Sequencing Data Discovery with MetaSeek

Adrienne Hoarfrost[1], Nick Brown[2], C. Titus Brown[3], Carol Arnosti[1],

[1] Department of Marine Sciences, University of North Carolina at Chapel Hill
[2] Independent Researcher, Durham, NC
[3] Department of Veterinary Medicine, University of California Davis

Corresponding Author:
Adrienne Hoarfrost
76 Lipman Dr., New Brunswick, NJ, 08901, USA
Email address: adrienne.l.hoarfrost@gmail.com

Abstract

Sequencing data resources have increased exponentially in recent years, as has interest in large-scale meta-analyses of integrated next-generation sequencing datasets. However, curation of integrated datasets that match a user's particular research priorities is currently a time-intensive and imprecise task. MetaSeek is a sequencing data discovery tool that enables users to flexibly search and filter on any metadata field to quickly find the sequencing datasets that meet their needs. MetaSeek automatically scrapes metadata from all publicly available datasets in the Sequence Read Archive, cleans and parses messy, user-provided metadata into a structured, standard-compliant database, and predicts missing fields where possible. MetaSeek provides a web-based graphical user interface and interactive visualization dashboard, as well as a programmatic API to rapidly search, filter, visualize, save, share, and download matching sequencing metadata.

The MetaSeek online interface is available at https://www.metaseek.cloud/. The MetaSeek database can also be accessed via API to programmatically search, filter, and download all metadata. MetaSeek source code, metadata scrapers, and documents are available at https://github.com/MetaSeek-Sequencing-Data-Discovery/metaseek/.

Additional guides, tutorials, and documents are available at https://github.com/MetaSeek-Sequencing-Data-Discovery/metaseek, and on the MetaSeek website, https://www.metaseek.cloud/. MetaSeek is distributed under an MIT license.

Introduction

Sequencing data generation is rapidly increasing, as of 2019 reaching more than 5 million sequencing datasets in the Sequence Read Archive (SRA), the primary repository for next-generation sequencing data in the International Nucleotide Sequence Database Collaboration

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27804v1 | CC BY 4.0 Open Access | rec: 16 Jun 2019, publ: 16 Jun 2019

(NCBI 2017). As research communities produce data at increasingly rapid rates, there is growing interest in leveraging these data resources for new insights into biological systems using comparative meta-analyses of large-scale integrated datasets. Data curation is the first step in this process, and generally requires identifying datasets in data repositories that match certain criteria that may be described within the datasets' metadata.

However, easy-to-use, flexible, and comprehensive tools for searching and filtering existing data repositories according to their metadata parameters are lacking. The e-utilities tool provided by the National Center for Biotechnology Information (NCBI), for example, is restricted to a free text search or exact string matching on a limited set of fields (Sayers 2017). A tool such as SRAdb (Zhu et al., 2013), in contrast, expands the searchability of metadata fields, but is specific to the R programming language and requires a local build of the SRAdb database. Neither of these tools, meanwhile, address the widespread errors in sequencing metadata, which is collected mainly via user-provided free text entries that result in frequent misspellings, missing fields, and nonobservance of existing metadata standards, the Minimum Information about any Sequence (MIxS) specification (Yilmaz et al. 2011).

MetaSeek provides a sequencing data discovery tool that facilitates easy and rapid curation of integrated sequencing datasets. The MetaSeek interface is intuitive, user-friendly, and flexible, allowing users to search on any metadata field in any of 10 ways. For programmatic access, MetaSeek exposes a simple API that is programming language agnostic.

Infrastructure & Implementation

MetaSeek automatically scrapes metadata from the SRA on a weekly basis. In the SRA and in MetaSeek, each '#RX' accession ID (SRX, ERX, or DRX depending on whether it originates from the USA NCBI, European EBI, or Japanese DDBJ databases respectively) is a unique metadata entry. Metadata for each dataset are gathered from across the SRA, BioSample, and PubMed databases and unified for each MetaSeek dataset entry.

As metadata are scraped from the SRA, they are cleaned and parsed to be compliant with MIxS metadata standards. Redundant fields are unified into a single field name, while fields with categorical inputs are parsed where possible to these values, rather than free text entries. Numerical fields that are gathered as free text, such as latitude and longitude, are parsed into numeric values as well. Finally, some fields with commonly missing metadata can be inferred from the other metadata context: investigation_type, an essential MIxS standard field, is often not provided by the user but can be predicted by logistic regression with 94.1% accuracy from the library_source, library_strategy, library_screening_strategy, and study_type fields.
The cleaned metadata are stored in the MetaSeek database, which is wrapped with an API implemented in Python's Flask library. The API interfaces communication between the database server and the MetaSeek web front-end, which is implemented in React, a popular JavaScript library for interactive web applications. The MetaSeek database, API, and front-end are hosted on an Amazon EC2 server.

Interfacing With MetaSeek

MetaSeek search, filter, and download functionality can be accessed via both the interactive online interface and a programmatic API. While the web interface emphasizes ease of use, the API emphasizes flexibility and comprehensiveness. Together, the online interface and API meet the needs of both casual and in-depth users.

The Online Interface. The main search and filter functionality is provided on the “Explore” page (www.metaseek.cloud/explore). A filter panel provides intuitive filter options for the most useful MetaSeek database fields. As users enter filter parameters, summary information of the datasets matching these filter parameters, such as counts of categorical variables, histograms of numeric fields, and a geographic map of dataset origins, are shown in real-time in an interactive visualization dashboard.

Users can save their configured filter parameters as “discoveries”, where they are accessible at a later date, shareable with other users, or able to be referenced in a publication. From the “Discovery Details” page, .csv files of either matching dataset IDs or all available metadata for each matching dataset can be downloaded directly. All user discoveries are made public and can be browsed on the “Browse” page, where previously saved discoveries can be used as a launching off point for other users. A video demonstrating the use of the MetaSeek website to identify marine metagenomes is linked to in the MetaSeek documents (https://github.com/MetaSeek-Sequencing-Data-Discovery/metaseek).

The MetaSeek API. The MetaSeek API provides a programmatic interface for querying the MetaSeek database. It is programming language agnostic and can be accessed via any HTTP POST request. The core API calls, SearchDatasetIds and SearchDatasetMetadata, take a set of filter parameters as input and return either a list of matching dataset IDs (SearchDatasetIds) or the full metadata (SearchDatasetMetadata) for every matching dataset. Filter parameters are flexible, such that any field in the MetaSeek database can be filtered by any value provided by the user, in any of 10 ways called “rule types”. Rule types are indicated by the user by an integer corresponding to the desired rule type, which consist of: “greater than”, “less than”, “greater than or equal to”, “less than or equal to”, “is equal to”, “is not equal to”, “is equal to any of a list of items”, “is not equal to any of a list of items”, “contains the partial text”, and “is not null”. Specific examples and tutorials for using the MetaSeek API are available in the API documentation (https://github.com/MetaSeek-Sequencing-Data-Discovery/metaseek/blob/master/APIdocs.md), and a glossary describing every MetaSeek field is also available on the website (www.metaseek.cloud/glossary).

Conclusions

MetaSeek fills a growing need in the bioinformatics community for faster, easier, and more accurate data discovery and integration. Future development will focus on curation of metadata from additional sequencing data repositories, direct integration with bioinformatics tools, and metadata inference from unstructured text data. Feature requests and input from the community are welcome and can be submitted via the website or directly to metaseek.cloud@gmail.com. Future updates and feature additions will be announced on the MetaSeek website.

Funding

This work has been supported by Gordon and Betty Moore Foundation GBMF4551 to CTB, NSF OCE-1736772 to CA, and a Deep Carbon Observatory Deep Life Modeling & Visualization Graduate Fellowship to AH.

Conflict of Interest

None declared.

Figures

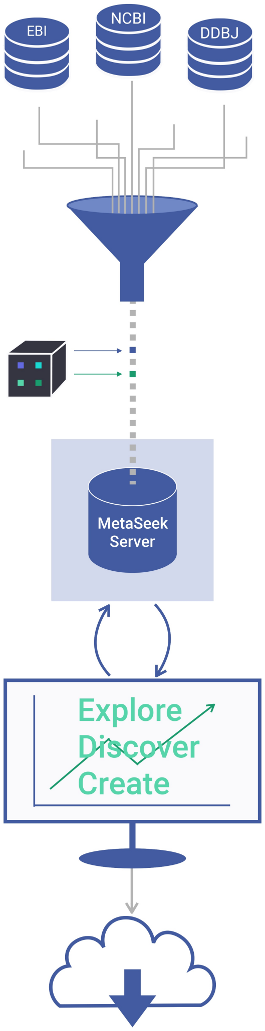

Figure 1 - The MetaSeek workflow.

References

NCBI Resource Coordinators. (2013). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 41, D8–D20. http://doi.org/10.1093/nar/gks1189.

Sayers E. (2017). The E-utilities In-Depth: Parameters, Syntax and More. 2009 May 29 [Updated 2017 Nov 1]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US).

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol, 29, 415–420.

Zhu Y, Stephens RM, Meltzer PS, Davis SR. (2013). SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics, 14(19).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27804v1 | CC BY 4.0 Open Access | rec: 16 Jun 2019, publ: 16 Jun 2019