

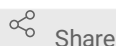


Jul 29, 2022

# BASIC PROTOCOL 1: Species Prescreening

 In 1 collectionmiriam.goldman<sup>1,2</sup>, chunyu.zhao<sup>3,4</sup><sup>1</sup>Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA,;<sup>2</sup>Biomedical Informatics, University of California San Francisco, San Francisco, CA;<sup>3</sup>Data Science, Chan Zuckerberg Biohub, San Francisco, CA, USA,;<sup>4</sup>Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA

1 Works for me

[dx.doi.org/10.17504/protocols.io.j8nlkkqj5l5r/v1](https://dx.doi.org/10.17504/protocols.io.j8nlkkqj5l5r/v1) miriam.goldman

## ABSTRACT

Reference-based metagenotyping depends crucially on the choice and customization of reference database. Therefore, a typical MIDAS2 workflow starts with a species prescreening step for each metagenome, which enables customization of the reference database to match the species in the sample. This protocol describes the species selection step: estimating species coverage per sample, merging the single-sample profiling results, and generating a list of species confidently detected in at least one sample. MIDAS2 estimates species coverage per sample by aligning reads to a database of sequences of 15 universal, single-copy genes (SCGs) and using the median (or mean) coverage of each species' SCGs.

## DOI

[dx.doi.org/10.17504/protocols.io.j8nlkkqj5l5r/v1](https://dx.doi.org/10.17504/protocols.io.j8nlkkqj5l5r/v1)

## PROTOCOL CITATION

miriam.goldman , chunyu.zhao 2022. BASIC PROTOCOL 1: Species Prescreening. **protocols.io**  
<https://dx.doi.org/10.17504/protocols.io.j8nlkkqj5l5r/v1>

## COLLECTIONS

**MIDAS 2 Protocol**

## LICENSE

————— This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

CREATED

Jul 21, 2022

LAST MODIFIED

Jul 29, 2022

PROTOCOL INTEGER ID

67273

PARENT PROTOCOLS

In steps of

[BASIC PROTOCOL 3: Population Single Nucleotide Variant Calling](#)

[BASIC PROTOCOL 3: Population Single Nucleotide Variant Calling](#)

[BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling](#)

[BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling](#)

Part of collection

[MIDAS 2 Protocol](#)

- 1 Install MIDAS2 (See Support Protocol 1)
- 2 Create a work folder containing the FASTQ files (here example input files are downloaded from Zenodo)

```
mkdir midas2_protocol
cd midas2_protocol
wget https://zenodo.org/record/6774633/files/reads.zip
unzip reads.zip
```

- 3 Initialize a local copy of a MIDAS Reference Database (MIDASDB). Here the SCG data from the UHGG MIDASDB is downloaded:

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

- 4 Run the single-sample species analysis to identify confidently detectable (i.e., relatively abundant) species in each sample, looping through samples. The output file is created automatically under the directories midas2\_output/SRR172902/species and midas2\_output/SRR172903/species

```
for sample_name in SRR172902 SRR172903  
do  
  midas2 run_species --sample_name ${sample_name} \  
  -1 reads/${sample_name}.fastq.gz \  
  --midasdb_name uhgg --midasdb_dir midasdb_uhgg \  
  --num_cores 4 midas2_output  
Done
```

- 5 Prepare the sample manifest file for the purpose of merging metagenotyping results across samples in the SNV and CNV modules. Generate the desired sample manifest file for SRR172902 and SRR172903.

```
echo -e "sample_name\tmidas_outdir" > list_of_samples.tsv  
ls reads | awk -F '.' '{print $1}' | awk -v OFS='\t' '{print $1,  
"midas2_output"}' >> list_of_samples.tsv
```

- 6 Merge species profiling results for the samples listed in the list\_of\_samples.tsv. The --min\_cov flag defines the minimum median\_marker\_coverage for estimating species prevalence. The output files are created automatically under the directory midas2\_output/merge/species.

```
midas2 merge_species --samples_list list_of_samples.tsv --min_cov 0.01
midas2_output/merge
```