`annonex2embl`: automatic preparation of annotated DNA sequences for bulk submissions to ENA

Michael Gruenstaeudl[1]

[1]**Institut für Biologie, Freie Universität Berlin, 14195 Berlin, Germany**
Email address: m.gruenstaeudl@fu-berlin.de

ABSTRACT

Motivation: The submission of annotated sequence data to public sequence databases constitutes a central pillar in biological research. The surge of novel DNA sequences awaiting database submission due to the application of next-generation sequencing has increased the need for software tools that facilitate bulk submissions. This need has yet to be met with a concurrent development of tools to automate the preparatory work preceding such submissions.

Results: I introduce `annonex2embl`, a Python package that automates the preparation of complete sequence flatfiles for large-scale sequence submissions to the European Nucleotide Archive. The tool enables the conversion of DNA sequence alignments that are co-supplied with sequence annotations and metadata to submission-ready flatfiles. Among other features, the software automatically accounts for length differences among the input sequences while maintaining correct annotations, automatically interlaces metadata to each record, and displays a design suitable for easy integration into bioinformatic workflows. As proof of its utility, `annonex2embl` is employed in preparing a dataset of more than 1,500 fungal DNA sequences for database submission.

INTRODUCTION

The submission of nucleotide sequence data to public sequence databases is a central pillar in the academic effort to share scientific data and make biological research more reproducible (Tenopir et al. 2011; Drew et al. 2013; Blaxter et al. 2016). In fact, submitting newly generated sequences to public repositories has become a mandatory practice for many scientific journals in the biological sciences (Fairbairn 2011; Vines et al. 2013; Roche et al. 2015). Several sequence databases accept nucleotide sequences for public storage, including the DNA Data Bank of Japan (DDBJ, Kodama et al. 2018), the European Nucleotide Archive (ENA, Harrison et al. 2019), and the nucleotide section of the National Center for Biotechnology Information (NCBI GenBank, Benson et al. 2018). These three archives operate under the umbrella of the International Nucleotide Sequence Database Collaboration (INSDC, Karsch-Mizrachi et al. 2018), which coordinates the standardized storage and public access of submitted sequences and has defined common data standards for the sequences and associated metadata (Cochrane et al. 2015).

The almost ubiquitous implementation of next-generation sequencing in biological research has led to a surge of novel DNA sequences from various organisms (Kress et al. 2015; Levy and Myers 2016) and made the generation and analysis of large-scale biological datasets commonplace (Hampton et al. 2013; Farley et al. 2018). Many contemporary molecular phylogenetic and population genetic studies now generate hundreds, if not thousands, of novel nucleotide sequences per investigation (e.g., Leebens-Mack et al. 2019; Li et al. 2019; Varga et al. 2019; Zhao et al. 2019). Accordingly, a considerable fraction of submissions to public databases represents sequence data from large-scale investigations (Kans and Ouellette 2001; Gruenstaeudl and Hartmaring 2019). Data accumulation statistics for nucleotide sequences at ENA corroborate this trend (Silvester et al. 2018; Cook et al. 2019). Most sequence databases have, consequently, adopted policies and methods to enable the submission of large-scale sequence datasets. GenBank, for example, has implemented a submission workflow to facilitate the sequence upload and submission for targeted locus studies containing 2,500 or more ribosomal RNA sequences (Sayers et al.

2019). It also expanded the functionality of its command-line submission tool (tbl2asn, Benson et al. 2006) to accept sequence data in different file formats (Sayers et al. 2019). Similarly, ENA has implemented measures to automate the sequence upload process through a new command-line submission tool (Webin-CLI, Harrison et al. 2019) and, thus, facilitate the streamlined submission of large numbers of annotated DNA sequences via its Webin submission portal (Silvester et al. 2018). However, the development of software tools necessary to format and prepare DNA sequences for large-scale submissions has not kept pace with the methodological advances to upload them. While command-line driven data transfer procedures can facilitate the upload of an almost infinite number of sequences to public sequence databases, the compilation of the necessary input files is a separate step and also requires automation.

Automating the preparation of novel sequence data for submission to public sequence databases is an important objective and particularly relevant in investigations that generate large numbers of similarly structured nucleotide sequences. For example, molecular phylogenetic and population genetic studies analyze differences in nucleotide sequence data among closely-related taxa (Yang and Rannala 2012; Casillas and Barbadilla 2017) and typically generate DNA sequence alignments (Morrison 2006). Such alignments represent ideal starting points for the automated preparation of submission input files. At that stage, the positional homology among the aligned sequences allows the bulk assignment of functional annotations across all sequences of a dataset, regardless of sequence number or length. Moreover, all sequences of a molecular phylogenetic or population genetic dataset must pass through the process of multiple sequence alignment (MSA) to establish positional homology (Morrison et al. 2015), rendering this stage optimal for the interlacing of sequence metadata. Preparing sequence submissions at the stage of MSA, thus, has important advantages compared to other submission preparation strategies. However, few, if any, software tools can facilitate the automated preparation of aligned and annotated DNA sequences for submission to public sequence databases. To the best of my knowledge, no stand-alone software tool currently exists that can automatically convert a large set of annotated DNA sequences and associated metadata into a file format compatible with the upload requirements of ENA. Bulk sequence submissions to ENA primarily operate with data files formatted according to the EMBL flatfile standard (Silvester et al. 2018); EMBL-formatted flatfiles are compact, yet human-readable, and contain the nucleotide sequences of an organism, functional annotations, and associated metadata (Stoesser et al. 2002). Similarly, no stand-alone file conversion tool currently exists that can generate EMBL formatted flatfiles and simultaneously be integrated into automated bioinformatic pipelines; the software tools available are primarily driven by graphical user interfaces (e.g., Artemis, Rutherford et al. 2000; DnaSP, Rozas et al. 2017) and, thus, mandate researchers to edit sequences by hand. Evidently, a software tool is needed that streamlines and automates the preparatory steps for the submission of large datasets of aligned and annotated DNA sequences to ENA. Specifically, the scientific community would benefit from an open-source, command-line driven software tool that automates the interlacing and conversion of MSAs, sequence annotations, and sequence metadata into submission-ready EMBL flatfiles.

In the present investigation, I introduce a software tool that automates the generation of submission-ready EMBL flatfiles from annotated DNA sequence alignments and associated metadata. The tool, titled `annonex2embl`, parses, splits and re-formats DNA sequences and their associated annotations from an MSA, supplements each sequence with the correct metadata of a co-supplied spreadsheet, and saves each sequence in the EMBL flatfile format. The software contains a series of features for the interlacing of nucleotide sequences, sequence annotations, and sequence metadata, and streamlines the process of preparing DNA sequences from molecular phylogenetic or population genetic investigations for submission to ENA. With the application of `annonex2embl`, a user can automatically generate a multi-record flatfile suitable for the bulk submission of large numbers of annotated DNA sequences without the need for manual data editing. To demonstrate the utility of `annonex2embl`, I employ the software to automatically prepare a dataset of more than 1,500 fungal DNA sequences for submission to ENA.

MATERIALS AND METHODS

Overall design

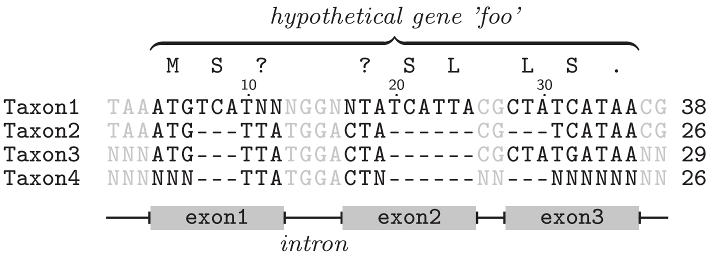
`annonex2embl` was written in Python (Python Software Foundation 2019) and is compatible with both Python versions 2 and 3. The software reads aligned DNA sequences and a set of sequence annotations from an MSA, splits the MSA into individual sequences, adjusts the annotations to the length of the individual sequences, connects each sequence to the correct metadata co-supplied in a data spreadsheet, and converts each sequence plus its associated annotations and metadata into a sequence record of an EMBL-formatted multi-record output file. The software is exclusively controlled via command-line parameters to facilitate its integration into automated bioinformatic pipelines, but also maintains full functionality as a stand-alone tool. The internal integrity of the software is confirmed via a series of unit tests (Pajankar 2017) which evaluate the proper construction of sequence records, feature tables and gene annotations, the length equilibration of individual sequence records, the automated search of gene product names, and the parsing and transfer of metadata to the sequence records. A display of the internal design of `annonex2embl` is given in Supplementary Figure S1.

Input and output

To execute `annonex2embl`, a minimum of six types of input information must be supplied by the user: (i) an MSA containing two or more aligned DNA sequences, (ii) annotations for the aligned sequences, (iii) a data table containing metadata for the sequences, (iv) a one-line description for the genomic region represented by the sequences, (v) the name of the output flatfile, and (vi) the name and email address of the person preparing the sequence submission. The MSA and the sequence annotations must be supplied as different blocks within a single NEXUS file, which can store sequence alignments, character sets, and phylogenetic trees as separate blocks of information (Maddison et al. 1997). Specifically, the MSA must be supplied as a 'data' block, the sequence annotations as an accompanying 'sets' block of the same NEXUS file. Multiple independent character sets can be specified per sets block, each of which defines the title, type, and position of an annotation feature across all sequences of the MSA (see the 'sets' block of 'input 1 of 2' in Figures 1 and 2). Every character set title must follow a specific format: it must encode - in that order - the name, the type and, optionally, the reading direction of the annotation feature, each separated by an underscore. The specification of the feature type is restricted to the vocabulary of feature keys defined by the INSDC (`http://www.insdc.org/files/feature_table.html#7.2`; accessed on 22-Sep-2019). The specification of the reading direction is limited to the keywords 'forward' and 'reverse', with the 5' to 3' direction ('forward') representing the default value. For example, the hypothetical coding sequence 'foo' with a forward reading direction would be represented by either character set definition `foo_CDS` or `foo_CDS_forward` (see the first character set of the 'sets' block of 'input 1 of 2' in Figures 1 and 2). Different character sets may overlap or even be identical in position with other character sets, and may be of any length ≥ 1. The metadata table must be supplied as a comma-delimited spreadsheet and contain a minimum of two data columns ('input 2 of 2' in Figure 1). The first column (titled 'isolate' by default) connects the individual rows of the metadata table to the DNA sequences of the MSA; its values must be identical to the sequence names of the MSA. The second column (titled 'organism' by default) contains the taxon names of the organisms represented by the sequences; its values must be identical to scientific taxon names registered at the NCBI taxonomy database (Federhen 2012). Next to these mandatory columns, the metadata table may comprise any number of additional data columns as long as the column titles are limited to the vocabulary of source feature qualifiers defined by the INSDC. The description of the genomic region represented by the MSA must be supplied as a short alphanumeric character string and characterize the entire genomic region delineated by the sequences (e.g., 'gene1 complete sequence, and gene2, partial sequence'). `annonex2embl` adds this description as part of the official description line to each sequence record of the resulting flatfile. The name and email address of the person preparing the sequence submission must be supplied as separate character strings and are necessary for both, their inclusion in the resulting flatfile as well as to enable optional features of `annonex2embl` such as the online gene product lookup. Next to this mandatory input, several optional input parameters can be specified to control specific processes during software execution.

Based on the aggregate of all input information, `annonex2embl` reads and parses the aligned DNA sequences and their annotations from the NEXUS file, as well as the sequence metadata from the comma-delimited spreadsheet, interlaces annotations, metadata, submitter info and general sequence description with the individual sequences, and saves of each DNA sequence plus its associated annotations and metadata one by one in the EMBL flatfile format. The output of `annonex2embl` is a multi-record EMBL flatfile as well as an accompanying manifest file. The flatfile comprises a set of individual sequence records whose number should be equal to the number of DNA sequences in the input MSA. If the number

INPUT 1 of 2: `TestData1.nex`



```
Taxon1  TAAATGTCATNNNGGNNTATCATTACGCTATCATAACG 38
Taxon2  TAAATG---TTATGGACTA------CG---TCATAACG 26
Taxon3  NNNATG---TTATGGACTA------CGCTATGATAANN 29
Taxon4  NNNNNN---TTATGGACTN------NN---NNNNNNNN 26
```

```
BEGIN SETS;
CHARSET foo_CDS = 4-12 17-25 28-36;
CHARSET foo_gene = 4-12 17-25 28-36;
END;
```

INPUT 2 of 2: `Metadata.csv`

isolate	*organism*	*country*	*specimen_voucher*
Taxon1	Arabidopsis thaliana		Museum1: Collection#
Taxon2	Nymphaea odorata	"Country: State, City"	Herbarium2: Lörrach9
Taxon3	Taxon three		
Taxon4	Betula pubescens	Iceland: Reykjavík	

OUTPUT: `TestData1.embl`

Taxon1 *[internal Ns]*

```
FT   source  1..38
FT           /isolate="Taxon1"
FT           /organism="Arabidopsis thaliana"
FT           /specimen_voucher="Museum1: Collection#"
FT           /mol_type="genomic DNA"
FT   CDS     join(4..12,17..25,28..36)
FT           /note="foo"
FT           /transl_table=11
FT           /translation="MSXXSLLS"
FT   gene    join(4..12,17..25,28..36)
FT           /note="foo"
FT   gap     11..13
FT           /estimated_length=3
FT   gap     16..17
FT           /estimated_length=2
```

Taxon2 *[deletions]*

```
FT   source  1..26
FT           /isolate="Taxon2"
FT           /organism="Nymphaea odorata"
FT           /country="Country: State, City"
FT           /specimen_voucher="Herbarium2: Lorrach9"
FT           /mol_type="genomic DNA"
FT   CDS     join(4..9,14..16,19..24)
FT           /note="foo"
FT           /transl_table=11
FT           /translation="MLLS"
FT   gene    join(4..9,14..16,19..24)
FT           /note="foo"
```

Taxon3 *[deletions; mutations (internal stop codon at end); terminal Ns]*

```
FT   source  1..24
FT           /isolate="Taxon3"
FT           /organism="Taxon three"
FT           /mol_type="genomic DNA"
FT   CDS     join(1..6,11..13,16..21)
FT           /note="foo"
FT           /transl_table=11
FT           /translation="MLLL"
FT   gene    join(1..6,11..13,16..21)
FT           /note="foo"
```

Figure 1. Schematic example of the input passed to, and the output received from, `annonex2embl` under test dataset 1. Only the matrix of the data block in the NEXUS input file is visualized. The country value for taxon 2 in the metadata input table is enclosed by double quotation marks because it contains a comma. Taxon 4 is not saved to the output by `annonex2embl` because the number of unambiguous nucleotides in its sequence is smaller than the minimum number required by the automated validation conducted by Webin. The sequence translation above the MSA refers to the sequence of taxon 1, is not part of the actual input, and provided to assist readers in relating the input to the output information.

of sequence records is lower than the number of sequences in the MSA, not all of the aligned sequences were processed successfully. annonex2embl informs the user of any issues encountered but does not necessarily terminate the overall processing of the input MSA. This design allows the rapid and efficient conversion of a large number of aligned DNA sequences, even if some of them exhibit incorrect specifications and require a separate treatment. Moreover, annonex2embl automatically skips the processing of all input sequences with a length of ten or fewer unambiguous nucleotides (e.g., taxon 4 in Figure 1), as such sequences do not pass the automated validation procedure of the Webin submission portal. The manifest file specifies the name, location, and associated study number of the flatfile and is required for the command-line driven submission of the flatfile via Webin-CLI. A set of example input and output files that illustrate the conversion process conducted by annonex2embl is co-supplied with the package (folder 'examples') and portrayed as Figures 1 and 2.

All input files required for the execution of annonex2embl can be generated via standard biological software applications. The NEXUS file, for example, can be generated by various user-friendly sequence alignment editors (e.g., PhyDE, Müller et al. 2010) as well as via the application of biological script libraries (e.g., BioPython, Cock et al. 2009). The metadata table can be generated by any common spreadsheet editor. Users can compile a wide variety of information in the metadata table for all or only some of the DNA sequences of the MSA as long as the sequence IDs and the organism names are present for all entries. Placeholders for empty table cells are unnecessary, as annonex2embl automatically removes vacant cells upon sequence processing.

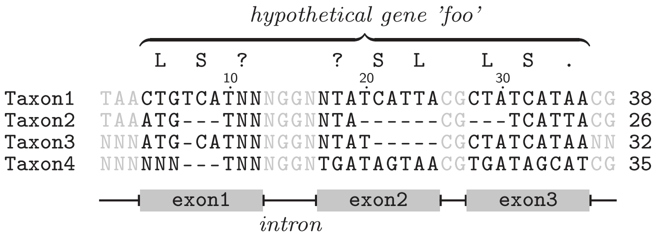
Adherence to INSDC conventions

annonex2embl strictly adheres to the naming and formatting conventions defined by the INSDC for feature keys and their qualifiers. The INSDC has defined a list of 52 feature keys and 102 associated qualifiers to represent functional annotations in DDBJ, EMBL and GenBank sequence flatfiles (feature table definition version 10.8, Karsch-Mizrachi et al. 2018). Any EMBL-formatted flatfile intended for submission to ENA must adhere to this terminology and the associated formatting rules to pass the automated validation procedure of the Webin submission portal (Gibson et al. 2015). By extension, any software tool that aims to convert DNA sequences to EMBL flatfiles must also enforce these conventions. Without adherence to these rules, users would need to post-process the resulting flatfiles to ensure compatibility with the EMBL flatfile standard, which can be prohibitively time-expensive. annonex2embl enforces the INSDC conventions by checking if all column names of the metadata table abide by the controlled vocabulary and, thus, match permissible qualifier names. For example, metadata information on the collection date and the collection locality of a sequenced organism must be named after, and formatted according to, the INSDC source feature qualifiers 'collection_date' and 'country', respectively. Users are, thus, compelled to employ the INSDC-approved terminology and its formatting rules when compiling the input data. Specifically, the character set names of the sequence annotations, as well as the column names of the metadata table, must conform to the approved terminology, and the content of the metadata table must adhere to the formatting rules. Similarly, annonex2embl enforces the correct character encoding for all metadata content. The INSDC has specified that feature qualifier values may only be composed of printable ASCII characters. Consequently, annonex2embl automatically replaces all non-ASCII characters in the metadata with ASCII-compatible characters before integrating the character strings as qualifier values in the feature table (see, for example, specimen voucher information of taxon 2 of 'input 2 of 2' in Figure 1 and compare it with the corresponding feature table line in the output section of the same figure). Likewise, annonex2embl transforms all user-supplied sequence annotations to appropriate feature table entries by converting the character set definitions to INSDC-approved feature keys with fitting qualifiers. Taken together, the strict adherence by annonex2embl to the naming and format conventions of the INSDC streamlines the generation of EMBL flatfiles without the need for manual post-processing.

Installation and usage

annonex2embl is accessible via the Python package index under http://pypi.python.org/pypi/annonex2embl and can be installed using the command `pip install annonex2embl`. It has been successfully tested on Linux (Arch Linux 4.18, Debian 9.9, and Ubuntu 16.04), MacOS (macOS 10.13), and Windows (Windows Server version 1803) under both Python 2.7 and 3.7.

INPUT 1 of 2: `TestData2.nex`



```
Taxon1  TAACTGTCATNNNGGNNTATCATTACGCTATCATAACG 38
Taxon2  TAAATG---TNNNGGNNTA------CG---TCATTACG 26
Taxon3  NNNATG-CATNNNGGNNTAT-----CGCTATCATAANN 32
Taxon4  NNNNNN---TNNNGGNTGATAGTAACGTGATAGCATCG 35
```

```
BEGIN SETS;
CHARSET foo_CDS_forward = 4-12 17-25 28-36;
CHARSET bar_intron = 13-16;
CHARSET baz_CDS_reverse = 20-25 28-36;
END;
```

OUTPUT: `TestData2.embl`

Taxon1 *[mutations; no start codon]*

```
FT   source   1..38
FT            /isolate="Taxon1"
FT            /organism="Arabidopsis thaliana"
FT            /specimen_voucher="Museum1: Collection#"
FT   CDS      join(<4..12,17..25,28..36)
FT            /note="foo"
FT            /transl_table=11
FT            /codon_start=1
FT            /translation="LSXXSLLS"
FT   gap      11..13
FT            /estimated_length=3
FT   intron   13..16
FT            /note="bar"
FT   gap      16..17
FT            /estimated_length=2
```

Taxon2 *[deletions; mutations; no stop codon]*

```
FT   source   1..26
FT            /isolate="Taxon2"
FT            /organism="Nymphaea odorata"
FT            /country="Country: State, City"
FT            /specimen_voucher="Herbarium2: Lorrach9"
FT   CDS      join(4..9,14..16,19..>24)
FT            /note="foo"
FT            /transl_table=11
FT            /codon_start=1
FT            /translation="MXXSL"
FT   gap      8..10
FT            /estimated_length=3
FT   intron   10..13
FT            /note="bar"
FT   gap      13..14
FT            /estimated_length=2
```

Taxon3 *[deletions; mutations (internal stop codon at start); terminal Ns]*

```
FT   source   1..27
FT            /isolate="Taxon3"
FT            /organism="Taxon three"
FT   CDS      join(1..8,13..16,19..27)
FT            /note="foo"
FT            /transl_table=11
FT            /translation="MHXYLS"
FT   gap      7..9
FT            /estimated_length=3
FT   intron   9..12
FT            /note="bar"
FT   gap      12..13
FT            /estimated_length=2
```

Taxon4 *[reverse gene]*

```
FT   source   1..29
FT            /isolate="Taxon4"
FT            /organism="Betula pubescens"
FT            /country="Iceland: Reykjavik"
FT   gap      2..4
FT            /estimated_length=3
FT   intron   4..7
FT            /note="bar"
FT   gap      7
FT            /estimated_length=1
FT   CDS      complement(join(19..27,<11..16))
FT            /note="baz"
FT            /transl_table=11
FT            /codon_start=1
FT            /translation="MLSLL"
```

Figure 2. Schematic example of the input passed to, and the output received from, `annonex2embl` under test dataset 2. Input 2 of 2 is the same as in Figure 1 and, thus, not displayed. To conserve space, the source qualifier `/mol_type="genomic DNA"` was removed from each output feature table.

The usage of `annonex2embl` is exclusively controlled via command-line parameters. At a minimum, a user must enter six input parameters: the name of, and file path to, a NEXUS file (command-line parameter `-n`); the name of, and file path to, a comma-delimited metadata table (`-c`); a quotation-mark enclosed text string that describes the genomic region represented by the DNA sequences (`-d`); the name of, and file path to, the output flatfile (`-o`); and the name (`-a`) and email address (`-e`) of the person preparing the sequence submission, both as quotation-mark enclosed text strings. In addition to these mandatory parameters, users may invoke up to eleven optional parameters. For example, users may wish to have `annonex2embl` test if each taxon name supplied via the metadata has been registered as a scientific taxon name with the NCBI taxonomy database (command-line parameter `--taxonomycheck`). ENA does not accept the submission of DNA sequences representing taxa that have not been registered with the NCBI taxonomy database, rendering an *a priori* evaluation of their registration critical. Similarly, users may wish to have the software automatically search and add standardized gene product names to the feature tables of the output flatfile (`--productlookup`), which is important for a complete representation of the DNA sequences (Pirovano et al. 2017). When selecting this option, `annonex2embl` parses the gene abbreviations from the character set definition titles of the sequence annotations, communicates with NCBI to identify the complete gene names, and adds appropriate product qualifiers to the relevant feature keys. For example, the gene abbreviation 'matK' would be identified as, and linked to, the complete gene name 'maturase K', which, in turn, would be added to the qualifier 'product' of the feature key 'gene'. The full set of available input parameters, their default values, and a short explanation of each parameter can be displayed by invoking the help command via `annonex2embl -h`.

To invoke `annonex2embl` on one of the example input files co-supplied with the package, a user may type the following command in a terminal:

```
annonex2embl \
  -n examples/input/TestData1.nex \
  -c examples/input/Metadata.csv \
  -o examples/output/TestData1.embl \
  -d "description of alignment here" \
  -a "your name here" \
  -e "your_email_here@yourmailserver.com"
```

To communicate any issues encountered while processing the input data, `annonex2embl` prints short warning or error messages to the screen. Warning messages indicate the skipping of individual sequence records due to issues restricted to these records, whereas error messages indicate the termination of the software prior to processing the last record of a file. For example, `annonex2embl` would provide a warning message if an annotation feature of type 'CDS' was not added to the feature table of a sequence record if the underlying amino acid sequence contained a stop codon immediately after the indicated start codon (e.g., hypothetical coding sequence 'baz' in taxa 1, 2, and 3 in Figure 2). Similarly, the software would provide a warning message if a sequence record was not saved to the output flatfile if an organism name was not encountered in the NCBI taxonomy database (e.g., taxon 3 in Figures 1 and 2), assuming that the optional parameter for taxonomy checks had been selected.

Upon execution, `annonex2embl` generates a submission-ready EMBL-formatted flatfile as well as an accompanying manifest file, which is specifically required for command-line driven sequence submissions to ENA. `annonex2embl` does not, however, conduct the sequence submission itself. Instead, the submission of the flatfile to ENA remains the responsibility of the user. This design decision was taken for two reasons: First, with Webin-CLI, ENA offers an official command-line software tool for the upload and submission of annotated DNA sequences via the Webin submission portal, whose usage shall be encouraged. Second, users should always validate the accuracy and content of the flatfiles before a submission via the proper validation tools (Harrison et al. 2019), such as Webin-CLI or the stand-alone ENA flatfile validator (Gibson et al. 2015). Upon submission of a flatfile to ENA, the user receives a list of unique accession numbers from ENA that permanently identify the sequences.

Application on empirical data

To demonstrate its utility, `annonex2embl` is employed to prepare a large-scale dataset of annotated fungal DNA sequences for submission to ENA. Specifically, the software is used to automatically convert an annotated MSA in NEXUS format comprising a total of 1,518 individual DNA sequences, and a corresponding metadata table, into a submission-ready EMBL-formatted flatfile. The DNA sequences were collected in an investigation on fungal molecular diversity of the Canarian archipelago (Gruenstaeudl et al. 2013) and have not been submitted to any sequence database. Upon conversion to an EMBL flatfile, the sequences are first validated and then uploaded to ENA, both via Webin-CLI v.1.8.11. The run-times of both `annonex2embl` and Webin-CLI are measured and compared. The input files for this test submission are available as references for format and content from Zenodo at `https://zenodo.org/record/3517124`.

RESULTS

Using `annonex2embl` to prepare a total of 1,518 aligned DNA sequences for submission to ENA was found to be rapid and reliable. Specifically, the software required 1.4 minutes for converting the annotated MSA and the co-supplied metadata into a submission-ready EMBL-formatted flatfile when only the mandatory parameters for a submission via the Webin portal were employed (Table 1). Thus, the conversion was an order of magnitude faster than the subsequent flatfile validation or the subsequent flatfile upload to ENA, both of which were conducted via Webin-CLI and required 15.9 minutes and 19.4 minutes, respectively. No data processing other than the invocation of `annonex2embl` was required before validating and uploading the flatfile, and not a single sequence record required post-processing, illustrating that `annonex2embl` can prepare a large number of annotated DNA sequences for submission to ENA in an automatic and streamlined fashion. The submitted fungal DNA sequences are available on ENA under accession numbers LR730418-LR731935.

Moreover, `annonex2embl` has been employed by several molecular phylogenetic investigations to facilitate their sequence submissions to ENA. Specifically, the software has so far been used in Roy et al. (2017), Korotkova et al. (2018), Canal et al. (2018), Borsch et al. (2018), and Falcon-Hidalgo et al. (2019) to submit DNA sequence data via the Webin submission portal.

Step	Overview of step	Details of step	Software employed	Command-line instruction	Time required
1	Flatfile generation	Conversion and interlacing of the aligned and annotated DNA sequences (in a single NEXUS file) with corresponding sequence metadata (in a single spreadsheet) into an EMBL flatfile	annonex2embl	`annonex2embl -n msa_plus_annotations.nex -c metadata.csv -d "28S large subunit ribosomal RNA gene, partial sequence" -e your_email_here@yourmailserver.com -a "Your name here" -o flatfile.embl --manifeststudy your_ENA_project_number_here --manifestname 28S_large_subunit_ribosomal_RNA --compress`	1.40 min
2	Flatfile validation	Automated validation to confirm flatfile format and content, including checks to confirm the adherence to INSDC formatting and naming conventions as well as the registration of scientific taxon names	Webin-CLI	`java -jar webin-cli-1.8.11.jar -context=sequence -manifest=flatfile.manifest -username your_Webin_ID_number_here -password your_Webin_password_here -validate`	15.93 min
3	Flatfile upload	Upload of the flatfile to ENA via the Webin submission portal, including a server-side validation of the submission data	Webin-CLI	`java -jar webin-cli-1.8.11.jar -context=sequence -manifest=flatfile.manifest -username your_Webin_ID_number_here -password your_Webin_password_here -submit`	19.43 min

Table 1. Overview of, the time required for, and the command-line instructions used in, the bioinformatic steps involved in the bulk preparation, validation, and submission of 1,518 fungal DNA sequences to ENA via `annonex2embl` and the command-line route of the Webin submission portal. Command-line instructions refer to usage on the Bash shell under Linux.

DISCUSSION

Several routes for preparing and submitting annotated DNA sequences to ENA exist. First, researchers can prepare annotated DNA sequences for submission as pre-formatted data spreadsheets (so-called 'checklists'), which are manually filled by the user with nucleotide sequences and associated metadata and then uploaded to ENA via the interactive route of the Webin submission portal (Silvester et al. 2018). Idiosyncrasies of the genomic regions necessitate the application of different checklist types, rendering the submission of diverse sets of sequences time- and labor-expensive. However, a recently developed software tool enables the automatic conversion of DNA sequences to some of the more commonly applied checklist types (Gruenstaeudl and Hartmaring 2019) and may, thus, be an option for small-scale submissions. Second, researchers can prepare annotated DNA sequences for submission as EMBL-formatted flatfiles, which can be uploaded to ENA via the command-line route (Harrison et al. 2019) or the programmatic route (Silvester et al. 2018) of the Webin submission portal. Programmatic submissions require that input flatfiles are accompanied by XML authentication and submission files which are difficult to generate for nonexperts. Hence, the command-line driven route of constructing, validating, and uploading EMBL flatfiles currently represents the most efficient and user-friendly method for large-scale sequence submissions to ENA.

Only few software applications exist that can generate EMBL-formatted flatfiles. Most of these tools operate only on individual sequences or require additional file conversions for their output to conform to the EMBL flatfile standard, rendering them unsuitable for automated, large-scale submission preparations. For example, the open-source genome viewer Artemis (Rutherford et al. 2000) can save annotated DNA sequences as flatfiles that abide by the EMBL flatfile format, but can only operate on one sequence at a time. Moreover, functional annotations must be added for each sequence individually in this viewer, rendering sequence submissions prohibitively time-expensive for all but the smallest datasets. Similarly, the commercial software suite Geneious (Kearse et al. 2012) can save DNA sequences in the GenBank flatfile format, which must be further converted to the EMBL flatfile format via additional software tools (e.g., EMBOSS Seqret, Olson 2002). While Geneious allows the semi-automatic propagation of annotations across aligned sequences, it is primarily driven via a graphical user interface, effectively precluding its integration into automated bioinformatic pipelines. Moreover, the conversion route enabled by Geneious is dependent on the continued compatibility of multiple, independent software tools and requires users to afford the licensing cost for the software suite. Similar functionality as provided by Geneious in conjunction with file converters can be achieved via the legacy NCBI tool Sequin (Benson et al. 2013), which can also transfer annotations across DNA sequences but, in like manner, requires additional conversion steps to generate EMBL-formatted flatfiles. Next to these stand-alone software tools, several online services offer the automatic submission of DNA sequences to ENA; however, the functionality of these web services is difficult to assess, as many of them are inaccessible in their original form (e.g., CDinFusion, Hankeln et al. 2011, attempt to access on 24-Oct-2019) or have not made their source code and, thus, the details of their underlying data processing publicly available. `annonex2embl`, by contrast, represents an open-source software tool that is transparent in its underlying processes, easily customizable, and available to users without charge. It enables the automated, command-line driven processing of large-scale sequence datasets directly into EMBL-formatted flatfiles and can be easily integrated into automated bioinformatic pipelines.

With the application of `annonex2embl`, users can prepare bulk submissions of DNA sequences without the need for additional data processing, particularly regarding the assignment of sequence annotations. Previously, researchers had to invest considerable time to manually prepare their sequence data for submission to public sequence databases (Pirovano et al. 2017; Gruenstaeudl and Hartmaring 2019). By employing `annonex2embl`, users can harness the advantages inherent to aligned sequences for their automated, streamlined conversion to records of a submission-ready flatfile because aligned sequences are uniform in length, making them amenable to automatic processing. For example, by defining annotations on aligned sequences, the annotations do not need to be specified one sequence at a time, but can be simultaneously assigned across all input sequences. Specifically, the annotations can be propagated to each sequence of an MSA due to the positional homology among them. Such a bulk assignment provides greater consistency between annotations and is an efficient means for preparing large-scale submissions. However, the bulk assignment of annotations on aligned sequences is only possible as long as the implicit length differences among the individual sequences are accommodated. MSAs typically comprise sequences of various length, with spaces inserted within sequences to enable the alignment of homologous nucleotides (Morrison et al. 2015). Accounting for these length differences while maintaining correct annotation assignments is a complex process and becomes especially intricate when the differences are the result of overlapping but non-identical insertions and deletions. annonex2embl maintains correct annotation assignments while accounting for implicit length differences among the sequences by employing a process that automatically transfers all length changes of the sequences to the location positions of their annotation features, thus keeping them in sync.

Another advantage of using MSAs for preparing bulk submissions of DNA sequences to public databases is the streamlined addition of metadata to the individual sequences. Sequence records in public databases should contain as much metadata information as possible, allowing the cross-linking of the submitted data with, and its re-usability by, other analyses and, thus, enhancing its value to other researchers (Meyer et al. 2019). In particular, a detailed description of the identity and location of the organism from which a sequence was generated, references to biological collections that preserve its voucher specimens, and bibliographic information that link the sequence to the published description of the dataset are essential for the scientific re-use of these sequences (Kans and Ouellette 2001). Preparing large-scale sequence sets for submission to public sequence databases should, consequently, include an automated interlacing of metadata to each sequence record, which annonex2embl implements via a sequence-specific data integration from a co-supplied metadata table.

In summary, annonex2embl enables the automated preparation of annotated DNA sequences plus associated metadata for bulk submission to the online sequence archive ENA and will likely accelerate the flow of novel sequence data to this archive, particularly for sequences from phylogenetic and population genetic investigations.

Acknowledgments

The author thanks Gabriele Dröge of the Botanischer Garten und Botanisches Museum Berlin for valuable discussions as well as Yannick Hartmaring and Nils Jenke of the Freie Universität Berlin for assistance with testing and debugging the software.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 418670221 – and by a start-up grant of the Freie Universität Berlin (Initiativmittel der Forschungskommission), both to MG.

Conflict of Interest: none declared

REFERENCES

Benson, D., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Sayers, E. (2013). GenBank. *Nucleic Acids Research*, 41:D36–D42.

Benson, D., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K., and Sayers, E. (2018). GenBank. *Nucleic Acids Research*, 46:D41–D47.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. (2006). GenBank. *Nucleic Acids Research*, 34:D16–D20.

Blaxter, M., Danchin, A., Fukami-Kobayashi, K., Sugano, S., Richard, J., Salzberg, S., Goushcha, A., Galina, Z., Kotsyuba, O., and Dmitrii, F. (2016). Reminder to deposit DNA sequences. *Science*, 352:780.

Borsch, T., Flores-Olvera, H., Zumaya, S., and Müller, K. (2018). Pollen characters and DNA sequence data converge on a monophyletic genus *Iresine* (Amaranthaceae, Caryophyllales) and help to elucidate its species diversity. *Taxon*, 67:944–976.

Canal, D., Koester, N., Jones, K., Korotkova, N., Croat, T., and Borsch, T. (2018). Phylogeny and diversification history of the large neotropical genus *Philodendron* (Araceae): Accelerated speciation in a lineage dominated by epiphytes. *American Journal of Botany*, 105:1035–1052.

Casillas, S. and Barbadilla, A. (2017). Molecular population genetics. *Genetics*, 205:1003–1035.

Cochrane, G., Karsch-Mizrachi, I., and Takagi, T. (2015). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 44:D48–D50.

Cock, P., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and De Hoon, M. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422–1423.

Cook, C., Lopez, R., Stroe, O., Cochrane, G., Brooksbank, C., Birney, E., and Apweiler, R. (2019). The European Bioinformatics Institute in 2018: tools, infrastructure and training. *Nucleic Acids Research*, 47:D15–D22.

Drew, B., Gazis, R., Cabezas, P., Swithers, K., Deng, J., Rodriguez, R., Katz, L., Crandall, K., Hibbett, D., and Soltis, D. (2013). Lost branches on the Tree of Life. *PLOS Biology*, 11:e1001636.

Fairbairn, D. (2011). The advent of mandatory data archiving. *Evolution*, 65:1–2.

Falcon-Hidalgo, B., Fuentes-Bazan, S., Berazain-Iturralde, R., and Borsch, T. (2019). Phylogenetic relationships and character evolution in neotropical *Phyllanthus* (Phyllanthaceae), with a focus on the Cuban and Caribbean taxa. *International Journal of Plant Sciences*, in press.

Farley, S., Dawson, A., Goring, S., and Williams, J. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, 68:563–576.

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, 40:D136–D143.

Gibson, R., Alako, B., Amid, C., Cerdeno-Tarraga, A., Cleland, I., Goodgame, N., Ten Hoopen, P., Jayathilaka, S., Kay, S., Leinonen, R., Liu, X., Pallreddy, S., Pakseresht, N., Rajan, J., Rossello, M., Silvester, N., Smirnov, D., Toribio, A., Vaughan, D., and Cochrane, G. (2015). Biocuration of functional annotation at the European Nucleotide Archive. *Nucleic Acids Research*, 44:D58–D66.

Gruenstaeudl, M. and Hartmaring, Y. (2019). EMBL2checklists: A Python package to facilitate the user-friendly submission of plant and fungal DNA barcoding sequences to ENA. *PLOS ONE*, 14:e0210347.

Gruenstaeudl, M., Santos-Guerra, A., Hawkes, C., and Jansen, R. (2013). Molecular survey of arbuscular mycorrhizal fungi associated with *Tolpis* on three Canarian islands (Asteraceae). *Vieraea*, 41:233–252.

Hampton, S., Strasser, C., Tewksbury, J., Gram, W., Budden, A., Batcheller, A., Duke, C., and Porter, J. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11:156–162.

Hankeln, W., Wendel, N., Gerken, J., Waldmann, J., and Buttigieg, P. (2011). CDinFusion - Submission-ready, on-line integration of sequence and contextual data. *PLOS ONE*, 6:e24797.

Harrison, P., Alako, B., Amid, C., Cerdeno-Tarraga, A., Cleland, I., Holt, S., Hussein, A., Jayathilaka, S., Kay, S., Keane, T., Leinonen, R., Liu, X., Martinez-Villacorta, J., Milano, A., Pakseresht, N., Rajan, J., Reddy, K., Richards, E., Rosello, M., Silvester, N., Smirnov, D., Toribio, A.-L., Vijayaraja, S., and Cochrane, G. (2019). The European Nucleotide Archive in 2018. *Nucleic Acids Research*, 47:D84–D88.

Kans, J. and Ouellette, B. (2001). Submitting DNA sequences to the databases. In *Bioinformatics: A practical guide to the analysis of genes and proteins*, chapter 4, pages 65–81. Wiley Online Library, New York, 2nd ed. edition.

Karsch-Mizrachi, I., Takagi, T., and Cochrane, G. (2018). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 46:D48–D51.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28:1647–1649.

Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., and Takagi, T. (2018). DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Research*, 46:D30–D35.

Korotkova, N., Parolly, G., Khachatryan, A., Ghulikyan, L., Sargsyan, H., Akopian, J., Borsch, T., and Gruenstaeudl, M. (2018). Towards resolving the evolutionary history of Caucasian pears (Pyrus, Rosaceae) - Phylogenetic relationships, divergence times and leaf trait evolution. *Journal of Systematics and Evolution*, 56:35–47.

Kress, W. J., Garcia-Robledo, C., Uriarte, M., and Erickson, D. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution*, 30:25–35.

Leebens-Mack, J., Barker, M., Carpenter, E., Deyholos, M., Gitzendanner, M., Graham, S., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S., Soltis, D., Soltis, P., Stevenson, D., Ullrich, K., Wickett, N., DeGironimo, L., Edger, P., Jordon-Thaden, I., Joya, S., Liu, T., Melkonian, B., Miles, N., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J., Augustin, M., Barrett, M., Baucom, R., Beerling, D., Benstein, R., Biffin, E., Brockington, S., Burge, D., Burris, J., Burris, K., Burtet-Sarramegna, V., Caicedo, A., Cannon, S., Cebi, Z., Chang, Y., Chater, C., Cheeseman, J., Chen, T., Clarke, N., Clayton, H., Covshoff, S., Crandall-Stotler, B., Cross, H., DePamphilis, C., Der, J., Determann, R., Dickson, R., Di Stilio, V., Ellis, S., Fast, E., Feja, N., Field, K., Filatov, D., Finnegan, P., Floyd, S., Fogliani, B., Garcia, N., Gateble, G., Godden, G., Goh, Q.-Y., Greiner, S., Harkess, A., Heaney, J., Helliwell, K., Heyduk, K., Hibberd, J., Hodel, R., Hollingsworth, P., Johnson, M., Jost, R., Joyce, B., Kapralov, M., Kazamia, E., Kellogg, E., Koch, M., Von Konrat, M., Konyves, K., Kutchan, T., Lam, V., Larsson, A., Leitch, A., Lentz, R., Li, F.-W., Lowe, A., Ludwig, M., Manos, P., Mavrodiev, E., McCormick, M., McKain, M., McLellan, T., McNeal, J., Miller, R., Nelson, M., Peng, Y., Ralph, P., Real, D., Riggins, C., Ruhsam, M., Sage, R., Sakai, A., Scascitella, M., Schilling, E., Schlosser, E.-M., Sederoff, H., Servick, S., Sessa, E., Shaw, A., Shaw, S., Sigel, E., Skema, C., Smith, A., Smithson, A., Stewart, C., Stinchcombe, J., Szovenyi, P., Tate, J., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.-N., Weller, S., Wenzel, M., Weststrand, S., Westwood, J., Whigham, D., Wu, S., Wulff, A., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M., Pires, J., Rothfels, C., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T., Nguyen, N.-P., Matasci, N., Nelson, D., Sayyari, E., Wafula, E., Walls, R., Warnow, T., An, H., Arrigo, N., Baniaga, A., Galuska, S., Jorgensen, S., Kidder, T., Kong, H., Lu-Irving, P., Marx, H., Qi, X., Reardon, C., Sutherland, B., Tiley, G., Welles, S., Yu, R., Zhan, S., Gramzow, L., Theissen, G., and Wong, G. K.-S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, in press.

Levy, S. and Myers, R. (2016). Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, 17:95–115.

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., Gitzendanner, M., Fritsch, P., Cai, J., Luo, Y., Wang, H., van der Bank, M., Zhang, S.-D., Wang, Q.-F., Wang, J., Zhang, Z.-R., Fu, C.-N., Yang, J., Hollingsworth, P., Chase, M., Soltis, D., Soltis, P., and Li, D.-Z. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants*, 5:461–470.

Maddison, D., Swofford, D., and Maddison, W. (1997). NEXUS: An extensible file format for systematic information. *Systematic Biology*, 46:590–621.

Meyer, F., Bagchi, S., Chaterji, S., Gerlach, W., Grama, A., Harrison, T., Paczian, T., Trimble, W., and Wilke, A. (2019). MG-RAST version 4 - lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Briefings in Bioinformatics*, 20:1151–1159.

Morrison, D. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*, 19:479–539.

Morrison, D., Morgan, M., and Kelchner, S. (2015). Molecular homology and multiple-sequence alignment: an analysis of concepts and practice. *Australian Systematic Botany*, 28:46–62.

Müller, J., Müller, K., Neinhuis, C., and Quandt, D. (2010). PhyDE: Phylogenetic Data Editor. Available from: `http://www.phyde.de/`. Accessed 04-Aug-2019.

Olson, S. (2002). EMBOSS opens up sequence analysis. *Briefings in Bioinformatics*, 3:87–91.

Pajankar, A. (2017). *Python unit test automation - Practical techniques for Python developers and testers*. Apress, New York, USA.

Pirovano, W., Boetzer, M., Derks, M., and Smit, S. (2017). NCBI-compliant genome submissions: Tips and tricks to save time and money. *Briefings in Bioinformatics*, 18:179–182.

Python Software Foundation (2019). Python Language Reference. Available from: `http://www.python.org/`.

Roche, D., Kruuk, L., Lanfear, R., and Binning, S. (2015). Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*, 13:e1002295.

Roy, J., Reichel, R., Bruggemann, N., Hempel, S., and Rillig, M. (2017). Succession of arbuscular mycorrhizal fungi along a 52-year agricultural recultivation chronosequence. *FEMS Microbiology Ecology*, 93:fix102.

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J., Guirao-Rico, S., Librado, P., Ramos-Onsins, S., and Sanchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, 34:3299–3302.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M., and Barrell, B. (2000).

Artemis: Sequence visualization and annotation. *Bioinformatics*, 16:944–945.

Sayers, E., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K., and Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research*, 47:D94–D99.

Silvester, N., Alako, B., Amid, C., Cerdeno-Tarraga, A., Clarke, L., Cleland, I., Harrison, P., Jayathilaka, S., Kay, S., Keane, T., Leinonen, R., Liu, X., Martinez-Villacorta, J., Menchi, M., Reddy, K., Pakseresht, N., Rajan, J., Rossello, M., Smirnov, D., Toribio, A., Vaughan, D., Zalunin, V., and Cochrane, G. (2018). The European Nucleotide Archive in 2017. *Nucleic Acids Research*, 46:D36–D40.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli, M., Tzouvara, K., and Vaughan, R. (2002). The EMBL nucleotide sequence database. *Nucleic Acids Research*, 30:21–26.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., Wu, L., Read, E., Manoff, M., and Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, 6:e21101.

Varga, T., Krizsan, K., Foldi, C., Dima, B., Sanchez-Garcia, M., Sanchez-Ramirez, S., Szollosi, G., Szarkandi, J., Papp, V., Laszlo, A., Andreopoulos, W., Angelini, C., Antonin, V., Barry, K., Bougher, N., Buchanan, P., Buyck, B., Bense, V., Catcheside, P., Chovatia, M., Cooper, J., Dämon, W., Desjardin, D., Finy, P., Geml, J., Haridas, S., Hughes, K., Justo, A., Karasinski, D., Kautmanova, I., Kiss, B., Kocsube, S., Kotiranta, H., LaButti, K., Lechner, B., Liimatainen, K., Lipzen, A., Lukacs, Z., Mihaltcheva, S., Morgado, L., Niskanen, T., Noordeloos, M., Ohm, R., Ortiz-Santana, B., Ovrebo, C., Racz, N., Riley, R., Savchenko, A., Shiryaev, A., Soop, K., Spirin, V., Szebenyi, C., Tomsovsky, M., Tulloss, R., Uehling, J., Grigoriev, I., Vagvolgyi, C., Papp, T., Martin, F., Miettinen, O., Hibbett, D., and Nagy, L. (2019). Megaphylogeny resolves global patterns of mushroom evolution. *Nature Ecology and Evolution*, 3:668–678.

Vines, T., Andrew, R., Bock, D., Franklin, M., Gilbert, K., Kane, N., Moore, J.-S., Moyers, B., Renaut, S., Rennison, D., Veen, T., and Yeaman, S. (2013). Mandated data archiving greatly improves access to research data. *The FASEB Journal*, 27:1304–1308.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13:303–314.

Zhao, Y.-P., Fan, G., Yin, P.-P., Sun, S., Li, N., Hong, X., Hu, G., Zhang, H., Zhang, F.-M., Han, J.-D., Hao, Y.-J., Xu, Q., Yang, X., Xia, W., Chen, W., Lin, H.-Y., Zhang, R., Chen, J., Zheng, X.-M., and Song, G. (2019). Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. *Nature Communications*, 10:4201.