BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling

Jul 29, 2022

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

dx.doi.org/10.17504/protocols.io.n92ldzke7v5b/v1

miriam.goldman

ABSTRACT

This protocol describes the CNV module of MIDAS2, which takes as input metagenomic sequencing reads from a set of samples and generates files with CNV genotypes for each sample for all detected species. There are two steps for population CNV calling: (1) single-sample quantification of copy number for each gene in the pangenome of each species with the midas2 run_genes command and (2) population CNV calling with the midas2 merge_genes command. Basic Protocols 1 (Species) and 2 (MIDASDB) should be run before this protocol.

DOI

dx.doi.org/10.17504/protocols.io.n92ldzke7v5b/v1

PROTOCOL CITATION

miriam.goldman , chunyu.zhao 2022. BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling. **protocols.io** https://dx.doi.org/10.17504/protocols.io.n92ldzke7v5b/v1

COLLECTIONS

MIDAS 2 Protocol

CREATED

Jul 28, 2022

LAST MODIFIED

Jul 29, 2022

PROTOCOL INTEGER ID

67835

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_genes command for each sample

1. Conceptually, a typical invocation of the run_genes command proceeds by four steps:
2. 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_genes expect to find the species profiling results at midas2_output/SRR172902/species/species_profile.tsv.
3. Compile the pangenomes for these species and build a sample-customized pan-genome bowtie2 index.
4. Align reads to this index with bowtie2.
5. For each gene in the pan-genome, normalize gene coverage by the mean coverage of all that species' SCGs to estimate copy number per cell [11].
6. Output a read mapping summary and CNV estimates for each species.

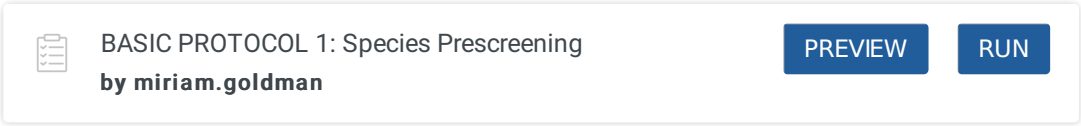
Citation: miriam.goldman , chunyu.zhao BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling
https://dx.doi.org/10.17504/protocols.io.n92ldzke7v5b/v1

midas2 run_genes

```
for sample_name in SRR172902 SRR172903
do
   midas2 run_genes \
   --sample_name ${sample_name} \
   -1 reads/${sample_name}.fastq.gz \
   --midasdb_name uhgg --midasdb_dir midasdb_uhgg \
   --species_list 100122,100277 \
   --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.

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



5 Upon the completion of run_genes for all the samples listed in the list_of_samples.tsv, MIDAS2 merges the CNV profiles across samples with the merge_genes command.

```
midas2 merge_genes --samples_list list_of_samples.tsv \
--midasdb_name uhgg --midasdb_dir midasdb_uhgg \
--min_copy 0.5 \
--num_cores 2 midas2_output/merge
```


Citation: miriam.goldman , chunyu.zhao BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling
https://dx.doi.org/10.17504/protocols.io.n92ldzke7v5b/v1

6 Population pangenome CNV analysis has finished successfully when all the following output files are created under the directory midas2_output/merge/genes/ without any error message

- genes_summary.tsv: merged single-sample CNV summary containing information such as mean_coverage.
- <species_id>/<species_id>.genes_copynum.tsv.lz4: gene-by-sample matrix of copy-number estimates
- <species_id>/<species_id>.genes_preabs q.tsv.lz4: gene-by-sample matrix of gene presence / absence
- <species_id>/<species_id>.genes_depth.tsv.lz4: gene-by-sample read coverage matrix
- <species_id>/<species_id>.genes_reads.tsv.lz4: gene-by-sample read counts matrix

Citation: miriam.goldman , chunyu.zhao BASIC PROTOCOL 4: Pan-genome Copy Number Variant Calling
https://dx.doi.org/10.17504/protocols.io.n92ldzke7v5b/v1