The Smithsonian Sequence Hub: a dashboard to illuminate collection impacts via genomic sequencing

Mike Trizna[1], Mirian T. N. Tsuchiya[1], Niamh Redmond[2] & Rebecca B. Dikow[1]

[1] Data Science Lab, Smithsonian Institution, [2] Smithsonian DNA Barcode Network, Smithsonian Institution

Abstract

Genomic sequencing is becoming an integral component of most collections-based biodiversity research. However, it is difficult to connect sequence and genome records in NCBI with specimen collections database entries -- if they exist at all. If best practices are followed using well-formatted BioSample records or the specimen voucher field in the sequence record itself, then NCBI can automatically index these values for searching and linking. There are thousands of existing sequence records derived from Smithsonian specimens that were published without following these best practices, so the Smithsonian Sequence Hub uses machine learning algorithms to identify them. The combined datasets are displayed as several visualizations in dashboard format to show distributions across multiple dimensions, such as sequencing completion date, what sequencing technology was used, taxonomic groups targeted, etc. The infrastructure for the Smithsonian Sequence Hub will be general enough that it can serve as a model for other natural history collections and will showcase the value of natural history collections for genetic and genomic research. The Smithsonian Data Science Lab is also building the Smithsonian Biodiversity Genome Hub as a collaborative platform for analyzing and annotating whole genome projects, and the Sequence Hub will act as a foundation for connecting those projects to specimen records.

Datasets

DarwinCore Archive

The Smithsonian National Museum of Natural History (NMNH) packages its public specimen data in a DarwinCore Archive (DwCA) format package through its hosted Integrated Publishing Toolkit (IPT). This file contains standardized data fields for 7,714,460 specimen records, which includes a field for "associatedSequences." Only 24,123 have a value for this field, of which 24,020 are valid GenBank identifiers. Installing and maintaining an IPT instance is a fairly involved process, and NMNH is the only Smithsonian unit that publishes specimen records in this manner.

BOLD

The Barcode of Life Database (BOLD) contains DNA barcode sequence records from eukaryotic organisms. BOLD also acts as a "workbench" for users to upload working data and gradually refine specimen and sequence data before publishing. his results in a large amount of fully complete records that are left in "private" status and that cannot be accessed by the public. The "institution_storing" field values of "Smithsonian Institution National Museum of Natural History" and "Smithsonian Tropical Research Institute" match 149,248 total public and private records, but public searches only match 60,934 records. Of these, only 23,610 records have associated GenBank identifiers.

US or USNM institution code

It is a best practice to attribute the specimen identifier from which the sequence was derived in a format that includes a standardized collection code (e.g. USNM). NCBI has a BioCollections database (https://www.ncbi.nlm.nih.gov/biocollections/) that indexes all sequence records that follow this format, which enables searching and linking by collection code. The USNM (Smithsonian National Museum of Natural History), US (Smithsonian Natural History Herbarium), STRI (Smithsonian Tropical Research Institute), and SERC (Smithsonian Environmental Research Center) codes together are listed on 37,690 GenBank records. For genome-level sequence records, specimen linking is usually done via separate BioSample entries, which allows for linking several sequence types (i.e. whole genome, transcriptome, and UCE) to a single specimen.

Methods

Downloading GenBank and BOLD records

The Python library "genetic_collections" (https://github.com/MikeTrizna/genetic_collections) is a wrapper around the NCBI eutils API (https://www.ncbi.nlm.nih.gov/books/NBK25500/) that downloads GenBank records in XML format and parses out specimen data fields into a tabular format. It also wraps the BOLD API and can download public data as well as perform some of the complicated screen scraping required to estimate "private" record counts. The library also has functions to convert GenBank and BOLD specimen data fields into DarwinCore datatypes so they can be directly compared, with differences tracked. There are also command-line versions of most functions, allowing access to users without Python knowledge.

Machine Learning to match sequence records to specimen records

The Python library "dedupe" (https://github.com/dedupeio/dedupe) utilizes an active learning approach to calibrate the weights of a machine learning model to perform fuzzy matching and entity resolution. Active learning is a specific type of machine learning that interactively queries a user to provide labels for algorithmically-chosen data points. In the case of the Sequence Hub, dedupe is used to ask a user whether a GenBank record matches a specimen record from the NMNH DwCA. The possible answers for each "query" are "yes", "no", or "unsure". The labeled data are then used to train a logistic regression model (or any other model enabled by the Python scikit-learn library) by adjusting weights of field differences (e.g. a taxonomic name difference might have a higher penalty than a misspelled collector name). Since the algorithm targets examples in an intelligent manner, the number of examples required to train the model is lower than other supervised learning methods.

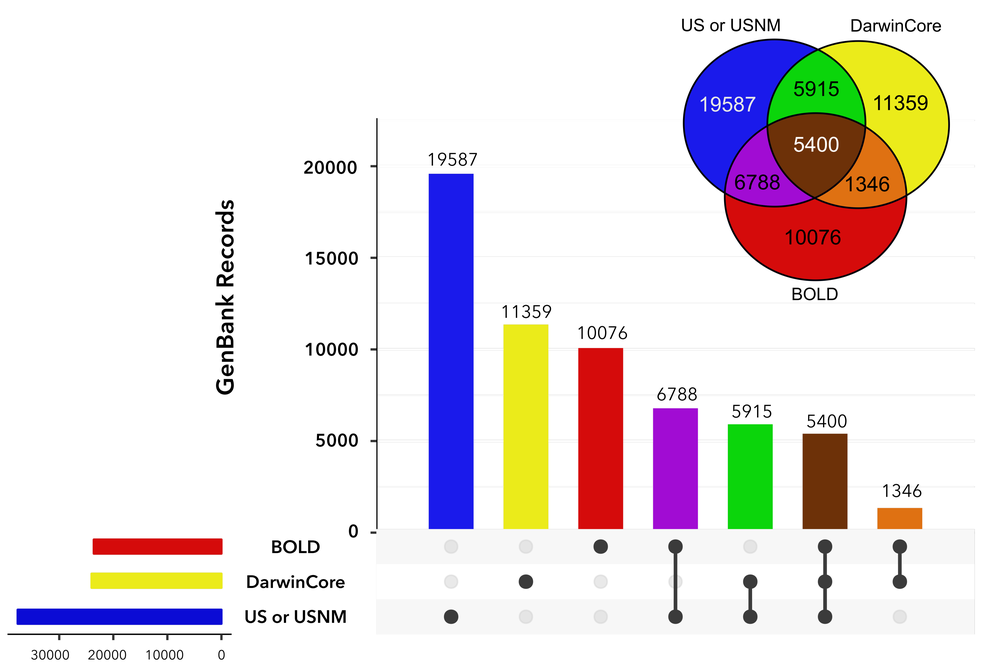

Figure 1: Smithsonian Genbank records with correct USNM designation and their presence in two other datasets: DarwinCoreArchive and BOLD. The colors in the UpSet plot and Venn Diagram correspond to the same dataset overlaps.

NCBI LinkOut

NCBI LinkOut is a service that allows external databases to add links to GenBank and other NCBI database records. The Smithsonian has registered as a LinkOut "provider", which lets us insert GUID links to specific specimen records into GenBank records that are derived from Smithsonian specimens. Currently, only specimen records that have an "associatedSequences" entry (the DarwinCore Archive dataset) have LinkOuts, but we intend to expand to include the other datasets that are described here.

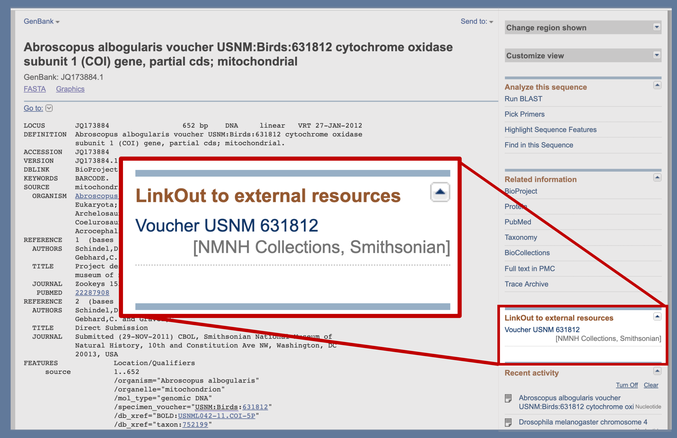

Figure 2: Example of GenBank LinkOut to a Smithsonian specimen record. The record shown is GenBank accession JQ173884.

Future Directions

Online Dashboard

In its current state, the Smithsonian Sequence Hub exists as a collection of Jupyter notebooks that document dataset construction and create visualizations to provide insights into the data. In the coming months, we intend to build an online dashboard website using the Dash library (https://github.com/plotly/dash) from Plotly to organize these various plots, and also allow for data exploration via filters and search tools. Dash is a Flask-powered framework for building analytical web applications. Since Plotly has equivalent libraries for both Python and R, it will also be possible to create the dashboard as an RShiny (https://github.com/rstudio/shiny) application, which would give users the option of choosing their preferred programming language to make modifications.

Integration with Smithsonian Genome Hub

The Smithsonian Biodiversity Genome Hub consists of a web-based platform, with the primary goal of making biodiverse genomes more accessible, analytical tools more available and visualizations and analyses more reproducible for researchers and their collaborators. It will also serve as the data repository of all de novo assemblies funded or generated by Smithsonian researchers and projects. The Smithsonian Sequence Hub will incorporate with the Genome Hub by connecting genomes with Smithsonian specimens. Once all genomic data for a collection are available, they can be used in a number of ways, including calculating the impact and economic value of a collection, tracking use of these data by both researchers (internal and external) and institutions, reducing duplication of experimental effort, protection of rare specimens,