifiers, variables, categories, arrays of categories or variables and arrays of arrays), it is possible to incorporate results from new analyses without making significant changes to the codebase. Field metadata are collated in a single JSON file allowing basic dataset information to be accessed without loading the full set of values. JSON is the native format for the JavaScript-based BlobToolKit Viewer and the typical patterns of use require computation across all data for a given attribute at once. Because the Viewer architecture inverts the usual server-client model, pushing computation to the client, this *BlobDir* format was favored for efficiency of data access over alternatives such as SQLite or HDF5.

Adding data to a BlobDir

Assembly: The minimum input required to create a new *BlobDir* dataset is a FASTA format assembly sequence file. This is parsed to generate a list of sequence identifiers, along with a set of basic, per-sequence statistics (length, GC proportion and undefined [N] bases). Additional metadata, including assembly accessions and taxonomic information can be provided for inclusion in the dataset metadata and, if an NCBI taxonomy ID (taxid) is provided, expanded taxonomic lineage details will be included. The *BlobDir* can be modified, for example, to add attributes based on new analyses, using the BlobTools2 *add* command.

Coverage: Both base and read coverage are calculated for each contig by parsing read alignment files in BAM, SAM or CRAM formats using the pysam library (https://github.com/pysam-developers/pysam).

Taxonomy: Taxonomy information is assigned to contigs and scaffolds through parsing of similarity searches of taxonomically-annotated sequence databases. Rather than simply use a single, top-scoring hit for each contig or scaffold, BlobTools2 uses simple taxonomy rules (taxrules) to deliver a best-supported assignment. BlobTools2 deploys taxrules introduced in BlobTools to assign putative taxonomic associations to sequence contigs: *bestsum* (total bitscore of all hits across all databases) and *bestsumorder* (total bitscore from a single database search, with scores taken from successive databases for contigs or scaffolds that failed to identify hits in the first). In a typical use case a file of NCBI BLAST+ *blastn* hits to the NCBI nt nucleotide database and a file of Diamond *blastx* hits to the UniProt/SwissProt database are supplied to be processed under one of these taxrules to generate a set of JSON files. For each of eight taxonomic ranks from superkingdom to species, files are generated containing the most likely taxon name, the summed bitscore of all hits to that taxon, a c-index value indicating the number of alternate taxa at that rank, and taxon names for every hit to each contig or scaffold. An additional file shows the location, score and taxid for every hit, information that is independent of the taxonomic rank under consideration. Results are split across multiple files to allow faster access to individual components during subsequent analyses.

BUSCO: As an example of the incorporation of new analyses, BUSCO (Benchmarking Universal Single-Copy Orthologs), a widely used tool for quality assessment of genome assemblies (Waterhouse *et al.* 2018) generates a sparse annotation where a few contigs are decorated with the presence of a BUSCO reference gene. BlobTools2 incorporates BUSCO using the same basic datatype as BLAST hit distributions. The only unique code occurs in a specially written parser module for the BUSCO file format.

Hyperlinks: Hyperlink templates can also be added to the *BlobDir* metadata to allow hyperlinks from assembly/taxon identifiers, individual sequence identifiers or individual BLAST/Diamond hits to external resources.

Applying filters

BlobTools2 supports filtering based on any of the contig/scaffold level attributes in a dataset. All filters can be applied to assembly, read or analysis files or to *BlobDir* datasets. This allows subsets of both input files and datasets to be obtained without the need to use external command line tools as was the case for previous blobology/BlobTools implementations.

Variable attributes, such as GC proportion or length can be filtered to include or exclude scaffolds with a given range of values by setting a maximum and/or minimum. Category attributes (principally taxonomic assignments) can be filtered by presence or absence of one or more keys. For example, to exclude all scaffolds with 'no-hit' at the phylum level, or to include only scaffolds assigned to the superkingdom 'Eukaryota'. A list of scaffold identifiers can also be used as the basis for filtering to keep or exclude records associated with specific sequences.

Filtering of assembly and read files can assist in the process of iterative assembly improvement, while filtering of analysis files and datasets may allow more detailed interrogation of subsets of the data in the BlobToolKit viewer without the need to repeat analyses or filter analysis outputs for re-importing.

Specification

The BlobToolKit Specification describes the file formats required by BlobTools2 and the BlobToolKit Viewer and includes a validator that tests a *BlobDir* dataset for departures from the specification. Use of JSON format allows validation with JSON-schema. While basic validation is possible with a static schema, the validator generates and tests against dynamically generated schemas to allow for the dependence of some metadata values on the presence and content of data in field-specific files. Validation includes type checking, testing for presence and content of expected files and assessing metadata ranges against the values present in corresponding field files.

BlobToolKit Viewer

The BlobToolKit Viewer allows interactive exploration of *BlobDir* datasets produced by BlobTools2.

Application programming interface

All data in a *BlobDir* can be made available through an application programming interface (API) implemented using the Express Node.js web framework (https://expressjs.com/). The API provides search functionality against entries in the *assembly* and *taxon* sections of the metadata along with direct access to datasets, fields and individual records within fields. Full API documentation is available at https://blobtoolkit.genomehubs.org/api-docs/.

Interactive data exploration

The BlobToolKit Viewer presents data retrieved via the API in a set of interactive views for dataset visualization, exploration and filtering. The Viewer is built on the React (https://reactjs.org) JavaScript library. It makes extensive use of Redux and reselect frameworks to allow real-time interaction with genome-scale datasets in client web browsers. This makes it practical to host large numbers of publicly accessible datasets on a server with a relatively small footprint. For datasets that are too large to be processed on the fly (including those with millions of contigs), pre-generated static image files can be served in place of the interactive views. Interactive plots are powered by the d3 data visualization library (https://d3js.org) and all plots can be exported directly as PNG or SVG image files.

Filters view: The Viewer supports the same set of filter parameters as BlobTools2. Filter controls provide a graphic representation of category or variable distributions. To reduce network overheads, only data for active fields are loaded into the browser and the filter view provides an indication of which data are currently available. All views update instantly based on changes to filters.

Blob view: Blobology and BlobTools introduced the blob plot in which contigs are represented as circles with areas proportional to contig length. This representation has several computational and interpretation issues. Circles are computationally expensive to plot, and rendering of datasets with many contigs (some published assemblies have over 1 million) makes it impossible to see all the data. While the scaled circle view is available in the Viewer, a square-binned blob plot of GC proportion *vs.* coverage is the default view when opening a new dataset. The squares are scaled to the square-root of the sum of lengths of contigs within each bin and colored by best-matching phylum (Figure 1A). Binning resolves problems with scaled circles by limiting the number of data points that need to be plotted, reducing both the computational expense and the potential for data to be obscured.

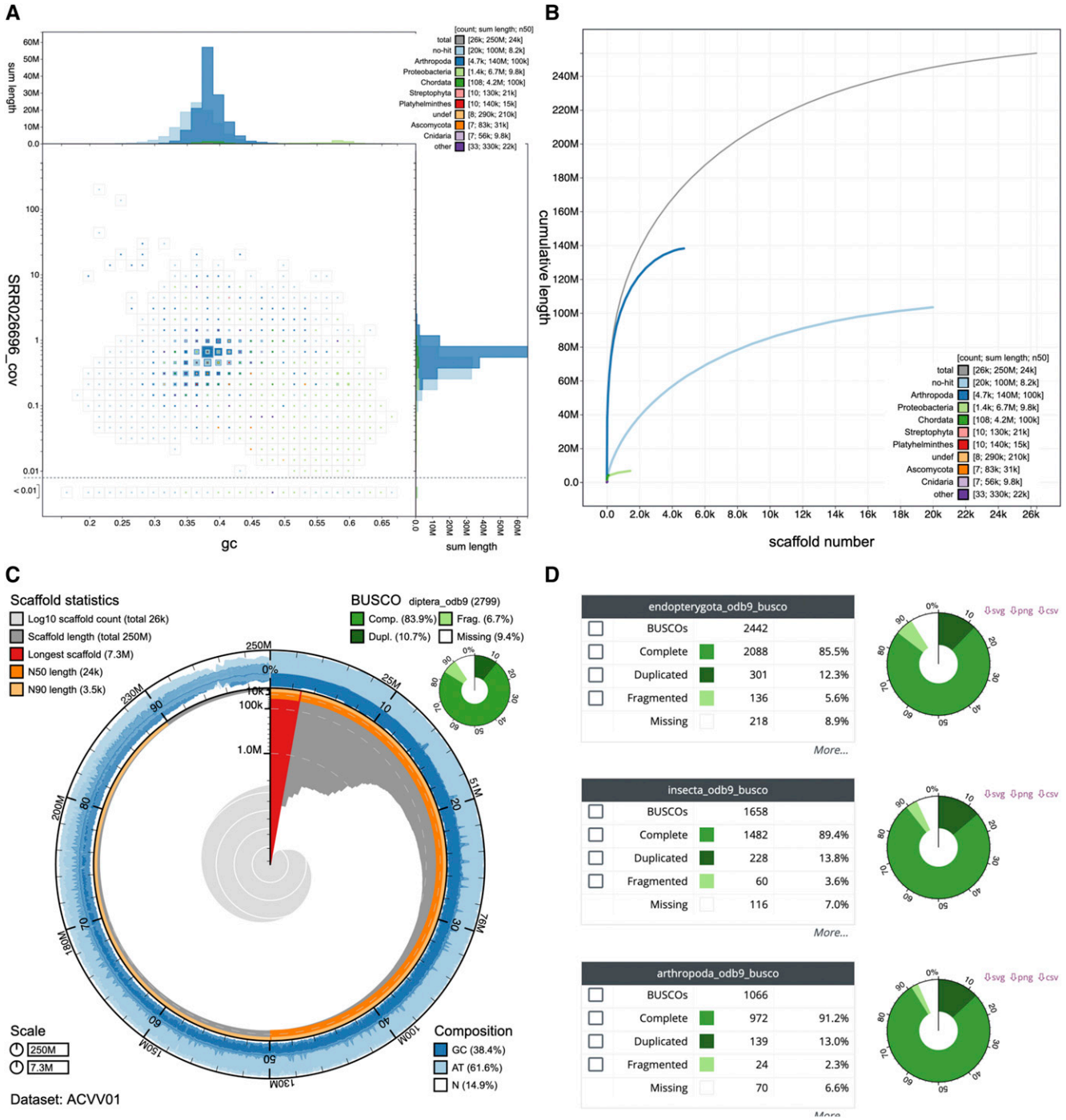

Figure 1 Assembly views available in the BlobToolKit Viewer, illustrated using the *Drosophila albomicans* assembly ACVV01 (Zhou *et al*. 2012). (A) Square-binned blob plot showing the distribution of assembly scaffolds on GC proportion and coverage axes. Squares within each bin are colored according to taxonomic annotation and scaled according to total span. Scaffolds within each bin can be selected for further investigation. (B) Cumulative assembly span plot showing curves for subsets of scaffolds assigned to each phylum relative to the overall assembly. (C) Snail plot summary of assembly statistics. (D) BUSCO scores allow selection of all scaffolds with a BUSCO reference gene in each category. These images derive from analyses of the whole assembly. Each view updates automatically in response to any filters or selections that are applied to the dataset. This figure can be regenerated, and explored further, using the URLs given in File S1.

Binning effectively divides the plot into an evenly-spaced grid (of squares or hexagons) so the area of each color on the plot is proportional to the total span of contigs assigned to each category. Where multiple categories are represented in a single bin, the squares are rendered in size order so each category remains visible in a set of concentric squares.

Binning parameters such as the resolution (how many divisions on each axis) and scaling function (alternatives to square-root are linear and logarithmic scales) can be changed interactively and update on the plot in real time. This can be a useful way to explore features that are not immediately clear in a static image. Alternative scaling functions can

increase or reduce the prominence of smaller values, for example using a log scale can increase the relative size of squares for categories with a small total span relative to the target taxon, potentially making it easier to identify contaminants or cobionts.

Binning effectively divides the plot into an evenly-spaced grid (of squares or hexagons), making it straightforward to select all contigs within a chosen GC proportion-coverage bin. Adjusting the resolution of the plot can make each bin larger (*e.g.*, to facilitate the selection of major features), or smaller to highlight fine-scale patterns, such as the off-axis bimodality associated with heterozygosity.

These options are available for plots of any variable in the dataset against any other variable, for example, to allow coverage *vs.* coverage plots to identify contigs that are only supported by one sequencing library. Categories may be assigned based on any of the taxonomic ranks that have been calculated.

Cumulative view: The Cumulative view is a commonly used representation of the fraction of the genome that is represented as size-ordered contigs are added to the assembly (Figure 1B). These plots also show this cumulative distribution broken down according to taxonomic assignments (the default is by phylum) and allow these separate curves to be stacked to show cumulative span by taxon.

Snail view: The Snail view is a reimplementation of interactive assembly statistic plots introduced in the Lepbase project (Challis *et al.* 2016). These capture a rich variety of assembly properties in a single dynamic graphic (Figure 1C). Snail plots can highlight specific features of an assembly that may not be immediately apparent from tabulated data.

BUSCO view: If BUSCO scores are added to a dataset, the BUSCO view shows a summary of the counts in each BUSCO category (complete, fragmented, etc.) under the current set of filters Figure 1D. It also allows selection of all contigs within a BUSCO category so that their distribution can be seen in the Blob view or the contigs can be inspected in the Table view. These interactions with other views make it possible to assess the impact of possible cobionts on the overall BUSCO score for an assembly.

Table view: The Table view shows information for each contig for each currently active attribute. The available columns can be controlled by activating or deactivating individual attributes in the Filters view. The default columns show the GC proportion, length, coverage and taxonomic assignment that are used to generate plots in the Blob and Cumulative views. Individual records can be selected (either to view their position in the Blob view or for use in filtering) and rows can be sorted according to selected status or by any of the attribute values.

Hit view: The Hit view shows the distribution of sequence similarity hits to sequence databases along a single contig and can be accessed from the Table view of contigs, and is particularly useful for investigating contigs or scaffolds with unexpected or conflicting taxonomic attribution. The hyperlink functionality can be used to embed links to associated records in public sequence databases.

Detail view: A subset of dataset metadata is presented in a tabular Detail view, together with optional links to external resources. Full dataset metadata can be retrieved in JSON format.

Reproducible analyses: Sharing analyses reproducibly is critical, particularly when many choices have been made to generate a particular filtered dataset or image. To aid in reproducibility the Viewer encodes query parameters within the URL for the displayed data. Parameters developed during interactive filtering can be applied in BlobTools2 (specified individually or using the entire URL or query string) to filter input files and *BlobDir* datasets. Selection-based filters are not stored in the URL due to the potential number of identifiers involved. Selections can be exported and imported via a List menu, which will export a JSON format file that includes a complete list of identifiers based on the current filters, including selections. This file also contains a summary of URL parameters and filtered dataset statistics (including BUSCO scores, span and N50 by taxon, etc.) and can be used to specify filter parameters used within BlobTools2.

Access to views: BlobTools2 provides a *view* command that uses the Selenium WebDriver to provide non-interactive access to all plot types. For datasets with millions of contigs that are too large for practical interactive exploration, use of *view* provides a way to generate static images that will not display in the interactive mode.

INSDC-pipeline

INSDC-pipeline is a reusable Snakemake (Köster and Rahmann 2018) pipeline to run analyses on publicly available, International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/) public eukaryotic genome assemblies. We built the pipeline to automate the generation of *BlobDir* datasets from the available data, including retrieval and formatting of database files, retrieval of sequences for each assembly and the associated raw read files, read mapping, BLAST and Diamond searches, and BUSCO analyses (Figure 2). We have made the results available on a public instance of the BlobToolKit Viewer at https://blobtoolkit.genomehubs.org/view (Table 1).

This workflow broadly follows the BlobTools workflow (Laetsch and Blaxter 2017), but with some changes to increase efficiency. For example, Diamond searches against UniProt are only run for contigs with no BLAST hit to the nt database, and the addition of BUSCO analyses. Query genomes are masked using *windowmasker* to reduce spurious matches to interspersed repeats. A wrapper script for *blastn* splits contigs longer than 100 kb into chunks before running BLAST, to avoid taxonomic inference for longer contigs being dependent on a single region. Since this pipeline was run on public datasets extracted from the same databases that are used to infer taxonomic affiliation, all sequences belonging to the same genus as the query assembly were excluded either before (Diamond) or during (BLAST) sequence similarity searches. Details of settings used and configuration options for the tools used in the main analysis steps are given in Table 2.

The pipeline uses Conda (https://docs.conda.io/projects/conda/en/latest/index.html) environments to load all external dependencies. These are stored as YAML-format files within the INSDC-pipeline repository. The *generate_metadata* step of the pipeline includes the current git commit hash in an extended version of the input configuration file so the specific versions of each program used can be determined from the *BlobDir* metadata. A record of database versions is maintained by including the date of creation in the local database directory names.

ENA integration

We have worked to integrate the analyses generated by BlobToolKit with the genome presentations of the European Nucleotide Archive (ENA) (Amid *et al.* 2019), to enhance understanding and utility of submitted data. Importantly, ENA holds both deposited raw sequence read and genome assembly data and it is possible to mine these data to discover relationships describing which read sets were used in given assemblies. At the time of analysis, of the 7,632 eukaryotic genome assemblies

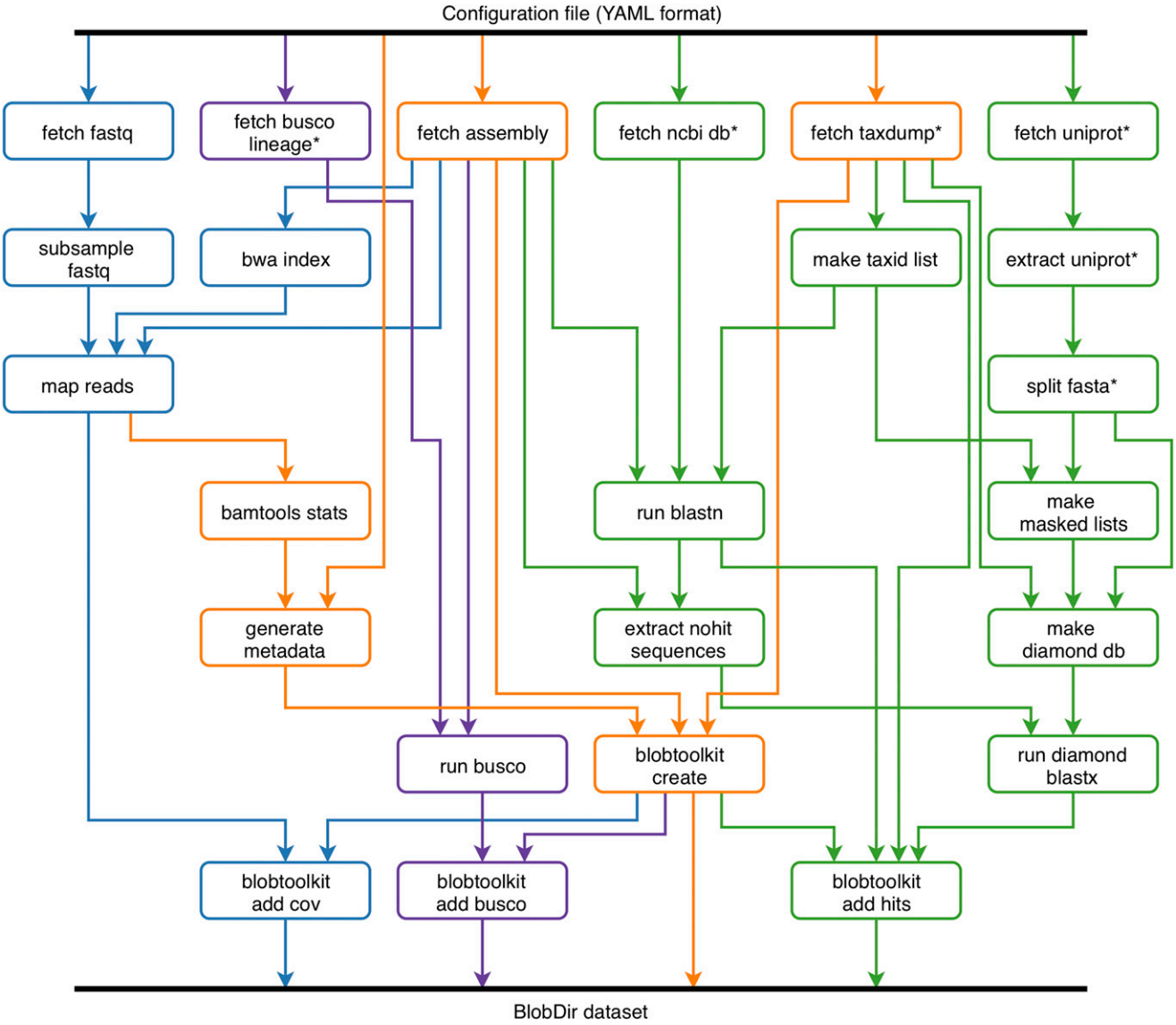

Figure 2 Depiction of the snakemake workflow used to analyze publicly available (INSDC-registered) eukaryotic genome assemblies. The workflow is run once for each assembly. Each box represents a Snakemake rule that may be run one or more times during workflow execution. The workflow can be logically divided into four parts: (i) creation of a minimal *BlobDir* dataset based on a single assembly with metadata derived from the configuration file and additional taxonomic annotation from the NCBI taxdump, shown in orange; (ii) addition of sequence-similarity search results based on *blastn* and Diamond *blastp* searches of the nt and refseq databases, shown in green; (iii) addition of read coverage data based on minimap2 alignment of read files linked to the assembly record (where available), shown in blue; and (iv) addition of BUSCO results based on analyses with all relevant BUSCO lineages, shown in purple. Rules marked with an asterisk are typically only run the first time the pipeline is executed as they generate local copies of relevant database files used elsewhere in the pipeline.

present within the ENA that could be associated with read sets, 585 (8%) were associated with a single run in the raw sequence data, 875 (11%) with between two to four runs, and 6,172 (81%) associated with four or more runs. None of the eukaryotic genome assemblies explicitly referenced the run(s) used to create the assembly within the relevant metadata. Values differ from those presented in Table 1, which uses only data available through the API to make associations between genome assemblies and read sets. We note that the 585 assemblies associated with a single run derived from 266 unique species, potentially permitting the identification of common contaminants in frequently-reassembled taxa. The species with the most independent assemblies were *Saccharomyces cerevisiae*, *Homo sapiens* and *Pyricularia oryzae*. These findings led to the inclusion of user documentation for the process of referencing reads during eukaryotic genome assembly submission to the ENA (https://ena-docs.readthedocs.io/en/latest/submit/assembly/genome.html#submitting-isolate-genome-assemblies). This will encourage future assemblies to be submitted with a referenced run, thereby increasing the number of assemblies for which BlobToolKit can report contamination.

A cross-reference service was set up in conjunction with in-house cloud services for the purpose of processing eukaryotic genome assemblies hosted on the ENA via BlobToolKit, as well as hosting the resulting visual and textual data. The BlobToolKit API was used to access relevant data for each assembly in coordination with Jupyter Notebooks, generating hypertext markup language (HTML) documents for assemblies with links out to associated interactive BlobToolKit Viewer analyses.

Table 1 Summary of assemblies analyzed and available[a] at https://blobtoolkit.genomehubs.org/view on 6th February 2020

Kingdom	Species	Assemblies		
	Total	Total	With reads	Without reads
Fungi	1240/2094	2738/5551	2267/3257	471/2294
Metazoa	938/1838	1311/2900	750/1813	561/1087
Viridiplantae	361/622	655/1214	404/737	251/477
Other Eukaryota	336/421	711/959	350/489	361/470
Total	**2875**/4975	**5415**/10624	**3771**/6296	**1644**/4328

a. For each kingdom within Eukaryota, the numbers of assemblies analyzed/available are shown. Values were obtained through using a scripted query of the ENA and BlobToolKit APIs described in File S1.

Each of these documents displays the respective blob, snail and cumulative length of scaffold by phylum plots, along with assembly statistics directly from the ENA website (see, for example, https://www.ebi.ac.uk/ena/browser/view/GCA_000298335). The generation of these documents is modified autonomously based upon the data available via the API, and uploaded to GitHub Pages respectively.

Data availability

All BlobToolKit code is freely available under open source licenses from https://github.com/blobtoolkit. Current release versions of each of the repositories at the time of writing have been deposited in the Zenodo open access repository: BlobTools2 v2.1, https://doi.org/10.5281/zenodo.3531583; BlobToolKit-Docker v1.0, https://doi.org/10.5281/zenodo.3660946; INSDC-pipeline v1.0, https://doi.org/10.5281/zenodo.3533168; Specification v1.0, https://doi.org/10.5281/zenodo.3531846; and Viewer v1.0, https://doi.org/10.5281/zenodo.3533128. The Docker image is available from https://hub.docker.com/r/genomehubs/blobtoolkit.

File S1 contains URLs and/or commands required to reproduce each of the Figures and Tables in this manuscript alongside full automatically generated captions for each Figure. Files S2 and S3 contain lists of selected scaffolds to allow presented selections to be reproduced. Files S4, S5 and S6 contain the full analyzed datasets for the three assemblies presented in Results. Supplemental material available at figshare: https://doi.org/10.25387/g3.10303865.

RESULTS

The following case studies highlight some of the features of BlobToolKit and the ways it may be used in assessment of published assemblies.

Identification of common cobionts

The *Drosophila albomicans* assembly ACVV01 (GCA_000298335.1) contains 1,440 scaffolds that have greatest sequence similarity to Proteobacteria sequences in the reference databases (nt and UniProt; see File S4 for the analyzed dataset). On a blob plot of GC proportion *vs.*. coverage, many of these scaffolds are found in a distinct blob with higher GC proportion and lower coverage than the majority of the assembled scaffolds (Figure 3A and B). The difference in the distributions of the two sets is highlighted in a kite representation of the data (Figure 3B; see File S2 for a list of selected scaffolds).

When analyzed at higher taxonomic resolution, the scaffolds assigned to Proteobacteria derive from several distinct species. The majority of proteobacterial scaffolds (representing 4.3 Mb of 6.7 Mb) are assigned to *Acetobacter*, and there are 1.8 Mb of scaffolds assigned to *Gluconobacter* (Figure 3C). The *Gluconobacter* scaffolds have a lower

Table 2 External program versions, settings and configuration options used during the main analysis steps

Rule	Program (version)	settings
subsample fastq	Seqtk sample (1.2)[a]	Subsample by proportion calculated from configurable maximum coverage (default: 100x).
bwa index	Bwa index (0.7.17)[b]	Algorithm 'bwtsw' used for all assemblies.
map reads	Minimap2 (2.11)[c]	Preset ('sr', 'map-pb' or 'map-nt') based on input read type.
bamtools stats	Bamtools stats (2.5.1)[d]	Insert size option used for paired end reads.
run busco	BUSCO (3.0.2)[e]	Genome mode using default settings, lineages are configurable.
run windowmasker	Windowmasker (2.9.0)[f]	Generate counts in binary output format.
run blastn	NCBI blastn (2.9.0)[g]	Uses windowmasker database and specific output format ('6 qseqid staxids bitscore std'). Uses configurable 'max-target-seqs' and 'evalue'. A configurable wrapper script around the blastn executable splits long (default: 100kb) sequences into (default: up to 10) chunks. An optional filter excludes a configurable list of NCBI taxIDs (default: excludes query genus). Requires v5 BLAST database.
extract nohit sequences	Seqtk sample (1.2)[a]	Subsample assembly sequences based on list of IDs with no blastn hit.
make diamond db	Diamond makedb (0.9.19)[h]	Includes taxonomic information for each sequence. Input sequences optionally filtered to exclude configurable list of NCBI taxIDs (default: excludes query genus).
run diamond blastx	Diamond blastx (0.9.19)[h]	Uses 'sensitive' option with specific output format ('6 qseqid staxids bitscore qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore'). Uses configurable 'max-target-seqs' and 'evalue'.

a. https://github.com/lh3/seqtk
b. (Li and Durbin 2010)
c. (Li 2018)
d. https://github.com/pezmaster31/bamtools
e. (Waterhouse *et al.* 2017)
f. (Morgulis *et al.* 2006)
g. (Altschul 1997)
h. (Buchfink *et al.* 2015)

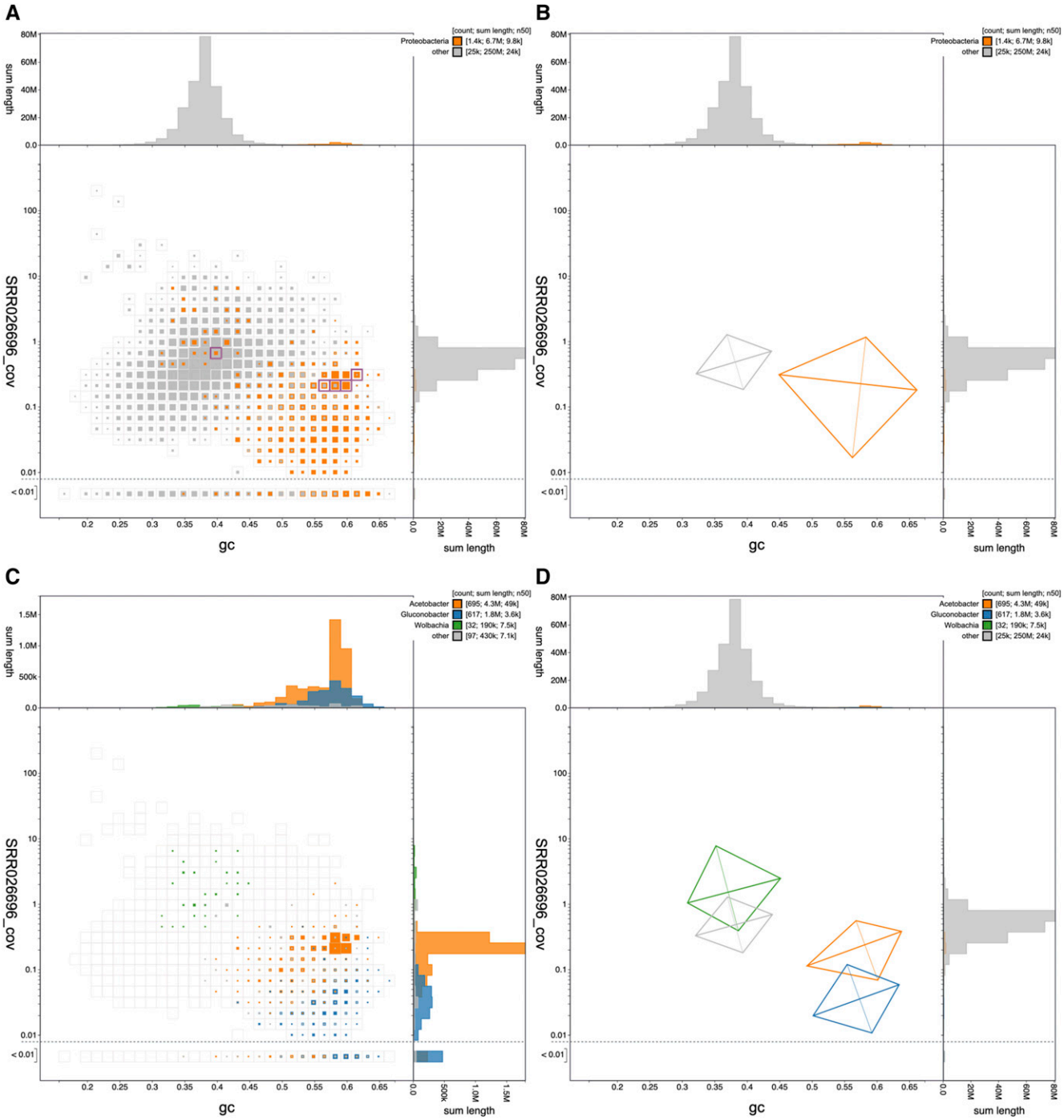

Figure 3 Blobplot of base coverage in read set SRR026696 against GC proportion for scaffolds in *Drosophila albomicans* assembly ACVV01. (A & B) Scaffolds are colored by phylum with Proteobacteria highlighted in orange and all other phyla grouped together in gray. Histograms show the distribution of scaffold length sum along each axis. (A) Square-binned blob plot at a resolution of 30 divisions on each axis. Colored squares within each bin are sized in proportion to the sum of individual scaffold lengths on a logarithmic scale, ranging from 867 to 40,536,114. The bins highlighted in pink contain a total of 5 scaffolds that have been annotated as Proteobacteria but that contain BUSCOs using the diptera_odb9 BUSCO set. (B) A simplified representation of the distributions of scaffolds assigned to each phylum highlights the difference in GC proportion and coverage of Proteobacteria scaffolds. Each kite has a pair of lines representing two standard deviations about the mean on each axis (weighted to account for scaffold lengths) that intersect at a point representing the weighted median. They are angled according to a weighted linear regression equation to indicate the relationship between coverage and GC proportion. (C) Assembly filtered to exclude non-proteobacterial scaffolds. Scaffolds are colored by genus with *Acetobacter* highlighted in orange, *Gluconobacter* shown in blue and *Wolbachia* shown in green. Colored squares within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 1,005 to 771,195. (D) A simplified representation of the distributions of scaffolds assigned to each genus highlights the difference in GC proportion and coverage of *Acetobacter*, *Gluconobacter* and *Wolbachia* scaffolds. This figure can be regenerated, and explored further, using the URLs given in File S1. The list of scaffolds highlighted in (A) is available in File S2.

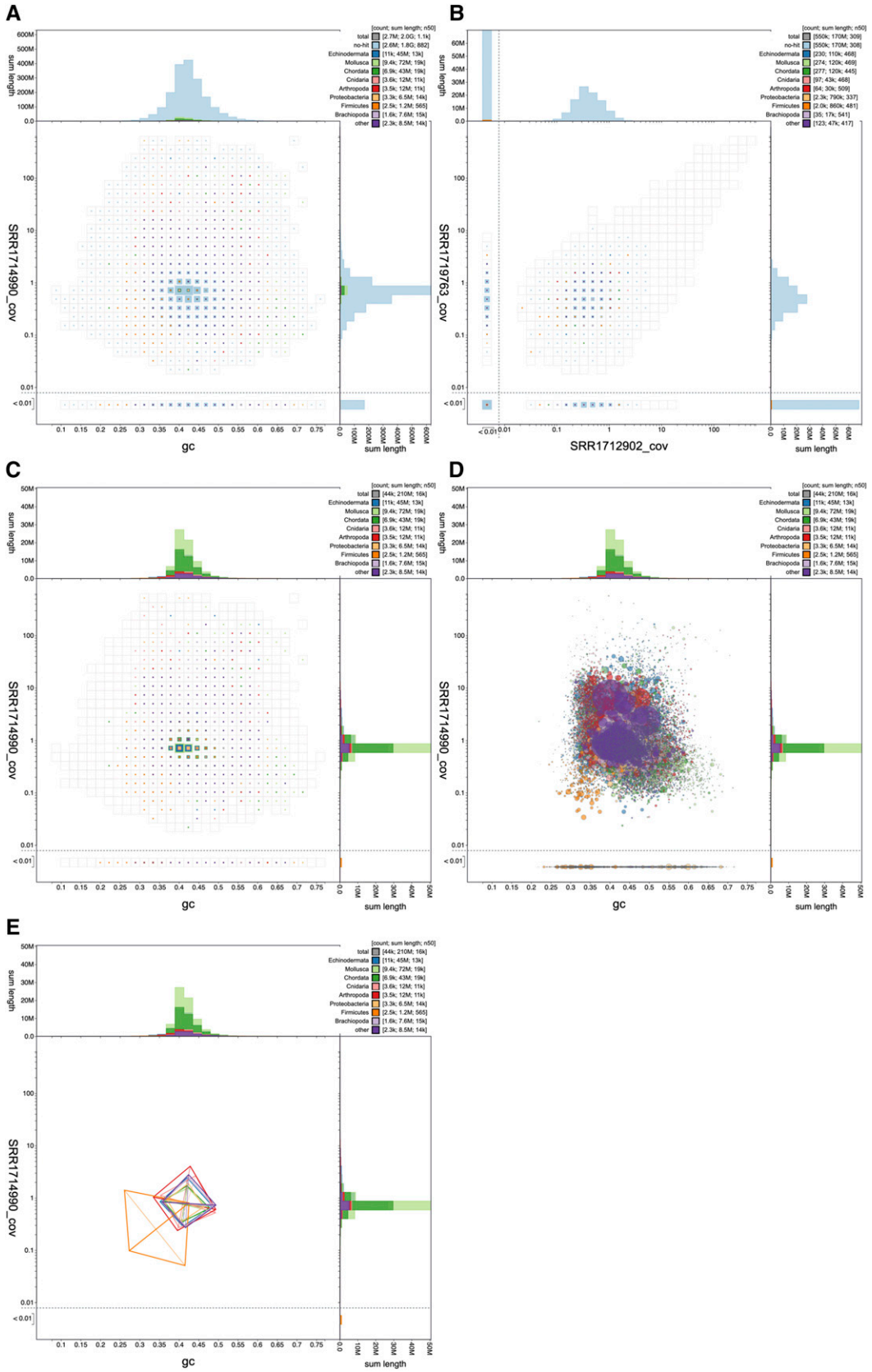

Figure 4 Visualization of the highly fragmented *Conus consors* assembly SDAX01. (A) Binned distribution of all 2,688,687 assembly scaffolds shows unimodal distributions in GC proportion and coverage axes. The majority of scaffolds lack a taxonomic annotation (assigned to "no-hit"). (B) Square-binned plot of coverage in read set SRR1719763 against coverage in SRR1712902 for scaffolds with coverage <= 0.01 in read set

coverage than the *Acetobacter* scaffolds, and thus the assembly is, as expected, less complete. *Acetobacter* and *Gluconobacter* species are common cobionts of *Drosophila* (Crotti *et al.* 2010) and usually have genomes of 3-4 Mb. A third group of scaffolds is assigned to the alphaproteobacterial genus *Wolbachia* (Figure 3D). *Wolbachia* are intracellular symbionts that commonly manipulate the reproductive biology of their hosts (Werren *et al.* 2008), and insect-infecting strains have genomes of ~1.4 Mb. However, the cumulative span of scaffolds assigned to *Wolbachia* is only 190 kb. The GC proportion and coverage of these scaffolds is more congruent with that of the bulk, *Drosophila*-assigned scaffolds. Collectively, these data suggest that the *Wolbachia*-assigned scaffolds are likely to represent nuclear insertions of *Wolbachia* fragments. Such insertions are common in insect genomes, and derive from previous colonization of the species by this endosymbiont (Dunning Hotopp *et al.* 2007).

It is notable that some of the loci identified using the diptera_odb9 BUSCO set (EOG091502LX, EOG091505EO, EOG091502SD, EOG091504TW, EOG09150B43, EOG09150529) are annotated as being present in scaffolds that have been assigned to Proteobacteria. Five of these scaffolds have GC proportions and coverages consistent with their being part of the bacterial rather than the *Drosophila* genomes. Thus the BUSCO assessment of ACVV01 is compromised by the presence in the bacteria of loci which are recognized as being members of the BUSCO dipteran reference gene set. While excluding the BUSCOs identified in the proteobacterial genomes makes a very small difference to the overall BUSCO completeness score of assembly ACVV01 (83.7% *vs.* 83.9% complete; diptera_odb9; BUSCO 3.0.2), their inclusion in, for example, phylogenomic analyses would lead to erroneous inferences. Similar patterns are observed in other *Drosophila* assemblies. For example, in *Drosophila elegans* assembly AFFF02, two diptera_odb9 BUSCOs (EOG0915021D, EOG091501A1) are present on scaffolds assigned to Proteobacteria. The mis-annotated BUSCOs from proteobacterial scaffolds in ACVV01 are found within core Arthropoda scaffolds in AFFF02 and *vice versa*. This highlights the importance of determining assembly integrity and contamination before assessing quality and completeness, and before proceeding to downstream analyses.

Visualization of highly fragmented assemblies

Conus consors is a cone snail studied for its production of neurotoxins (Andreson *et al.* 2019). The *C. consors* assembly SDAX01 (GCA_004193615.1; see https://www.ncbi.nlm.nih.gov/genome/24193) highlights the challenges associated with visualization of highly fragmented datasets. The 2 Gb assembly is split into 2,688,687 scaffolds with an N50 length of 1,128 bp (see File S5 for the analyzed dataset). While the full dataset can be viewed in the BlobToolKit Viewer, interactive visualization of so many contigs requires use of a device with a relatively high-specification (at least 8 GB RAM) and a browser that does not limit the amount of available RAM (*e.g.*, Firefox). To allow such assemblies to be viewed on any device, we have set default parameters to limit the computation required.

The default, binned view (Figure 4A) ensures that the number of graphic elements that must be rendered by the browser does not increase linearly with dataset size as would be the case if each scaffold were plotted individually. This representation is sufficient to show that SDAX01 has a unimodal distribution on both the GC proportion and coverage axes. However 550,837 scaffolds with a total span of over 170 Mbp have coverage below 0.01 with the selected read set (SRR1714990). An assembly of this size is typically based on a number of sequencing runs and in this case nine short read accessions are associated with the same bioproject (PRJNA267645) as the assembly. The largest three of these read sets were mapped to the assembly, allowing comparison of coverage across libraries. For scaffolds with coverage <= 0.01 in SRR1714990, a coverage *vs.*. coverage plot of the remaining two libraries (SRR1719763 and SRR1712902; Figure 4B) shows the majority of these scaffolds (433,970 scaffolds with a total span of over 136 Mb) have coverage in at least one other library. Some have no coverage in any of the three libraries. It might be prudent to consider all these contigs as questionable components of the *C. consors* genome, or artifacts due to heterozygosity or misassembly.

On the public BlobToolKit Viewer site, all datasets with over 1 million scaffolds are presented with a set of pre-generated images so users not wishing to explore beyond the default visualizations have no need to download or process the data files. In interactive mode, the same threshold is used to filter out scaffolds that lack a taxonomic annotation (those assigned to the "no-hit" category) so the default interactive view emphasizes the portion of the dataset that provides most information for contaminant screening (Figure 4C). For this assembly, filtering out "no-hit" scaffolds leaves 43,857 scaffolds (1.6% of all contigs) with a total span of 209 Mb (10.2% of the total span). Below a default threshold of 100,000 scaffolds, it is computationally reasonable to plot individual scaffolds as scaled circles, even on relatively low-powered devices. However, the resulting image can be difficult to interpret as the visibility of specific features becomes dependent on plotting order with the last plotted scaffolds having greatest prominence (Figure 4D). Using a kite representation highlights a distinct distribution of Firmicute scaffolds in the *C. consors* assembly (Figure 4E) suggesting that these represent a contaminant.

Identification of mis-annotated records in public databases

The genomes of many bird species are being generated to understand the evolution of this important group, and to explore the evolutionary genomics of particular phenotypes (Jarvis *et al.* 2014). While most other paleognath birds (kiwis, ostriches, rheas and their kin) are flightless, tinamous can fly, and genomic analyses are exploring the biology of this phenotypic shift (Sackton *et al.* 2019). The genome assembly of the thicket tinamou, *Crypturellus cinnamomeus* (PTEZ01; GCA_003342915.1) (Sackton *et al.* 2019) was analyzed using BlobToolKit (see File S6). We noted that this assembly (total span 1.1 Gb) contained ~130 Mb of scaffolds that had coverage an order of magnitude lower than that of the main part of the assembly (Figure 5A). This blob of scaffolds also had a mean GC proportion of 0.52, contrasting with the main assembly GC proportion of 0.42. Exploring the biology of this set of scaffolds revealed several interesting features.

SRR1714990. The extent of the unfiltered distribution is indicated by the empty square bins. (C) In the interactive browser datasets with over 1,000,000 scaffolds are presented with the "no-hit" scaffolds filtered out to reduce computation. In this case, 43,857 scaffolds are plotted in the filtered dataset. (D) A non-binned presentation of the same data shows the challenges of interpreting a dataset plotted as a large number of overlapping circles, even after filtering "no-hit". (E) A simplified representation of the distributions of scaffolds assigned to each phylum highlights the difference in GC proportion and coverage of scaffolds assigned to Firmicute. This figure can be regenerated, and explored further, using the URLs given in File S1.

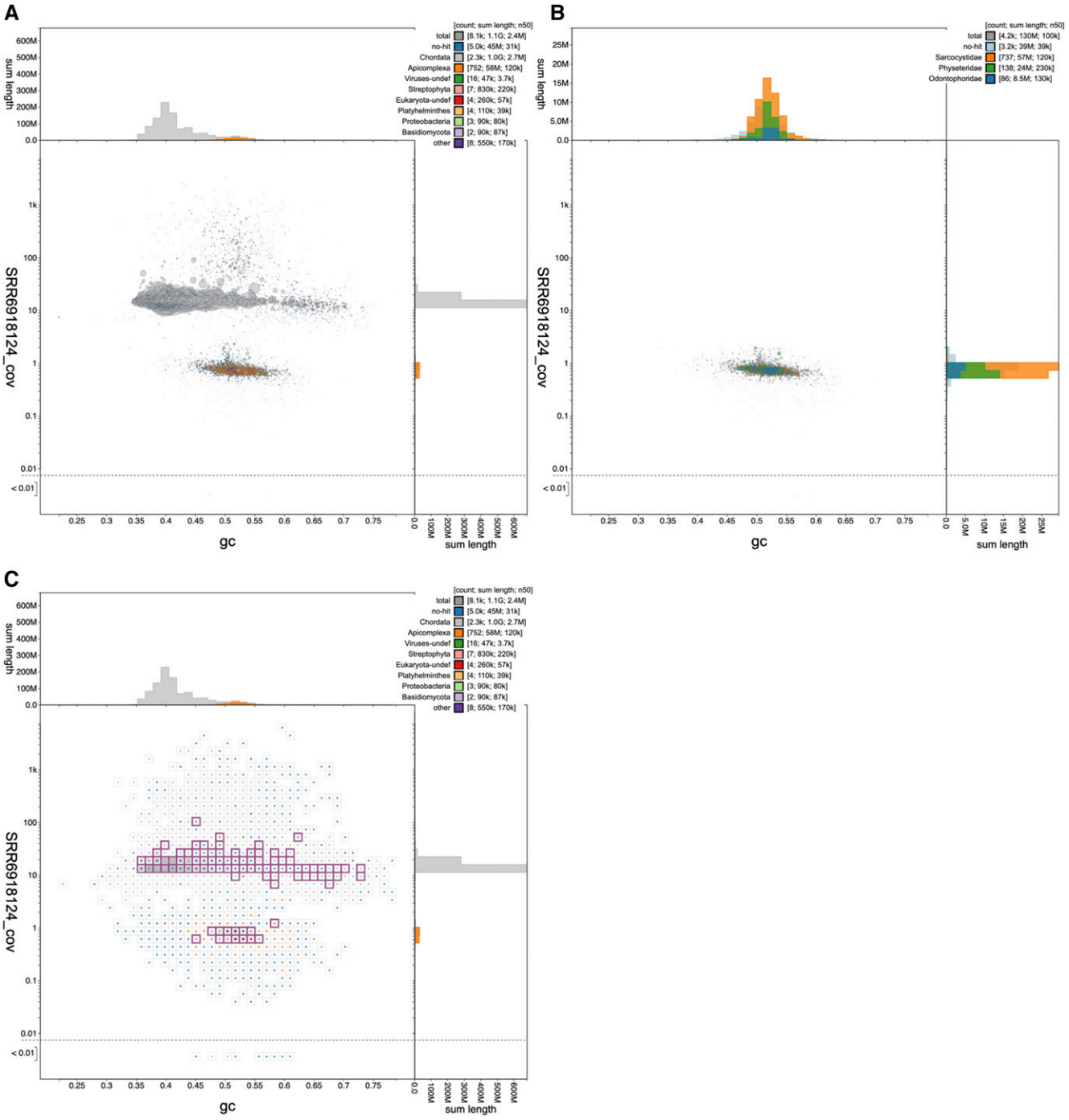

Figure 5 Blob plots of the *Crypturellus cinnamomeus* assembly PTEZ01 showing the presence of an apicomplexan parasite. (A) Circles are scaled with area proportional to scaffold length and colored by phylum. Scaffolds assigned to the phylum Apicomplexa are colored orange and form a distinct blob relative to the majority of Chordata-assigned scaffolds, shown in gray. (B) Circles are colored by family and scaffolds assigned to families other than Physeteridae, Odontophoridae or Sarcocystidae have been filtered out. Scaffolds with coverage greater than 2 in the SRR6918124 read set have also been excluded. (C) A square-binned plot in which bins containing scaffolds with BUSCO annotations using any of the applicable reference gene sets are outlined in pink. This figure can be regenerated, and explored further, using the URLs given in File S1. The list of scaffolds highlighted in (C) is available in File S3.

Half of the span of the low-coverage scaffolds (58 Mb) was assigned to the protist group Eucoccidiorida, and more specifically had high-scoring matches to *Sarcocystis* species (Figure 5B). *Sarcocystis* are apicomplexan parasites that infect a wide range of vertebrate and non-vertebrate hosts. Sarcocystidae, which includes the important pathogens *Neospora* and Toxoplasma, have genomes that range from ~60 Mb to 127 Mb (*Sarcocystis neurona*). The other scaffolds in the low-coverage blob either had no annotation (39 Mb) or were annotated as deriving from a cetacean, *Physeter catodon* (Physeteridae; the sperm whale; 24 Mb) or a galliform bird, *Colinus virginianus* (Odontophoridae;

Table 3 BUSCO scores[a] for the *Crypturellus cinnamomeus* assembly PTEZ01

Lineage	Complete	Duplicated	Fragmented	Missing	Single copy	Total
aves_odb9	92.8% (-0.1%)	1.0% (-0.1%)	4.1% (+0.0%)	3.1% (+0.1%)	91.8% (+0.0%)	4915
tetrapoda_odb9	96.1% (-0.1%)	0.3% (-0.0%)	2.3% (+0.0%)	1.6% (+0.1%)	95.8% (-0.1%)	3950
vertebrata_odb9	97.4% (-0.2%)	0.2% (-0.1%)	1.4% (-0.0%)	1.2% (+0.2%)	97.1% (-0.1%)	2586
metazoa_odb9	91.6% (-3.2%)	1.2% (-0.7%)	3.5% (-0.5%)	4.9% (+3.7%)	90.4% (-2.5%)	978
eukaryota_odb9	91.4% (-8.3%)	3.6% (-3.3%)	4.3% (-1.7%)	4.3% (+9.9%)	87.8% (-5.0%)	303

a. BUSCO analyses were performed using BUSCO 3.0.2 and the indicated ortholog group sets. Numbers in parentheses show the change in score when scaffolds with a coverage below 2 in read set SRR6918124 are removed from the assembly.

the northern bobwhite quail; 8 Mb). While it is possible that a bird genomics laboratory might contaminate across species, the northern bobwhite genome was not sequenced by the same team that sequenced the tinamou, and contamination with sperm whale is hard to imagine. Instead, we infer that the bobwhite and sperm whale genomes are also contaminated by co-assembled genomes from *Sarcocystis*-like apicomplexans. Available *C. virginianus* and *P. catodon* assemblies were analyzed with BlobToolKit to determine the presence of Apicomplexan-assigned scaffolds in these assemblies (Table 4). A total of 48 Mb of the 1.2Gb (4%) of the *C. virginianus* assembly AWGT02 (GCA_000599465.2 (Oldeschulte *et al.* 2017)) is inferred to be derived from an apicomplexan parasite. For *P. catodon*, the only published assembly, AWZP01 (GCA_000472045.1 (Warren *et al.* 2017)) is inferred to be free of contamination with sequences of apicomplexan origin. However, two more recent assemblies, including a chromosome-level assembly PGGR02 (GCA_002837175.2), which is tagged as the RefSeq (Pruitt *et al.* 2005) representative genome, each contain 4.3 Mb of sequence assigned to Apicomplexa.

Thus 11% of the thicket tinamou genome assembly appears to derive not from the target species but rather from a parasite, and sequence from this group of parasites is also present in other genome assemblies from diverse target species. This contamination of the INSDC databases with whole genomes mistakenly attributed to their host species identity means that the public commons becomes an untrustworthy substrate for discovery research. Critically, as with the *D. albomicans* example above (Figure 1), the likely *Sarcocystis*-derived scaffolds contained many BUSCO annotations (Figure 5C; see File S3 for a list of selected scaffolds), and contributed 6% of the unique eukaryote BUSCO hits in the assembly (Table 3).

We have identified additional examples of co-sequencing of apicomplexan pathogens with target species in other taxa (Table 4). These include early assemblies of the model organisms *Mus musculus* and Rattus norvegicus, for which subsequent revisions have been released that have shorter span and few or no remaining apicomplexan-assigned sequences. For non-model organisms the resources available for assembly revision are considerably smaller so it is important to have the means to identify co-sequencing with pathogens and other cobionts. BlobToolKit makes evident these fascinating biological juxtapositions, and facilitates evidence-led separation of host from cobiont. Indeed this task was one of the original motivations for the development of the blob plot: to separate symbiont genomes from those of their hosts (Kumar *et al.* 2013).

DISCUSSION

BlobToolKit is a significant extension of the approach launched in BlobTools. In particular, by permitting user interaction with the rich data associated with each contig in the **Viewer** mode, BlobToolKit can enhance discovery of novel biology. The addition of real-time interaction addresses a criticism of the approach, relative to cluster-based methods such as Anvi'o (Eren *et al.* 2015), that it limits the amount of supporting data that can be included (Delmont and Eren 2016). We envisage three main uses for BlobToolKit. The first is in the research laboratory aiming to sequence for the first time the genome of a new species. BlobToolKit can be used during the assembly process, to filter

Table 4 Presence of Apicomplexa-assigned sequences in selected chordate genome assemblies

	Accession			*Span (Mb)*		
Species	*Blob- ToolKit*	*GCA*	*Date*	*Apicomplexa*	*Chordata*	*Total*
Colinus virginianus	AWGT02	GCA_000599465.2	2017[a]	48.0	1074	1254
Crypturellus cinnamomeus	PTEZ01	GCA_003342915.1	2018[b]	59.9	1017	1122
Mus musculus	AAHY01	GCA_000002165.1	2009[c]	188.9	2738	3251
Mus musculus	LXEJ02	GCA_003774525.2	2018[d]	0.0	2687	2801
Physeter catodon	AWZP01	GCA_000472045.1	2013[e]	0.0	2279	2280
Physeter catodon	UEMC01	GCA_900411695.1	2018[f]	4.3	2472	2512
Physeter catodon	PGGR02	GCA_002837175.2	2019[g]	4.3	2472	2512
Piliocolobus tephrosceles	PDMG02	GCA_002776525.2	2018[h]	33.3	2976	3038
Rattus norvegicus	AAHX01	GCA_000002265.1	2006[i]	15.4	2699	2932
Rattus norvegicus	AABR07	GCA_000001895.4	2014[j]	0.2	2869	2870

a. (Oldeschulte *et al.* 2017)
b. (Sackton *et al.* 2019)
c. (Mural *et al.* 2002)
d. Most recent non-chromosomal assembly (see https://www.ncbi.nlm.nih.gov/assembly/GCA_003774525.2)
e. (Warren *et al.* 2017)
f. (see https://www.ncbi.nlm.nih.gov/assembly/GCA_900411695.1)
g. (see https://www.ncbi.nlm.nih.gov/assembly/GCA_000472045.1)
h. (see https://www.ncbi.nlm.nih.gov/assembly/GCA_000472045.1)
i. (Florea *et al.* 2005)
j. (Gibbs *et al.* 2004)

contaminants and cobionts, and to explore issues such as haploid *vs.* diploid contigs, and patterns of coverage in different sequence read datasets (for example, comparing male and female read sets in heterogametic organisms). As part of an assembly workflow, BlobToolKit should ensure better quality assemblies with higher biological credibility.

The second use is in publication and visualization of published assemblies. The BlobToolKit Viewer generates publication quality images that are fully reproducible via the embedding of control parameters in the URL. These images should, we believe, become standard in reporting genome assemblies, and thus enhance the ease of assessment of assembly quality. We have worked to embed BlobToolKit views into the presentation of genome assemblies at the ENA for just this reason and believe that we have demonstrated that collaboration between tools developers and public databases is important in refining best practice in data publication. Journals may generate (or request that authors supply) BlobToolKit assessments of new assemblies submitted for publication, to aid review and speed publication of high quality data.

The third is in comparative and evolutionary genomics. With ongoing improvements in sequencing technologies and assembly software, genome assemblies are improving in quality and contiguity. Among other players, the Earth Biogenome Project (Lewin *et al.* 2018), 10K Vertebrate Genome Project (Genome 10K Community of Scientists 2009) and Tree of Life project (https://www.sanger.ac.uk/science/programmes/tree-of-life) collectively aim to generate chromosomally-contiguous reference genomes for (in the first instance) all known families of Eukaryota. BlobToolKit protocols can be used to explore these genomes for evidence of past horizontal gene transfer, for the presence of symbionts and parasites, and to explore chromosomal patterns of gene expression.

The difficulty we experienced in associating raw sequence read sets with submitted assemblies has led ENA to include a more apparent and thorough explanation of the benefits of and process for referencing reads during eukaryotic genome assembly submission to the repository. We advocate the practice of assembly submission along with associated reads to INSDC to enable downstream analysis and assembly contamination detection.

We aim to complete analysis of all public genomes in INSDC and post them to the BlobToolKit Viewer website at https://blobtoolkit.genomehubs.org/view in the near future, and then maintain currency with the flow of new genomes. The toolkit is under active development (see https://github.com/blobtoolkit) and we welcome feature requests and collaborations to expand and improve its capabilities.

ACKNOWLEDGMENTS

BlobToolKit is based on Blobology by Sujai Kumar and BlobTools by Dominik Laetsch and we thank both, and other colleagues in the Blaxter lab, for their comments and criticisms. This work was funded by a BBSRC Bioinformatics and Biological Resources award BB/P024238/1.

LITERATURE CITED

Altschul, S., 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amid, C., B. T. F. Alako, V. Balavenkataraman Kadhirvelu, T. Burdett, J. Burgin *et al.*, 2019 The European Nucleotide Archive in 2019. Nucleic Acids Res.

Andreson, R., M. Roosaare, L. Kaplinski, S. Laht, T. Kõressaar *et al.*, 2019 Gene content of the fish-hunting cone snail Conus consors. bioRxiv. 590695. https://doi.org/10.1101/590695

Arakawa, K., 2016 No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc. Natl. Acad. Sci. USA 113: E3057. https://doi.org/10.1073/pnas.1602711113

Artamonova, I. I., T. Lappi, L. Zudina, and A. R. Mushegian, 2015 Prokaryotic genes in eukaryotic genome sequences: when to infer horizontal gene transfer and when to suspect an actual microbe. Environ. Microbiol. 17: 2203–2208. https://doi.org/10.1111/1462-2920.12854

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte *et al.*, 2002 Recent segmental duplications in the human genome. Science 297: 1003–1007. https://doi.org/10.1126/science.1072047

Bakker, F. T., D. Lei, J. Yu, S. Mohammadin, Z. Wei *et al.*, 2016 Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. Biol. J. Linn. Soc. Lond. 117: 33–43. https://doi.org/10.1111/bij.12642

Bohlin, J., L. Snipen, S. P. Hardy, A. B. Kristoffersen, K. Lagesen *et al.*, 2010 Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics 11: 464. https://doi.org/10.1186/1471-2164-11-464

Buchfink, B., C. Xie, and D. H. Huson, 2015 Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12: 59–60. https://doi.org/10.1038/nmeth.3176

Challis, R. J., S. Kumar, K. Dasmahapatra, C. D. Jiggins, and M. Blaxter, 2016 Lepbase: the Lepidopteran genome database. bioRxiv 056994. https://doi.org/10.1101/056994

Crotti, E., A. Rizzi, B. Chouaia, I. Ricci, G. Favia *et al.*, 2010 Acetic acid bacteria, newly emerging symbionts of insects. Appl. Environ. Microbiol. 76: 6963–6970. https://doi.org/10.1128/AEM.01336-10

Delmont, T. O., and A. M. Eren, 2016 Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. PeerJ 4: e1839. https://doi.org/10.7717/peerj.1839

Dunning Hotopp, J. C., M. E. Clark, D. C. Oliveira, J. M. Foster, P. Fischer *et al.*, 2007 Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317: 1753–1756. https://doi.org/10.1126/science.1142490

Ekblom, R., and J. B. W. Wolf, 2014 A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. 7: 1026–1042. https://doi.org/10.1111/eva.12178

Eren, A. M., Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison *et al.*, 2015 Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3: e1319. https://doi.org/10.7717/peerj.1319

Florea, L., V. Di Francesco, J. Miller, R. Turner, A. Yao *et al.*, 2005 Gene and alternative splicing annotation with AIR. Genome Res. 15: 54–66. https://doi.org/10.1101/gr.2889405

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret, 2001 GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159: 907–911.

Genome 10K Community of Scientists, 2009 Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100: 659–674.

Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren *et al.*, 2004 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521. https://doi.org/10.1038/nature02426

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde *et al.*, 2014 Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346: 1320–1331. https://doi.org/10.1126/science.1253451

Köster, J., and S. Rahmann, 2018 Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 34: 3600. https://doi.org/10.1093/bioinformatics/bty350

Kumar, S., and M. L. Blaxter, 2011 Simultaneous genome sequencing of symbionts and their hosts. Symbiosis 55: 119–126. https://doi.org/10.1007/s13199-012-0154-6

Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter, 2013 Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front. Genet. 4: 237. https://doi.org/10.3389/fgene.2013.00237

Laetsch, D. R., and M. L. Blaxter, 2017 BlobTools: Interrogation of genome assemblies. F1000 Res. 6: 1287. https://doi.org/10.12688/f1000research.12232.1

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921. https://doi.org/10.1038/35057062

Lewin, H. A., G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington *et al.*, 2018 Earth BioGenome Project: Sequencing life for the future of life. Proc. Natl. Acad. Sci. USA 115: 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589–595. https://doi.org/10.1093/bioinformatics/btp698

Lightfield, J., N. R. Fram, and B. Ely, 2011 Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. PLoS One 6: e17677. https://doi.org/10.1371/journal.pone.0017677

López-García, P., L. Eme, and D. Moreira, 2017 Symbiosis in eukaryotic evolution. J. Theor. Biol. 434: 20–33. https://doi.org/10.1016/j.jtbi.2017.02.031

Merchant, S., D. E. Wood, and S. L. Salzberg, 2014 Unexpected cross-species contamination in genome sequencing projects. PeerJ 2: e675. https://doi.org/10.7717/peerj.675

Morgulis, A., E. M. Gertz, A. A. Schäffer, and R. Agarwala, 2006 WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22: 134–141. https://doi.org/10.1093/bioinformatics/bti774

Mural, R. J., M. D. Adams, E. W. Myers, H. O. Smith, G. L. G. Miklos *et al.*, 2002 A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 296: 1661–1671. https://doi.org/10.1126/science.1069193

Oldeschulte, D. L., Y. A. Halley, M. L. Wilson, E. K. Bhattarai, W. Brashear *et al.*, 2017 Annotated Draft Genome Assemblies for the Northern Bobwhite (*Colinus virginianus*) and the Scaled Quail (*Callipepla squamata*) Reveal Disparate Estimates of Modern Genome Diversity and Historic Effective Population Size. G3 (Bethesda) 7: 3047–3058. https://doi.org/10.1534/g3.117.043083

Pruitt, K. D., T. Tatusova, and D. R. Maglott, 2005 NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33: D501–D504. https://doi.org/10.1093/nar/gki025

Romiguier, J., V. Ranwez, E. J. P. Douzery, and N. Galtier, 2010 Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20: 1001–1009. https://doi.org/10.1101/gr.104372.109

Sackton, T. B., P. Grayson, A. Cloutier, Z. Hu, J. S. Liu *et al.*, 2019 Convergent regulatory evolution and loss of flight in paleognathous birds. Science 364: 74–78. https://doi.org/10.1126/science.aat7244

Salinas, J., G. Matassi, L. M. Montero, and G. Bernardi, 1988 Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res. 16: 4269–4285. https://doi.org/10.1093/nar/16.10.4269

Salter, S. J., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson *et al.*, 2014 Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12: 87. https://doi.org/10.1186/s12915-014-0087-z

Salzberg, S. L., J. C. Dunning Hotopp, A. L. Delcher, M. Pop, D. R. Smith *et al.*, 2005 Serendipitous discovery of Wolbachia genomes in multiple Drosophila species. Genome Biol. 6: R23. https://doi.org/10.1186/gb-2005-6-3-r23

Šmarda, P., P. Bureš, L. Horová, I. J. Leitch, L. Mucina *et al.*, 2014 Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc. Natl. Acad. Sci. USA 111: E4096–E4102. https://doi.org/10.1073/pnas.1321152111

Tomaszkiewicz, M., P. Medvedev, and K. D. Makova, 2017 Y and W Chromosome Assemblies: Approaches and Discoveries. Trends Genet. 33: 266–282. https://doi.org/10.1016/j.tig.2017.01.008

Warren, W. C., L. Kuderna, A. Alexander, J. Catchen, J. G. Pérez-Silva *et al.*, 2017 The Novel Evolution of the Sperm Whale Genome. Genome Biol. Evol. 9: 3260–3264. https://doi.org/10.1093/gbe/evx187

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35: 543–548. https://doi.org/10.1093/molbev/msx319

Werren, J. H., L. Baldo, and M. E. Clark, 2008 Wolbachia: master manipulators of invertebrate biology. Nat. Rev. Microbiol. 6: 741–751. https://doi.org/10.1038/nrmicro1969

Zhou, Q., H.-M. Zhu, Q.-F. Huang, L. Zhao, G.-J. Zhang *et al.*, 2012 Deciphering neo-sex and B chromosome evolution by the draft genome of Drosophila albomicans. BMC Genomics 13: 109. https://doi.org/10.1186/1471-2164-13-109

Communicating editor: B. Andrews