RevMet Documentation

Release 0.0.1

Ned Peel

Mar 10, 2021

Contents

CHAPTER 1

Tutorial

This tutorial will run the RevMet method on a set of 600 nanopore reads using Illumina genome skims for 7 species at around 0.2x coverage. The coverage level has been reduced in order to make the tutorial run quicker - for real data, we recommend higher coverage.

The example below has been designed to run on a laptop/desktop computer, with alignments running in series. Much greater performance can be achieved by running in parallel on High Performance Computing (HPC).

On a 2017 2.9Ghz i7 quad core MacBook Pro, this example takes around 8 minutes to run.

A number of dependencies should be installed prior to this tutorial - , , and .

1. Download the example RevMet dataset from and uncompress:

```
tar -xvf revmet_example.tar.gz
```

2. Change into the directory:

```
cd revmet_example
```

3. Run the revmet example bash script:

```
sh revmet_example.sh
```

4. To view the read counts for each of the constituent species:

```
cat output/mock_mix_1_1_bin_counts.txt
```

5. To view the species composition of the sample as percentages:

```
cat output/mock_mix_1_1_bin_percentges.txt
```

1.1 How the example RevMet script works

1. At the top of the revmet_example.sh script, file location and mapping variables are assigned:

```
skim_refs_dir=skim_refs
nanopore_reads=nanopore_reads/mock_mix_1_1.fasta
output_dir=output
scripts_dir=scripts
mapq=0
include_flag_f=0
exclude_flag_F=2308
```

2. The script then loops through the reference genome skims and maps each of them to the long reads of a nanopore-sequenced sample, which for this example is mock_mix_1_1.fasta, a 600 read DNA mock mix subset. We are using here due to its speed. However, in our we assigned a greater number of nanopore reads and experienced fewer false positive results by mapping with using a strict MAPping Quality (MAPQ) of 60.

3. filters the alignment files based on the include (-f) and exclude (-F) flags set in the variables section. In this case we use exclude 2308, therefore SAMtools removes unmapped, secondary, and supplementary alignments (For more information, see).

4. SAMtools then sorts and indexes each alignment file before calculating the depth of mapping coverage at each long-read position using the SAMtools depth function.

5. A custom python script, 'percent_coverage_from_depth_file.py', uses these depth files to calculate 'percent coverage' for each long read, defined as the fraction of nucleotide positions that were mapped to by one or more reference-skim Illumina reads.

6. The percent coverage files, each of which contain the % coverage values for every mock mix 1.1 nanopore read from a particular skim dataset, are concatenated into one file. The python script 'minion_read_bin_from_perc_cov.py' uses this concatenated file to uniquely assign each nanopore read to the reference species that mapped with the highest % coverage.

7. The number of long reads binned to each species is counted with the 'minion_read_bin_counts_from_perc_cov_binned_with_threshold.py' script, which can also filter reads into "unassigned" based on % coverage thresholds. By default, if the highest percent coverage for a read is <15% its identity is judged to be ambiguous and it is left unassigned.

8. Finally, the 'convert_minion_read_counts_to_percentages.py' script implements a 1% minimum-abundance filter, which sets plant species represented by fewer than 1% of the total assigned long reads to zero before converting the remaining read counts to percentages. The minimum-abundance filter threshold can be altered with the "-t" flag.

RevMet (Reverse Metagenomics) is a method that allows reliable and semi-quantitative characterization of the species composition of mixed-species eukaryote samples, such as bee-collected pollen, without requiring assembled reference genomes. Instead, reference species are represented only by 'genome skims': low-cost, low-coverage, short-read datasets. The skims are mapped to long reads sequenced from mixed-species samples, using nanopore sequencing, and the long reads are uniquely assigned to eukaryote species.

CHAPTER 2

Example

- To learn how to implement the RevMet method, see the *Tutorial page*.