BASIC PROTOCOL 2: Download MIDAS Reference Database

In 1 collection

Jul 29, 2022

miriam.goldman [1,2], chunyu.zhao [3,4]

[1]Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA,;

[2]Biomedical Informatics, University of California San Francisco, San Francisco, CA;

[3]Data Science, Chan Zuckerberg Biohub, San Francisco, CA, USA,;

[4]Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA

1 *Works for me* Share

dx.doi.org/10.17504/protocols.io.kxygxz6xwv8j/v1

miriam.goldman

ABSTRACT

This protocol describes how to download all or part of a MIDASDB, a set of custom files constructed from microbial genome sequences and containing all the information needed to metagenotype the species detected in a set of shotgun-metagenomic samples. MIDAS2 provides two prebuilt MIDASDBs sourced from large, public microbial genome collections: MIDASDB-UHGG (4,644 species / 286,997 genomes) based on the Unified Human Gastrointestinal Genome catalog (v1) [9] and MIDASDB-GTDB (47,893 species / 258,405 genomes) based on the Genome Taxonomy Database (v202) [10]. Support Protocol 2 describes how to build a new MIDASDB locally from a custom genome collection. A MIDASDB should be downloaded or built before any other MIDAS2 protocols can be run.

There are three components in a MIDASDB: single-copy marker genes (SCGs), representative genomes (rep-genome), and pangenomes (pan-genome). Each species contributes sequences to all three components. By preloading the MIDASDB, individual calls to MIDAS2 commands do not need to automatically download the necessary files. As a result, with a preloaded MIDASDB, per-sample analyses can be run in parallel without a risk of processes interfering with one another.

DOI

dx.doi.org/10.17504/protocols.io.kxygxz6xwv8j/v1

PROTOCOL CITATION

miriam.goldman , chunyu.zhao 2022. BASIC PROTOCOL 2: Download MIDAS Reference Database. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.kxygxz6xwv8j/v1

COLLECTIONS

MIDAS 2 Protocol

CREATED

Jul 28, 2022

LAST MODIFIED

Jul 29, 2022

PROTOCOL INTEGER ID

67832

PARENT PROTOCOLS

In steps of

BASIC PROTOCOL 3: Population Single Nucleotide Variant Calling

BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling

Part of collection

MIDAS 2 Protocol

1 Initialize a local copy of MIDASDB-UHGG

```
midas2 database --init --midasdb_name uhgg --midasdb_dir
midasdb_uhgg
```

This command creates the local directory midasdb_uhgg/ if it doesn't exist, and downloads the following files/directories:

- genomes.tsv: the table-of-contents file assigning genomes to species and denoting the representative genome for each species.
- metadata.tsv: tab-deliminated table specifying the six-digit numeric species identifiers (species_id) with taxonomic assignments.

Citation: miriam.goldman , chunyu.zhao BASIC PROTOCOL 2: Download MIDAS Reference Database https://dx.doi.org/10.17504/protocols.io.kxygxz6xwv8j/v1

- md5sum.json: md5sum cache for database files; used internally by MIDAS2 during downloading.
- markers/: SCG data needed for species prescreening.
- markers_models/: SCG profile hidden Markov model.
- chunks/: design cache for parallelizing the SNV module over partitions of the representative genome (“chunks”).

2 Customize the MIDASDB download. In Basic Protocol 1, 22 species were present in at least one sample list_of_species.tsv). Now those will be downloaded in the database components (both rep-genome and pan-genome) only for these 22 species.

```
midas2 database --download --midasdb_name uhgg --midasdb_dir
midasdb_uhgg --species_list list_of_species.tsv
```

3 The download has completed successfully when the command midas2 database --download finishes and no error is reported.

Citation: miriam.goldman , chunyu.zhao BASIC PROTOCOL 2: Download MIDAS Reference Database https://dx.doi.org/10.17504/protocols.io.kxygxz6xwv8j/v1