Jul 29, 2022

BASIC PROTOCOL 3: Population Single Nucleotide Variant Calling

In 1 collection

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.81wgb638nlpk/v1

miriam.goldman

ABSTRACT

This protocol describes the SNV module of MIDAS2, which takes as input metagenomic sequencing reads from a set of samples and generates files with SNV genotypes for each sample for all detected species. The SNV module has two steps: (1) single-sample allele tallying with the midas2 run_snps command and (2) population SNV calling with the midas2 merge_snps command. Basic Protocols 1 (Species) and 2 (MIDASDB) should be run before this protocol.

DOI

dx.doi.org/10.17504/protocols.io.81wgb638nlpk/v1

PROTOCOL CITATION

miriam.goldman , chunyu.zhao 2022. BASIC PROTOCOL 3: Population Single Nucleotide Variant Calling. **protocols.io** https://dx.doi.org/10.17504/protocols.io.81wgb638nlpk/v1

COLLECTIONS ⓘ

MIDAS 2 Protocol

LICENSE

CREATED

Jul 28, 2022

LAST MODIFIED

Jul 29, 2022

PROTOCOL INTEGER ID

67833

PARENT PROTOCOLS

Part of collection

MIDAS 2 Protocol

1 Perform species prescreening as described in Basic Protocol 1.

BASIC PROTOCOL 1: Species Prescreening
by miriam.goldman

PREVIEW RUN

2 Download MIDASDB as described in Basic Protocol 2.

BASIC PROTOCOL 2: Download MIDAS Reference Database
by miriam.goldman

PREVIEW RUN

3 Execute the run_snps command for each sample.

Conceptually, a typical invocation of the run_snps command proceeds by four steps:

1. Select the list of species abundant enough for accurate metagenotyping based on the species profiling results and user- defined species selection criterion. Taking SRR172902 as an example, run_snps expects to find the species profiling results at midas2_output/SRR172902/species/species_profile.tsv.
2. Compile the representative genomes for these species and build a sample-customized rep-genome bowtie2 index.
3. Align reads to this index with bowtie2
4. Output a read alignment summary and pileup result for each species.

Citation: miriam.goldman , chunyu.zhao BASIC PROTOCOL 3: Population Single Nucleotide Variant Calling https://dx.doi.org/10.17504/protocols.io.81wgb638nlpk/v1

```
for sample_name in SRR172902 SRR172903
do
    midas2 run_snps \
    --sample_name ${sample_name} \
    -1 reads/${sample_name}.fastq.gz \
    --midasdb_name uhgg --midasdb_dir midasdb_uhgg \
    --select_by median_marker_coverage,unique_fraction_covered \
    --select_threshold=0,0.6 \
    --num_cores 8 midas2_output
done
```

The number of CPUs used is specified via --num_cores 8. This step can also be parallelized over multiple samples (e.g., using shell background processes or xargs).

4 Prepare sample manifest file for merging pileup results across samples. We can use the same file list_of_samples.tsv generated by step 6 in Basic Protocol 1.

BASIC PROTOCOL 1: Species Prescreening PREVIEW RUN
by miriam.goldman

5 Upon the completion of run_snps for all the samples in the file list_of_samples.tsv, MIDAS2 compute the population SNVs with the merge_snps command.

1. There are three main steps for each species:
2. For each genomic site in the representative genomes, MIDAS2 determines the set of alleles present across all samples where the species is detected.
3. For each genomic site, population major and minor alleles are then identified based either on the accumulated reads counts or sample counts in step (1). The population major allele is the allele with highest frequency across samples, and the population minor allele is the second most frequent. In the case of ties, the alphabetically first allele is the major allele.
4. Finally, MIDAS2 reports the vertical coverage (read depth) and population minor allele frequency of each site in each sample.

Citation: miriam.goldman , chunyu.zhao BASIC PROTOCOL 3: Population Single Nucleotide Variant Calling https://dx.doi.org/10.17504/protocols.io.81wgb638nlpk/v1

```
midas2 merge_snps --samples_list list_of_samples.tsv \
--midasdb_name uhgg --midasdb_dir midasdb_uhgg \
--genome_coverage 0.4 --genome_depth 0.1 --sample_counts 2 \
--snp_type bi --num_cores 8 midas2_output/merge
```

This command selects species present in both samples with horizontal genome coverage > 40% and average vertical genome coverage > 0.1X. There are four species meeting these selection criteria. Only bi-allelic SNVs are metagenotyped (--snp_type bi). The number of CPUs used is specified via --num_cores 8. These parameters are all adjustable.

6 Population SNV analysis has finished successfully when all the following output files are created under the directory midas2_output/merge/snps/ without any error message.

- snps_summary.tsv: merged single-sample pileup summary containing information such as horizontal genome coverage (fraction_covered) and vertical genome coverage (mean_coverage) for each sample.

For each species passing the species selection filter, information about SNVs identified across samples are organized by species_id , with three LZ4 files per subdirectory:

- <species_id>/<species_id>.snps_info.tsv.lz4: metadata of population SNVs (e.g., biological annotations)
- <species_id>/<species_id>.snps_allele_freq.tsv.lz4: site-by-sample matrix of population minor allele frequencies
- <species_id>/<species_id>.snps_depth.tsv.lz4: site-by-sample read depth matrix