

Extending OBI Herbarium Records to include associated NCBI GenBank sequences

hash://md5/40b93e072ceb31bb9e78078b929f19d8

hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6

Jorrit Poelen

Katelin Pearson

Jenn Yost

2023-07-19

Abstract

Specimen from Natural History Collections are physical repositories of genetic information. Genetic sequences extracted from specimen are stored in genetic sequence databases like the openly accessible GenBank at NCBI, DNA DataBank of Japan, or the European Nucleotide Archive (ENA). While researchers and collection managers make efforts to associate (or link) Natural History Collection records with their derived genetic accession records, extra work is need to make these associations explicit. We describe how a collaboration between a biodiversity informatics expert and collection managers of the Hoover/OBI Herbarium at CalPoly, San Luis Obispo, CA was forged with the aim to extend OBI specimen records to include their associated GenBank records. In addition, we quantify the costs of creating these specimen extensions, and discuss the socio-economic capacity needed to repeat this digital specimen extension process for the hundreds of millions of specimen records available globally today.

Contents

Introduction	2
Example	3
Methods	5
Phase 1. Acquire and Version	6
Phase 2. Propose OBI associated GenBank Records	7
Results	8
Capture GenBank Candidate Records	8
Curatorial Candidate Record Review	9
Adding GenBank Links to Symbiota Records	9
Discussion	14
References	15
Appendix A	15

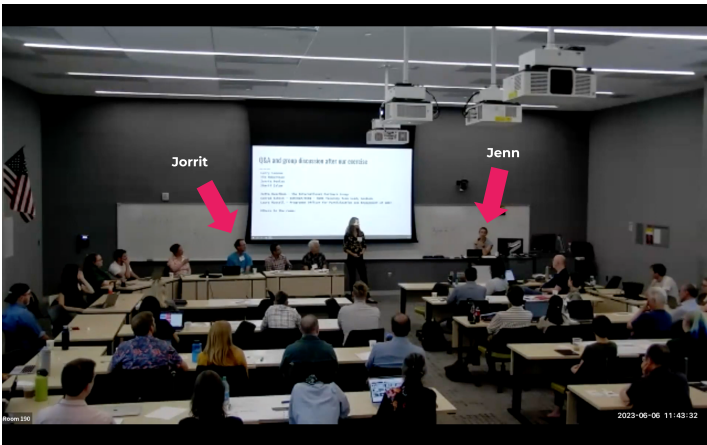
<https://linker.bio/hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6>



Introduction

Billions of biodiversity data records are made openly available by hundreds of Natural History Collections all over the world. Also, since 1982, National Institutes of Health have published versions nucleotide sequences through GenBank. Many specimen described in Natural History Collections have associated GenBank sequence accessions available in GenBank.

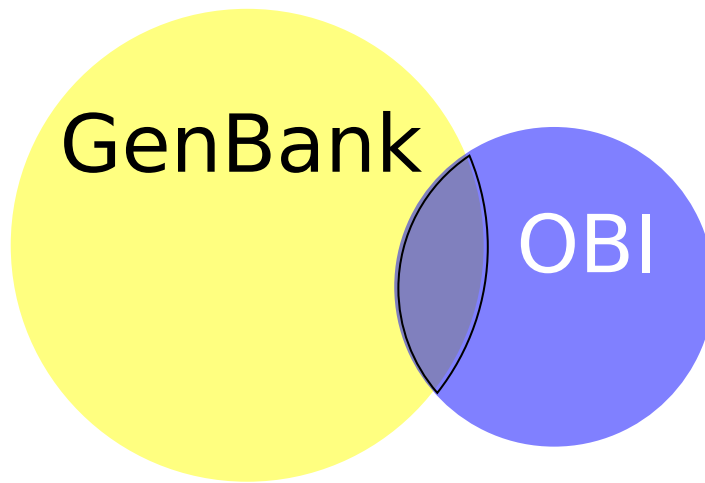
During the 2023 Annual Conference of Digital Data in Biodiversity Research hosted by Arizona State University, Jenn Yost expressed a desire to make it easier to link GenBank accession records to the specimen records the helps curate at the The Hoover Herbarium ({ "http://rs.tdwg.org/dwc/terms/institutionCode": "OBI"}), Cal Poly State University, San Luis Obispo, CA (Yost 2023).



29

Jenn Yost expressing her desire to better Link GenBank records with their associated specimen records (Yost 2023).

This repository is the outcome at a first prototype to help outline a process to discover OBI specimen record references in GenBank. With this, Jenn Yost and collaborators like Kate Pearson can link specimen records to the GenBank accession they are associated with.




Hoover Herbarium (OBI) at Cal Poly State University, San Luis Obispo, CA keeps herbarium specimen. Some of these specimen have associated record in GenBank. These GenBank records extend the OBI specimen additional information such as genetic sequences.

Example

The Hoover Herbarium hosts a preserved specimen of type *Angelica hendersonii* Coult. & Rose that was collected in 1966-07-05 by Tracey & Viola Call at the north end of Tomales Bay and 2 mi south of Tomales in Marin County, California with catalog number: OBI09031, collector number: 2490, occurrence id: 256368e3-f8d7-4028-8010-1a4ff3eb8111, and web reference <https://cch2.org/portal/collections/individual/index.php?occid=166203>.

🔍
🌐
cch2.org/portal/collections/individual/index.php?occid=166203

Details
Genetic
Duplicates
Comments
Linked Resources


OBI - Robert F. Hoover Herbarium, Cal Poly State University (OBI)


Occurrence ID: 256368e3-f8d7-4028-8010-1a4ff3eb8111
Secondary Catalog #: 9031
Taxon: *Angelica hendersonii* J.M. Coult. & Rose
Family: Apiaceae
📄 Show Determination History
Collector: Tracey Call
Number: 2490
Date: 1966-07-05
Verbatim Date: 5-Jul-66
Additional Collectors: Viola Call
Locality: United States, California, Marin, North end of Tomales Bay and 2 mi south of Tomales
Elevation: 15 meters **Verbatim Elevation:** 50ft.
Habitat: Low bluffs
Usage Rights: CC BY-NC (Attribution-Non-Commercial)
Record ID: 9a370197-6899-4072-8b17-4f2f043fbd54

For additional information about this specimen, please contact: Jenn Yost, Director and Associate Professor (jyost@calpoly.edu)

Webpage associated with OBI09031 as seen via <https://cch2.org/portal/collections/individual/index.php?occid=166203> on 2023-09-11.

GenBank hosts a accession record <https://www.ncbi.nlm.nih.gov/nuccore/MT735455>

with locus *Angelica hendersonii* voucher Tracey & V. Call 2490 (OBI09031) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.

	 ncbi.nlm.nih.gov/nucleotide/MT735455
LOCUS	MT735455 599 bp DNA linear PLN 23-MAY-2021
DEFINITION	<i>Angelica hendersonii</i> voucher Tracey & V. Call 2490 (OBI09031) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
ACCESSION	MT735455
VERSION	MT735455.1
KEYWORDS	.
SOURCE	<i>Angelica hendersonii</i>
ORGANISM	Angelica hendersonii Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; asterids; campanulids; Apiales; Apiaceae; Apioideae; apioid superclade; Selineae; <i>Angelica</i> .
REFERENCE	1 (bases 1 to 599)
AUTHORS	Liao,C.-Y., Gao,Q., Katz-Downie,D.S. and Downie,S.R.
TITLE	A systematic study of North American <i>Angelica</i> species (Apiaceae) based on nrDNA ITS and cpDNA sequences and fruit morphology
JOURNAL	J Syst Evol (2021) In press
REMARK	Publication Status: Available-Online prior to print DOI: 10.1111/jse.12702
REFERENCE	2 (bases 1 to 599)
AUTHORS	Liao,C. and Downie,S.
TITLE	Direct Submission
JOURNAL	Submitted (07-JUL-2020) College of Architecture and Environment, Sichuan University, No. 24 South Section 1, Yihuan Road, Chengdu, Sichuan 610065, China
FEATURES	Location/Qualifiers
source	1..599 /organism=" <i>Angelica hendersonii</i> " /mol_type="genomic DNA" /specimen_voucher="Tracey & V. Call 2490 (OBI09031)" /db_xref="taxon: 2831622 " /country="USA" /collection_date="05-Jul-1966" /collected_by="Tracey & V. Call" /identified_by="C.Y. Liao"
misc RNA	1..216 /product="internal transcribed spacer 1"
rRNA	217..378 /product="5.8S ribosomal RNA"
misc RNA	379..599 /product="internal transcribed spacer 2"
ORIGIN	1 tcgaatcctq caataacaga atgaccqct aacacattaa caatttqqc gaacatcggg

Webpage associated with GenBank accession MT735455 as seen via <https://www.ncbi.nlm.nih.gov/nucleotide/MT735455> on 2023-09-11.

Our desire is to develop a method to facilitate the discovery of this preserved specimen and their associated GenBank accession records. The annotated web page screenshots below gives some hints to what information elements may be used to help associated related records.

LOCUS	MT735455	599 bp	DNA	Linear	PLN 23-MAY-2021
DEFINITION	<i>Angelica hendersonii</i> voucher Tracey & V. Call 2490 (OBI09031) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.				
ACCESSION	MT735455				
VERSION	MT735455.1				
KEYWORDS					
SOURCE	<i>Angelica hendersonii</i>				
ORGANISM	<i>Angelica hendersonii</i> Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetales; asterids; campanulids; Apiales; Apiaceae; Apioideae; aploid superclade; Selineae; Angelica.				
REFERENCE	1 (bases 1 to 599)				
AUTHORS	Liao, C.-Y., Gao, Q., Katz-Downie, D.S. and Downie, S.R.				
TITLE	A systematic study of North American <i>Angelica</i> species (Apiaceae) based on nrDNA ITS and cpDNA sequences and fruit morphology				
JOURNAL	J Syst Evol (2021) In press				
REMARK	Publication Status: Available-Online prior to print DOI: 10.1111/jse.12792				
REFERENCE	2 (bases 1 to 599)				
AUTHORS	Liao, C. and Downie, S.				
TITLE	Direct Submission				
JOURNAL	Submitted (07-JUL-2020) College of Architecture and Environment, Sichuan University, No. 24 South Section 1, Yihuan Road, Chengdu, Sichuan 610065, China				
FEATURES	Location/Qualifiers				
source	1..599 /organism="Angelica hendersonii" /mol_type="genomic DNA" /specimen_voucher="Tracey & V. Call 2490 (OBI09031)" /db_xref="taxon:2831622" /country="USA" /collection_date="05-Jul-1966" /collected_by="Tracey & V. Call" /identified_by="C.Y. Liao"				
misc_RNA	1..216 /product="internal transcribed spacer 1"				
rRNA	217..378 /product="5.8S ribosomal RNA"				
misc_RNA	379..599 /product="internal transcribed spacer 2"				
ORIGIN	1. tgcatttc cctgacgc atgacgac cagccttc cactttgac cagccttcg				

Details	Genetic	Duplicates	Comments	Linked Resources
<p>CAL POLY OBI Robert F. Hoover Herbarium, Cal Poly State University (OBI)</p> <p>Occurrence ID: 256368e3-f8d7-4028-8010-1a4ff3eb8111 Secondary Catalog #: 9031 Taxon: <i>Angelica hendersonii</i> J.M. Coult. & Rose Family: Apiaceae Show Determination History</p> <p>Collector: Tracey Call Number: 2490 Date: 1966-07-05 Verbatim Date: 5-Jul-66</p> <p>Additional Collectors: Viola Call Locality: United States, California, Marin, North end of Tomales Bay and 2 mi south of Tomales Elevation: 15 meters Verbatim Elevation: 50ft. Habitat: Low bluffs Usage Rights: CC BY-NC (Attribution-Non-Commercial) Record ID: 9a370197-6899-4072-8b17-4f2f043fbd54</p> <p>For additional information about this specimen, please contact: Jenn Yost, Director and Associate Professor (jyost@calpoly.edu)</p>				

At first glance, the highlighted parts of the html pages appear to suggest evidence of association between specimen record OBI09031 and accession record MT735455. These associations include OBI (the institution code), *Angelica hendersonii* (taxonomic identification), 1966 (collection year), 2490 (collector number), 9031 (secondary catalog), and Tracy Call and Viola Call (collectors).

Methods

Instead of relying on visual inspection of individual html pages for herbarium specimen and GenBank accession records, an data-driven workflow was designed to first acquire and version GenBank and OBI records. Then, using these versioned archives, the records are analyzed and associated record candidates

are proposed.

The Hoover Herbarium publishes their digital collections using DwC-A through the CCH2 portal. And, they registered their collection with the GBIF dataset registry.

Phase 1. Acquire and Version

Acquire and Version GenBank Accession Records GenBank publishes their sequence accession records in flat file archives online via <https://ftp.ncbi.nlm.nih.gov/genbank/>. Their publications are published grouped by divisions. One of these divisions, the so-called PLN division, covers sequences of plants, fungi and algae.

We used Preston, a biodiversity dataset tracker, to track GenBank PLN sequence records and make them available for versioned archiving, and offline processing [1].

The following script was used to track the GenBank PLN sequence records:

```
#!/bin/bash
#
# Lists Genbank Plant sequence entries (including fungi and algae)
#
# For more info, see https://ftp.ncbi.nlm.nih.gov/genbank/README.genbank

preston track "https://ftp.ncbi.nlm.nih.gov/genbank/gbpln.txt" \
| preston cat \
| grep -oE "gbpln+[0-9]+[.]seq" \
| sed 's+^+https://ftp.ncbi.nlm.nih.gov/genbank/+g' \
| sed 's+$.gz+g'
```

At the time, this produced a list of resources starting with:

```
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln10.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln100.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1000.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1001.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1002.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1003.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1004.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1005.seq.gz
https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1006.seq.gz
```

These files ending with `seq.gz` were then tracked using command like:

```
preston track https://ftp.ncbi.nlm.nih.gov/genbank/gbpln1.seq.gz
```

A Preston package was built using these “track” commands to document where

and when genbank resources were accessed, and what they contained. In addition, copies of the resources were made. This package can be uniquely identified by the following content id:

hash://sha256/bc7368469e50020ce8ae27b9d6a9a869e0b9a2a0a9b5480c69ce6751fa4b870e

This resulting Preston package of GenBanks PLN division record was archived offline on an external harddisk and online at ASU's BioKiC (Biodiversity Knowledge integration Center) and made available via <https://linker.bio>. The total volume of the GenBank PLN records was a little over 200GB, small enough to fit on a personal computer, or external hard disk.

Acquire and Version OBI Herbarium Specimen Records Similarly, the OBI specimen records were tracked and archived using Preston [2]. Then, this versioned and offline enabled archive was used to query for identifiers found in candidate records.

For instance, GenBank accession record <https://www.ncbi.nlm.nih.gov/nuccore/MT735455> references numbers like “2490” and “9031” (from OBI09031) extracted from their locus. These numbers are then used to select records that contain both via query:

```
preston ls\
--anchor hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6\
--remote https://linker.bio\
--no-cache\
| preston dwc-stream\
--remote https://linker.bio\
--no-cache\
| grep -E "[^0-9a-zA-Z-] (2490) [^0-9a-zA-Z-]" \
| grep -E "[^0-9a-zA-Z-] (9031) [^0-9a-zA-Z-]"
```

where the lines with “grep” in is select only records that have the specified number (e.g., 2490, 9031) where the characters preceding and following are *not* alphanumeric characters. In this example, on only a single record has both numbers in it.

Phase 2. Propose OBI associated GenBank Records

Then the GenBank archive was processed to list all records that mention “OBI” in their (locus, voucher_specimen) descriptions using:

```
preston ls\
--anchor hash://sha256/bc7368469e50020ce8ae27b9d6a9a869e0b9a2a0a9b5480c69ce6751fa4b870e\
--remote https://linker.bio,https://zenodo.org/record/8117720/files/,https://biokic6.rc.asu.edu\
--no-cache\
| preston gb-stream\
--remote https://linker.bio,https://zenodo.org/record/8117720/files/,https://biokic6.rc.asu.edu
```

```
--no-cache\  
| grep "OBI"
```

The first command (i.e., `preston ls ... https://linker.bio`) lists the content of the package with id hash://sha256/bc7368469e50020ce8ae27b9d6a9a869e0b9a2a0a9b5480c69ce6751fa4b87 and downloads the necessary data via `https://linker.bio` if needed.

The second command (i.e., `preston gb-stream`) analyzes the package content as a stream, and generates metadata objects for each genbank accession encountered.

The third command (i.e., `grep "OBI"`) includes only those datadata records that contain “OBI”.

Results

Capture GenBank Candidate Records

On processing millions of GenBank accession records, 256 candidate genbank accession records with mention of OBI were shared with Katelin Pearson for review. By selecting the PLN division (plants), and selecting the OBI institutions code, we reduced the search space by a couple of order of magnitudes. With only a few hundred records, Kateline Pearson, an OBI curator, was able to make the candidate GenBank accession records that likely referenced OBI specimen records (see `genbank-associations-mentioning-OBI.csv` or associated online sheet for the table with manual review notes).

Following, Jorrit Poelen used the OBI preston archive and retrieved preserved specimen records that contained numbers and/or other identifying information (e.g., scientific name occurring in the genbank accession record) to select a candidate specimen record for each candidate accession record. In about 1.5 hours, he compiled this list of specimen record / accession records associations in the following format.

occid	url	resourcename	locus
4060422	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1025115	GenBank/Record/M1025115	<p>W025115 sp.</p> <p>SR-2020 voucher</p> <p>OBI161445 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence.</p>

Curatorial Candidate Record Review

With this information, Katelin Pearson, a OBI data curator, took about 15 minutes to annotate the specimen records in the CCH2 Symbiota database with their related GenBank Accession number. Most of this time was spent to convert the information provided by Jorrit Poelen into a more convenient format. The full table can be found in Appendix A and the first two lines of the OBI genetics table can be found below. Here, the occid is the record number unique to the CCH2 Symbiota database, url is the reference a GenBank accession, resourcename is the type of resource that Symbiota understands, and locus the optional information supported by Symbiota to include in an associated sequence record.

Adding GenBank Links to Symbiota Records

After Katelin Pearson upload the genbank link table into the CCH2 Symbiota database, she exported the updated records to the published DwC-A. Following, Jorrit Poelen tracked the updated version of the DwC-A and selected the records with associated GenBank sequence records. Following, he created a table (see Appendix B.) including the reference to the original record, a web url to a html record page, and the associated genbank record annotations. The first three lines of Appendix B. can be found below.

derivedFrom	reference	associatedSequences
https://linker.bio/line:zip:hash:https://a252.org/peer/91e9735/109766inB3394451chm9/486a4827621h3bca10398678037b99ab	https://a252.org/peer/91e9735/109766inB3394451chm9/486a4827621h3bca10398678037b99ab	
https://linker.bio/line:zip:hash:https://a252.org/peer/91e9735/109766inB3394451chm9/486a4827621h3bca10398678037b99ab	https://a252.org/peer/91e9735/109766inB3394451chm9/486a4827621h3bca10398678037b99ab	<p>Angelica hendersonii voucher Tracey & V. Call 2490 (OBI09031) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence., https://www.ncbi.nlm.nih.gov/nuccore/MT735455 GenE Record, Angelica hendersonii Tracey & V. Call 2490 (OBI09031) ndhF-rpl32 intergenic spacer, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MT765790 GenE Record, Angelica hendersonii Tracey & V. Call 2490 (OBI09031) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence., https://www.ncbi.nlm.nih.gov/nuccore/MT765975 GenE Record, Angelica hendersonii Tracey & V. Call 2490 (OBI09031) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MT766140</p>

Comparing Example Record Before and After Record Linking In our methods, we keep track of the versions of the datasets we work with. The OBI specimen records were versioned prior and after annotating OBI specimen records with their associated GenBank accessions. This means that the changes in the records, as published via the OBI DwC-A can be measured.

To demonstrate the changes to a specific record related to our example specimen record OBI09031, please consider the record prior to annotating the association:

```

preston ls\
--anchor hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6\
--remote https://linker.bio\
--no-cache\
| preston dwc-stream\
--remote https://linker.bio\
--no-cache\
| grep -E "[^0-9a-zA-Z-] (2490) [^0-9a-zA-Z-]" \
| grep -E "[^0-9a-zA-Z-] (9031) [^0-9a-zA-Z-]" \
| tail -n1\
| jq --raw-output '["http://www.w3.org/ns/prov#wasDerivedFrom"] '

```

which points us to the versioned records with identifier:

```
line:zip:hash://sha256/b60f9dd7868d6296ddea107219d41e5a92d55f1a5e0e5ee894c6e9977cb872cd!/occ
```

The content associated with this content identifier can be retrieved via preston

```
cat 'line:zip:hash://sha256/b60f9dd7868d6296ddea107219d41e5a92d55f1a5e0e5ee894c6e9977cb872cd'
or accessed via OBI09031@b60f9.csv.
```

A textual representation of the record is shown below.

id	166203
institutionCode	OBI
collectionCode	
ownerInstitutionCode	
collectionID	3818d95b-b6a4-11e8-b408-001a64db2964
basisOfRecord	PreservedSpecimen
occurrenceID	256368e3-f8d7-4028-8010-1a4ff3eb8111
catalogNumber	
otherCatalogNumbers	9031
higherClassification	Organism Plantae Viridiplantae Streptophyta Embryophyta Tracheophyta
kingdom	Plantae
phylum	Tracheophyta
class	Magnoliopsida
order	Apiales
family	Apiaceae
scientificName	Angelica hendersonii
taxonID	210544
scientificNameAuthorship	Coult. & Rose
genus	Angelica
subgenus	
specificEpithet	hendersonii
verbatimTaxonRank	
infraspecificEpithet	
taxonRank	Species
identifiedBy	
dateIdentified	

identificationReferences	
identificationRemarks	
taxonRemarks	
identificationQualifier	
typeStatus	
recordedBy	Tracey Call; Viola Call
recordNumber	2490
eventDate	1966-07-05
year	1966
month	7
day	5
startDayOfYear	186
endDayOfYear	
verbatimEventDate	5-Jul-66
occurrenceRemarks	
habitat	Low bluffs
fieldNumber	
eventID	
informationWithheld	
dataGeneralizations	
dynamicProperties	
associatedOccurrences	herbariumSpecimenDuplicate: https://cch2.org/portal/collection
associatedSequences	
associatedTaxa	
reproductiveCondition	
establishmentMeans	
lifeStage	
sex	
individualCount	
preparations	
locationID	
continent	
waterBody	
islandGroup	
island	
country	United States
stateProvince	California
county	Marin
municipality	
locality	North end of Tomales Bay and 2 mi south of Tomales
locationRemarks	
decimalLatitude	
decimalLongitude	
geodeticDatum	
coordinateUncertaintyInMeters	
verbatimCoordinates	

```

georeferencedBy
georeferenceProtocol
georeferenceSources
georeferenceVerificationStatus
georeferenceRemarks
minimumElevationInMeters      15
maximumElevationInMeters
minimumDepthInMeters
maximumDepthInMeters
verbatimDepth
verbatimElevation             50ft.
disposition
language
recordEnteredBy
modified                      2011-08-18 00:00:00
rights                        http://creativecommons.org/licenses/by-nc/4.0/
rightsHolder
accessRights
recordID                      9a370197-6899-4072-8b17-4f2f043fbd54
references                    https://cch2.org/portal/collections/individual/index.php?occ

```

Similarly, the record seen after the annotation can be retrieved using:

```

preston ls\
--anchor hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6\
--remote https://linker.bio\
--no-cache\
| preston dwc-stream\
--remote https://linker.bio\
--no-cache\
| grep -E "[^0-9a-zA-Z-] (2490) [^0-9a-zA-Z-]" \
| grep -E "[^0-9a-zA-Z-] (9031) [^0-9a-zA-Z-]" \
| head -n1\
| jq --raw-output '["http://www.w3.org/ns/prov#wasDerivedFrom"]'

```

yielding:

```
line:zip:hash://sha256/cd9de973510975dac3394952bba9c486a482762b3beab05ecb678037b99ab85b!/occ
```

Now, we can use a text comparison between the two versioned records, using diff, a widely available linux tool.

```
diff <(preston cat 'line:zip:hash://sha256/b60f9dd7868d6296ddea107219d41e5a92d55f1a5e0e5ee89
```

which results in

```

50c50
<  "associatedSequences": "",
---
>  "associatedSequences": "GenBank Record, Angelica hendersonii voucher Tracey & V. Call 29

```

: output of a commonly used programming tool `diff` as applied to our OBI09031 example.



output of a visual text comparison tool available via <https://commontools.org> as applied to our OBI09031 example.

Additionally, you can find the before/after example records in json OBI09031-before.json/ OBI09031-after.json or csv OBI09031-before.csv / OBI09031-after.csv formats.

Finally, because we have our versioned records available in text formats, the options for re-use, archiving, or other subsequent processing are plentiful, and is consistent with one of the unix principles.

Expect the output of every program to become the input to another, as yet unknown, program.

Discussion

We took a systematic approach to independently track natural history collection records and sequence records. Then, we used regular expressions (or queries) to select candidate GenBank accession records. Following, after manual review of candidate records, we extracted identifiers and names to link locate their associated specimen records in the Hoover Herbarium collection as tracked. While the method is not fully automated, our method reduced the number of candidate accession records from millions to hundreds. This many order of magnitude reduction of candidates made manual review was feasible. We expect that periodic revisiting of the available records in GenBank will yield additional associated genbank records. Also, we hope that this example show that links between existing GenBank accessions and their specimen records can be found without major technological investment. And, we hope that this example will help inspire to develop best practices to place identifying information in GenBank records such that collection managers can somewhat easily locate sequences associated to the specimen they keep.

References

- [1] Poelen, Jorrit H. (2023). GenBank PLN (Plantae, Fungi, Algae) Sequence Index in TSV, CSV, JSONL formats hash://sha256/bc7368469e50020ce8ae27b9d6a9a869e0b9a2a0a9b5480c69ce675 hash://md5/f6f78f64e3b3ff06adc3229badbd578b (0.1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8117720>
- [2] Jorrit Poelen, Katelin Pearson, and Jenn Yost. 2023. Extending OBI Herbarium Records to include associated NCBI GenBank sequences. <https://github.com/jhpoelen/obi-genbank> hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb1

Appendix A

GenBank link table created by Katelin Pearson to link OBI specimen records to their associated GenBank sequences.

See also OBI_genetics.csv.

occid	url	resourcename	locus
4060422	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1025115	GenBank/Record/M1025115	<p>sp.</p> <p>SR-2020 voucher</p> <p>OBI161445 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence.</p>
2186655	https://www.ncbi.nlm.nih.gov/GenBank/Record/M670383	GenBank/Record/M670383	<p>marlothii voucher</p> <p>Rodin 9194 (OBI) trnS-trnG intergenic spacer, partial sequence; chloroplast.</p>

occid	url	resourcename	locus
2186655	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1768100	GenBank/Record/M1768100	Chlamys marlothii voucher Rodin 9194 (OBI) ribosomal protein S16 (rps16) gene, intron; chloroplast.
2186655	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1768102	GenBank/Record/M1768102	Chlamys marlothii voucher Rodin 9194 (OBI) trnT-trnL intergenic spacer, partial sequence; chloroplast.
2186655	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1768101	GenBank/Record/M1768101	Chlamys marlothii voucher Rodin 9194 (OBI) trnL-trnF intergenic spacer, partial sequence; chloroplast.
2186655	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1768108	GenBank/Record/M1768108	Chlamys marlothii voucher Rodin 9194 (OBI) internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.

occid	url	resourcename	locus
214465	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF785179	GenBank/Record/AF785179	lucida voucher D. Smith 203 (OBI13881) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
214465	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765849	GenBank/Record/AF765849	lucida D. Smith 203 (OBI13881) ndhF-rpl32 intergenic spacer, partial sequence.
214465	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766044	GenBank/Record/AF766044	lucida D. Smith 203 (OBI13881) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
214465	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766264	GenBank/Record/AF766264	lucida D. Smith 203 (OBI13881) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.

occid	url	resourcename	locus
214463	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF735149	GenBank/Record/AF735149	scabrida voucher A.C. Sanders et al. 6885 (OBI044899) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
214463	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765845	GenBank/Record/AF765845	scabrida A.C. Sanders et al. 6885 (OBI044899) ndhF-rpl32 intergenic spacer, partial sequence.
214463	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766024	GenBank/Record/AF766024	scabrida A.C. Sanders et al. 6885 (OBI044899) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
214463	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766162	GenBank/Record/AF766162	scabrida A.C. Sanders et al. 6885 (OBI044899) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.

occid	url	resourcename	locus
211800	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF735148	GenBank/Record/AF735148	lucida voucher Tracey & V. Call 2507 (OBI081640) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
211800	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF763854	GenBank/Record/AF763854	lucida Tracey & V. Call 2507 (OBI081640) ndhF-rpl32 intergenic spacer, partial sequence.
211800	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF763855	GenBank/Record/AF763855	lucida Tracey & V. Call 2507 (OBI081640) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
211800	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF763856	GenBank/Record/AF763856	lucida Tracey & V. Call 2507 (OBI081640) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.

occid	url	resourcename	locus
198762	https://www.ncbi.nlm.nih.gov/GenBank/Record/JF951140	GenBank/Record/JF951140	lemmonii isolate LEM25383 trnT-trnL intergenic spacer, partial sequence; tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; plastid.
196156	https://www.ncbi.nlm.nih.gov/GenBank/Record/ON157416	GenBank/Record/ON157416	secundiflorus var. secundiflorus voucher OBI:29532 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
196156	https://www.ncbi.nlm.nih.gov/GenBank/Record/ON136165	GenBank/Record/ON136165	secundiflorus var. secundiflorus voucher OBI:DKeil29532 atpB-rbcL intergenic spacer region, partial sequence; chloroplast.

occid	url	resourcename	locus
184474	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1025106	GenBank/Record/M1025106	ojaiensis voucher OBI75168 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664492	GenBank/Record/M664492	scariosum var. scariosum voucher OBI 60356 external transcribed spacer, partial sequence.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664582	GenBank/Record/M664582	scariosum var. scariosum voucher OBI 60356 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.

occid	url	resourcename	locus
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M60468	GenBank/Record/M60468	scariosum var. scariosum voucher OBI 60356 maturase K (matK) gene, partial cds; chloroplast.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M60472	GenBank/Record/M60472	scariosum var. scariosum voucher OBI 60356 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M60482	GenBank/Record/M60482	scariosum var. scariosum voucher OBI 60356 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M60496	GenBank/Record/M60496	undulatum voucher OBI 60365 external transcribed spacer, partial sequence.

occid	url	resourcename	locus
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664610	GenBank/Record/M664610	undulatum voucher OBI 60365 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664651	GenBank/Record/M664651	undulatum voucher OBI 60365 maturase K (matK) gene, partial cds; chloroplast.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664726	GenBank/Record/M664726	undulatum voucher OBI 60365 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175596	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664735	GenBank/Record/M664735	undulatum voucher OBI 60365 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.

occid	url	resourcename	locus
175592	https://www.ncbi.nlm.nih.gov/GenBank/Record/M03275447	GenBank/Record/M03275447	scariosum var. citrinum voucher OBI 29634F photosystem II protein D1 (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175592	https://www.ncbi.nlm.nih.gov/GenBank/Record/M03344906	GenBank/Record/M03344906	scariosum var. citrinum voucher OBI 29634F tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; chloroplast.
175592	https://www.ncbi.nlm.nih.gov/GenBank/Record/M03351162	GenBank/Record/M03351162	scariosum var. citrinum voucher OBI 29634F internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.

occid	url	resourcename	locus
175592	https://www.ncbi.nlm.nih.gov/GenBank/Record/M0230952	GenBank/Record/M0230952	scariosum var. citrinum voucher OBI 29634F external transcribed spacer, partial sequence.
175581	https://www.ncbi.nlm.nih.gov/GenBank/Record/M0664493	GenBank/Record/M0664493	scariosum var. toiyabense voucher OBI 60380 external transcribed spacer, partial sequence.
175581	https://www.ncbi.nlm.nih.gov/GenBank/Record/M0664583	GenBank/Record/M0664583	scariosum var. toiyabense voucher OBI 60380 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
175581	https://www.ncbi.nlm.nih.gov/GenBank/Record/M0664639	GenBank/Record/M0664639	scariosum var. toiyabense voucher OBI 60380 maturase K (matK) gene, partial cds; chloroplast.

occid	url	resourcename	locus
175581	https://www.ncbi.nlm.nih.gov/GenBank/Record/M6417273	GenBank/Record/M6417273	scariosum var. toiyabense voucher OBI 60380 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175581	https://www.ncbi.nlm.nih.gov/GenBank/Record/M6417333	GenBank/Record/M6417333	scariosum var. toiyabense voucher OBI 60380 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.
175526	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664487	GenBank/Record/M664487	ochrocentrum voucher OBI 60392 external transcribed spacer, partial sequence.

occid	url	resourcename	locus
175526	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664551	GenBank/Record/M664551	ochrocentrum voucher OBI 60392 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
175526	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664674	GenBank/Record/M664674	ochrocentrum voucher OBI 60392 maturase K (matK) gene, partial cds; chloroplast.
175526	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664729	GenBank/Record/M664729	ochrocentrum voucher OBI 60392 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175526	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664734	GenBank/Record/M664734	ochrocentrum voucher OBI 60392 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.

occid	url	resourcename	locus
175241	https://www.ncbi.nlm.nih.gov/GenBank/Record/M230951	GenBank/Record/M230951	fontinale var. campylon voucher OBI 27922 external transcribed spacer, partial sequence.
175241	https://www.ncbi.nlm.nih.gov/GenBank/Record/M275341	GenBank/Record/M275341	fontinale var. campylon voucher OBI 27922 maturase K (matK) gene, partial cds; chloroplast.
175241	https://www.ncbi.nlm.nih.gov/GenBank/Record/M275448	GenBank/Record/M275448	fontinale var. campylon voucher OBI 27922 photosystem II protein D1 (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175241	https://www.ncbi.nlm.nih.gov/GenBank/Record/M334405	GenBank/Record/M334405	fontinale var. campylon voucher OBI 27922 tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; chloroplast.

occid	url	resourcename	locus
175241	https://www.ncbi.nlm.nih.gov/GenBank/Record/M335163	GenBank/Record/M335163	fontinale var. campylon voucher OBI 27922 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
175241	https://www.ncbi.nlm.nih.gov/GenBank/Record/M230952	GenBank/Record/M230952	scariosum var. citrinum voucher OBI 29634F external transcribed spacer, partial sequence.
175222	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664514	GenBank/Record/M664514	eatonii var. eatonii voucher OBI 64116 external transcribed spacer, partial sequence.
175222	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664550	GenBank/Record/M664550	eatonii var. eatonii voucher OBI 64116 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.

occid	url	resourcename	locus
175222	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664666	GenBank/Record/M664666	eatonii var. eatonii voucher OBI 64116 maturase K (matK) gene, partial cds; chloroplast.
175222	https://www.ncbi.nlm.nih.gov/GenBank/Record/M641734	GenBank/Record/M641734	eatonii var. eatonii voucher OBI 64116 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175222	https://www.ncbi.nlm.nih.gov/GenBank/Record/M641730	GenBank/Record/M641730	eatonii var. eatonii voucher OBI 64116 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.
175203	https://www.ncbi.nlm.nih.gov/GenBank/Record/M230934	GenBank/Record/M230934	cymosum var. canovirens voucher OBI 30302-8 external transcribed spacer, partial sequence.
175203	https://www.ncbi.nlm.nih.gov/GenBank/Record/M235314	GenBank/Record/M235314	cymosum var. canovirens voucher OBI 30302-8 maturase K (matK) gene, partial cds; chloroplast.

occid	url	resourcename	locus
175203	https://www.ncbi.nlm.nih.gov/GenBank/Record/NC_034484	GenBank/Record/NC_034484	175203cymosum var. canovirens voucher OBI 30302-8 photosystem II protein D1 (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.
175203	https://www.ncbi.nlm.nih.gov/GenBank/Record/NC_034484	GenBank/Record/NC_034484	175203cymosum var. canovirens voucher OBI 30302-8 tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; chloroplast.
175203	https://www.ncbi.nlm.nih.gov/GenBank/Record/NC_034484	GenBank/Record/NC_034484	175203cymosum var. canovirens voucher OBI 30302-8 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.

occid	url	resourcename	locus
175187	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664465	GenBank/Record/M664465	Seatonii var. clokeyi voucher OBI 62978 external transcribed spacer, partial sequence.
175187	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664549	GenBank/Record/M664549	Seatonii var. clokeyi voucher OBI 62978 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
175187	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664665	GenBank/Record/M664665	Seatonii var. clokeyi voucher OBI 62978 maturase K (matK) gene, partial cds; chloroplast.
175187	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664723	GenBank/Record/M664723	Seatonii var. clokeyi voucher OBI 62978 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.

occid	url	resourcename	locus
175187	https://www.ncbi.nlm.nih.gov/GenBank/Record/M6647809	GenBank/Record/M6647809	eatonii var. clokeyi voucher OBI 62978 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.
175185	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664450	GenBank/Record/M664450	ciliolatum voucher OBI 60321 external transcribed spacer, partial sequence.
175185	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664539	GenBank/Record/M664539	ciliolatum voucher OBI 60321 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
175185	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664661	GenBank/Record/M664661	ciliolatum voucher OBI 60321 maturase K (matK) gene, partial cds; chloroplast.

occid	url	resourcename	locus
175185	https://www.ncbi.nlm.nih.gov/GenBank/Record/M641721	GenBank/Record/M641721	<p>ciliolatum voucher OBI 60321 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.</p>
175185	https://www.ncbi.nlm.nih.gov/GenBank/Record/M641726	GenBank/Record/M641726	<p>ciliolatum voucher OBI 60321 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.</p>
175101	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664445	GenBank/Record/M664445	<p>arizonicum var. tenuisectum voucher OBI 62969 external transcribed spacer, partial sequence.</p>

occid	url	resourcename	locus
175101	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664609	GenBank/Record/M664609	arizonicum var. tenuisectum voucher OBI 62969 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence.
175101	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664612	GenBank/Record/M664612	arizonicum var. tenuisectum voucher OBI 62969 maturase K (matK) gene, partial cds; chloroplast.
175101	https://www.ncbi.nlm.nih.gov/GenBank/Record/M664724	GenBank/Record/M664724	arizonicum var. tenuisectum voucher OBI 62969 psbA (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast.

occid	url	resourcename	locus
175101	https://www.ncbi.nlm.nih.gov/GenBank/Record/M6417291	GenBank/Record/M6417291	arizonicum var. tenuisectum voucher OBI 62969 tRNA-Leu (trnL) gene and trnL-trnF intergenic spacer, partial sequence; chloroplast.
166210	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1735142	GenBank/Record/M1735142	lineariloba voucher Tracey & V. Call 2043 (OBI081607) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
166210	https://www.ncbi.nlm.nih.gov/GenBank/Record/M1765840	GenBank/Record/M1765840	lineariloba Tracey & V. Call 2043 (OBI081607) ndhF-rpl32 intergenic spacer, partial sequence.

occid	url	resourcename	locus
166210	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766049	lineariloba Tracey & V. Call 2043 (OBI081607) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.	lineariloba Tracey & V. Call 2043 (OBI081607) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
166210	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766157	lineariloba Tracey & V. Call 2043 (OBI081607) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.	lineariloba Tracey & V. Call 2043 (OBI081607) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.
166209	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765448	lineariloba voucher Tracey & V. Call 2321 (OBI09033) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.	lineariloba voucher Tracey & V. Call 2321 (OBI09033) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
166209	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765844	lineariloba Tracey & V. Call 2321 (OBI09033) ndhF-rpl32 intergenic spacer, partial sequence.	lineariloba Tracey & V. Call 2321 (OBI09033) ndhF-rpl32 intergenic spacer, partial sequence.

occid	url	resourcename	locus
166209	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT766023	GenBank/Record/MT766023	lineariloba Tracey & V. Call 2321 (OBI09033) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
166209	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT766159	GenBank/Record/MT766159	lineariloba Tracey & V. Call 2321 (OBI09033) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.
166208	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT707551	GenBank/Record/MT707551	dissectum voucher D. Keilet al. 30299 (OBI068349) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
166208	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT65778	GenBank/Record/MT65778	dissectum D. Keilet al. 30299 (OBI068349) ndhF-rpl32 intergenic spacer, partial sequence.

occid	url	resourcename	locus
166208	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT666091	GenBank/Record/MT666091	dissectum D. Keilet al. 30299 (OBI068349) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
166208	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT666279	GenBank/Record/MT666279	dissectum D. Keilet al. 30299 (OBI068349) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.
166207	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT735143	GenBank/Record/MT735143	lineariloba voucher D. Keil 21070 (OBI071409) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
166207	https://www.ncbi.nlm.nih.gov/GenBank/Record/MT765841	GenBank/Record/MT765841	lineariloba D. Keil 21070 (OBI071409) ndhF-rpl32 intergenic spacer, partial sequence.

occid	url	resourcename	locus
166207	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766022	lineariloba D. Keil 21070 (OBI071409) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.	lineariloba D. Keil 21070 (OBI071409) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
166207	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766158	lineariloba D. Keil 21070 (OBI071409) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.	lineariloba D. Keil 21070 (OBI071409) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.
166204	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765454	hendersonii voucher Tracey & V. Call 2071 (OBI09030) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.	hendersonii voucher Tracey & V. Call 2071 (OBI09030) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
166204	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765781	hendersonii Tracey & V. Call 2071 (OBI09030) ndhF-rpl32 intergenic spacer, partial sequence.	hendersonii Tracey & V. Call 2071 (OBI09030) ndhF-rpl32 intergenic spacer, partial sequence.

occid	url	resourcename	locus
166204	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765974	GenBank/Record/AF765974	hendersonii Tracey & V. Call 2071 (OBI09030) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
166204	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766189	GenBank/Record/AF766189	hendersonii Tracey & V. Call 2071 (OBI09030) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.
166203	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765455	GenBank/Record/AF765455	hendersonii voucher Tracey & V. Call 2490 (OBI09031) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence.
166203	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765760	GenBank/Record/AF765760	hendersonii Tracey & V. Call 2490 (OBI09031) ndhF-rpl32 intergenic spacer, partial sequence.

occid	url	resourcename	locus
166203	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF765975	henderonii	Tracey & V. Call 2490 (OBI09031) tRNA-Asp (trnD-GUC), tRNA-Tyr (trnY-GUA), tRNA-Glu (trnE-UUC), and tRNA-Thr (trnT-GGU) genes, complete sequence.
166203	https://www.ncbi.nlm.nih.gov/GenBank/Record/AF766140	henderonii	Tracey & V. Call 2490 (OBI09031) rpl32-trnL intergenic spacer and tRNA-Leu (trnL) gene, partial sequence.

Appendix B

References to specimen records with associated sequences after application of links of Appendix A.

generated using:

```
preston cat\
--remote https://linker.bio\
hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6\
| preston dwc-stream\
| jq -c 'select(["http://rs.tdwg.org/dwc/terms/associatedSequences"] != null)'\
| jq '{ derivedFrom: ["http://www.w3.org/ns/prov#wasDerivedFrom"], reference: ["http://p
| sed 's+line:zip+https://linker.bio/line:zip+g'\
| sed 's+occurrences.csv!/+occurrences.csv!/L1,+g'\
| mlr --ijson --ocsv cat
```

See also specimen-record-with-associated-sequences.csv.

derivedFrom	reference	associatedSequences
https://linker.bio/line:zip:hastp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000000000/GCF_000000000.186439721.3/bca-1075203678037b99ab	https://doi.org/10.1093/aob/mcy073/5100161 GenBank Cirsium cymosum var. canovirens voucher OBI 30302-8 external transcribed spacer, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN230934 GenERecord, Cirsium cymosum var. canovirens voucher OBI 30302-8 maturase K (matK) gene, partial cds; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN275314 GenERecord, Cirsium cymosum var. canovirens voucher OBI 30302-8 photosystem II protein D1 (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN275448 GenERecord, Cirsium cymosum var. canovirens voucher OBI 30302-8 tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN314894 GenERecord, Cirsium cymosum var. canovirens voucher OBI 30302-8 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN335114	

derivedFrom	reference	associatedSequences
https://linker.bio/line:zip:hash:https://he252.org/nc01735/1006613394/1521019/186432762135da107524b678037b99ab	https://he252.org/nc01735/1006613394/1521019/186432762135da107524b678037b99ab	Cirsium fontinale var. campylon voucher OBI 27922 external transcribed spacer, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN230951 GenE Record, Cirsium scariosum var. citrinum voucher OBI 29634F external transcribed spacer, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN230952 GenE Record, Cirsium fontinale var. campylon voucher OBI 27922 maturase K (matK) gene, partial cds; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN275341 GenE Record, Cirsium fontinale var. campylon voucher OBI 27922 photosystem II protein D1 (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN275438 GenE Record, Cirsium fontinale var. campylon voucher OBI 27922 tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN314905 GenE Record, Cirsium fontinale var. campylon voucher OBI 27922 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN335163

derivedFrom	reference	associatedSequences
https://linker.bio/line:zip:hash:https://a252.org/pep/109735/109735 B33944591 ma9/486448276213 da1075592	https://a252.org/pep/109735/109735 B33944591 ma9/486448276213 da1075592	<p>Cirsium scariosum var. citrinum voucher OBI 29634F external transcribed spacer, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN230952 GenE Record, Cirsium scariosum var. citrinum voucher OBI 29634F photosystem II protein D1 (psbA) gene, partial cds; psbA-trnH intergenic spacer, complete sequence; and tRNA-His (trnH) gene, partial sequence; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN275437 GenE Record, Cirsium scariosum var. citrinum voucher OBI 29634F tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN314906 GenE Record, Cirsium scariosum var. citrinum voucher OBI 29634F internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN335162</p>

derivedFrom	reference	associatedSequences
https://linker.bio/line:zip:hahdps:/line:zip:hahdps:/line:zip:hahdps/">https://linker.bio/line:zip:hahdps:/line:zip:hahdps/	https://doi.org/10.6073/pdb/1U9K	Cirsium scariosum var. scariosum voucher OBI 60356 external transcribed spacer, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN604492 GenE Record, Cirsium undulatum voucher OBI 60365 external transcribed spacer, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN604496 GenE Record, Cirsium scariosum var. scariosum voucher OBI 60356 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN604582 GenE Record, Cirsium undulatum voucher OBI 60365 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence., https://www.ncbi.nlm.nih.gov/nuccore/MN604610 GenE Record, Cirsium scariosum var. scariosum voucher OBI 60356 maturase K (matK) gene, partial cds; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN604638 GenE Record, Cirsium undulatum voucher OBI 60365 maturase K (matK) gene, partial cds; chloroplast., https://www.ncbi.nlm.nih.gov/nuccore/MN604651 GenE Record, Cirsium scariosum var. scariosum voucher OBI 60356 psbA (psbA) gene, partial cds; psbA-trnH

derivedFrom	reference	associatedSequences
https://linker.bio/line:zip:hash:https://a252.org/pev73/109741133914521bna/486448276213bda10876278037b99ab	https://www.ncbi.nlm.nih.gov/nuccore/JF951067 GenBank Record, Phalaris lemmonii isolate LEM25383 trnT-trnL intergenic spacer, partial sequence; tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; plastid., https://www.ncbi.nlm.nih.gov/nuccore/JF951103	isolate LEM25383ITS3, internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence, https://www.ncbi.nlm.nih.gov/nuccore/JF951067 GenBank Record, Phalaris lemmonii isolate LEM25383 trnT-trnL intergenic spacer, partial sequence; tRNA-Leu (trnL) gene, complete sequence; and trnL-trnF intergenic spacer, partial sequence; plastid., https://www.ncbi.nlm.nih.gov/nuccore/JF951103

