Mapping Metagenomic Reads to Reference Sequences (Cyverse)

Benjamin Bolduc

Abstract

Mapping metagenomic reads from Ocean Sampling Day (OSD) 2014 against NCBI's ViralRefSeq alongside viral sequences identified from the Tara Oceans survey using VirSorter.

Citation: Benjamin Bolduc Mapping Metagenomic Reads to Reference Sequences (Cyverse). **protocols.io** dx.doi.org/10.17504/protocols.io.evybe7w
Published: 22 Apr 2016

Guidelines

One of the most commonly used procedures for analyzing viral metagenomic data is to map their reads (or reads from another dataset) against a set of references, often those from the read assembly. For example, if one wanted to know how well-represented viruses in NCBI's Viral Reference Sequences (ViralRefSeq) were in ocean viromes, they could map reads from lots of ocean viral metagenomes against ViralRefSeq. This is generally done using Bowtie2 or BWA, by selecting a reference set of sequences, and then providing paired or unpaired reads to Bowtie2/BWA. *Then* the results must be processed/filtered to generate coverage tables. Dealing with setting up multiple reads files (10 paired metagenomes = 10 alignment runs) and the processing those read files can be challenging (not to mention requring computational resources).

In this protocol, we'□□ll be using reads fromÂ Ocean Sampling Day (OSD) 2014 and map them against ViralRefSeq alongside viral sequences identified from the Tara Oceans survey using VirSorter.

Before start

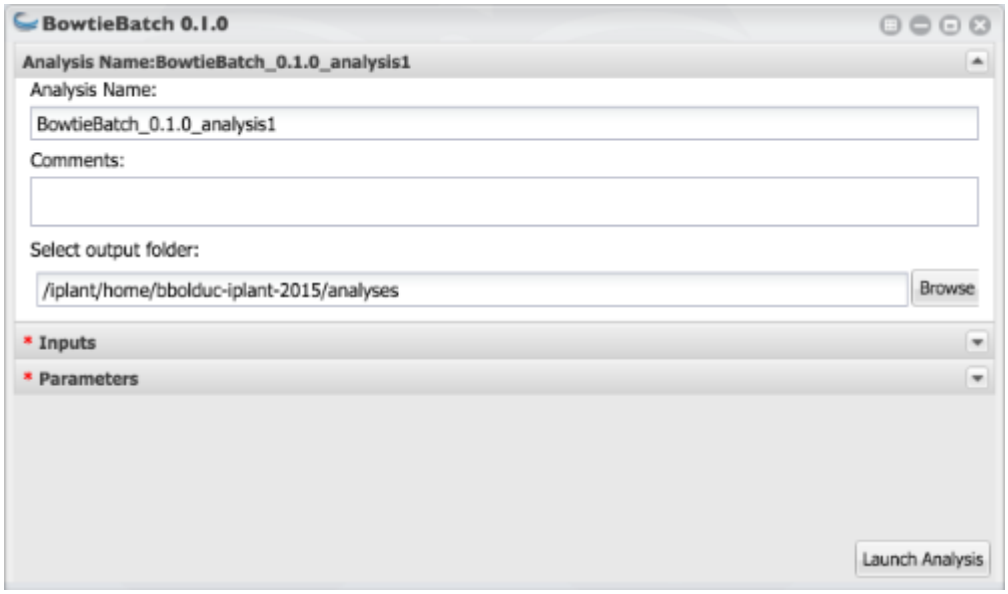
To run this protocol, users must first register for Cyverse account. All data (both inputs and outputs) are available within Cyverse's data store at /iplant/home/shared/iVirus/ExampleData/

Protocol

BowtieBatch
Step 1.

Open BowtieBatch

Open BowtieBatch from 'Apps'



BowtieBatch

Step 2.

Select Inputs

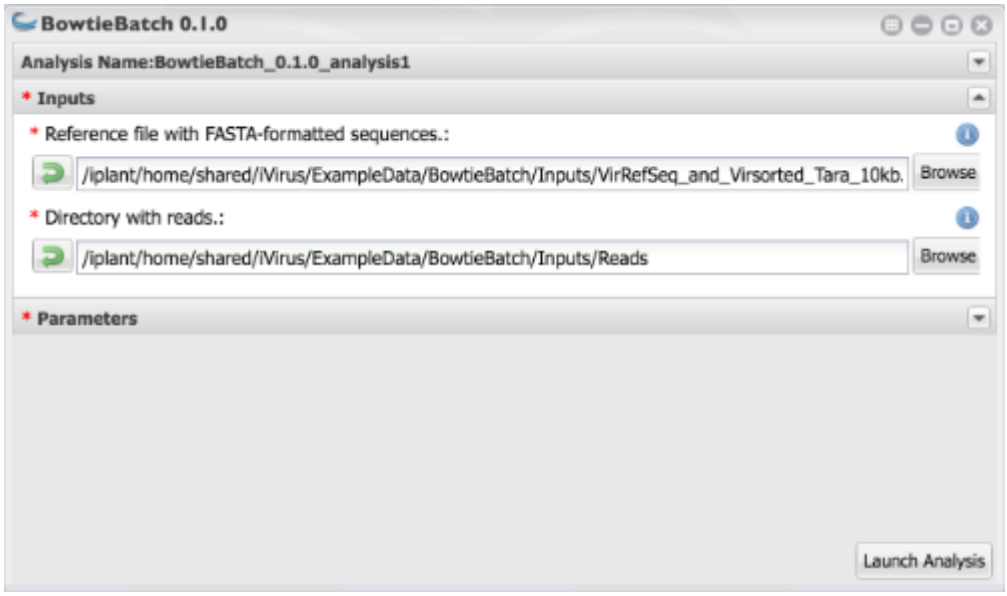
Select the 'Inputs' tab.

For **Reference file with FASTA-formatted sequences:** this is a FASTA-formatted file with reference sequences to map reads against

- Navigate to *Community Data --> iVirus --> ExampleData --> BowtieBatch --> Inputs*. Select *VirRefSeq_and_Virsorted_Tara_10kb.fasta* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/BowtieBatch/Inputs into the navigation bar and select the fasta file.

For **Directory with reads**: this is a directory containing the reads files that are to be mapped

- Navigate to *Community Data --> iVirus --> ExampleData --> BowtieBatch --> Inputs*. Select the *Reads* folder. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/BowtieBatch/Inputs into the navigation bar and select the reads folder.



BowtieBatch

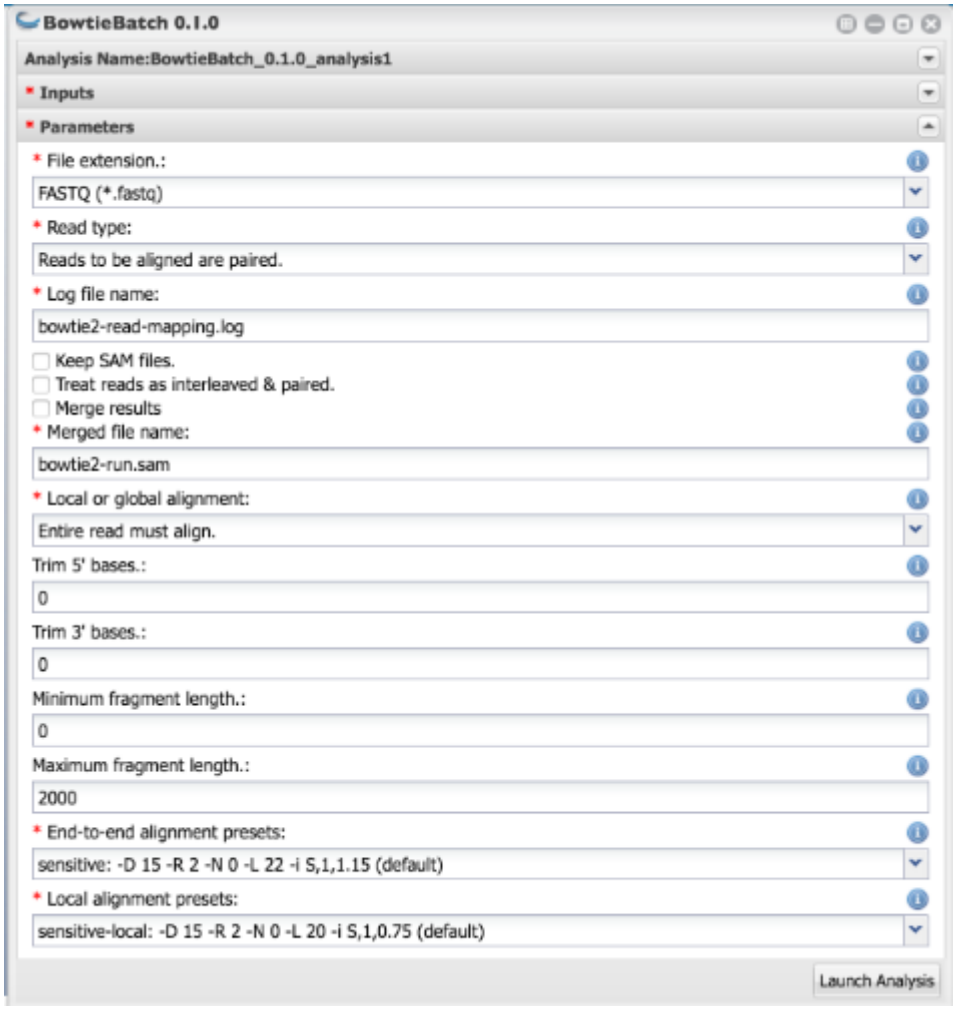
Step 3.

Select Parameters

The default options will be sufficient.

Read type: What type are the reads? Paired, unpaired?
Log file name: Information about the run's processing will be stored with this file name.
Keep SAM files:Bowtie2 generates SAM files as a result of mapping reads. However, these take up a lot of space and aren't kept if they're not needed. Users can select if they'd like to keep these files around in case they have other analyses that use SAM files. For this example, they are unnecessary.
Treat reads as interleaved & paired: Often times sequencing centers send *interleaved* reads to users. These read files contain both pairs, usually Forward1, Reverse1, Forward2, etc... By using this option users can automatically handle interleaved files w/out worrying about having another app handling this conversion.
Merge results: Bowtie2 can create a single SAM file that combines all the individual paired/unpaired alignments. Because the next tool in this pipeline (Read2RefMapper) requires individually separated files, do not select this.

The remaining options are bowtie specific, please consult its documentation.



NOTES

Benjamin Bolduc 20 Apr 2016

While the default options will suffice for this example, play around with a few of the options (excluding file extension).

BowtieBatch

Step 4.

Launch Analysis

Run the job!

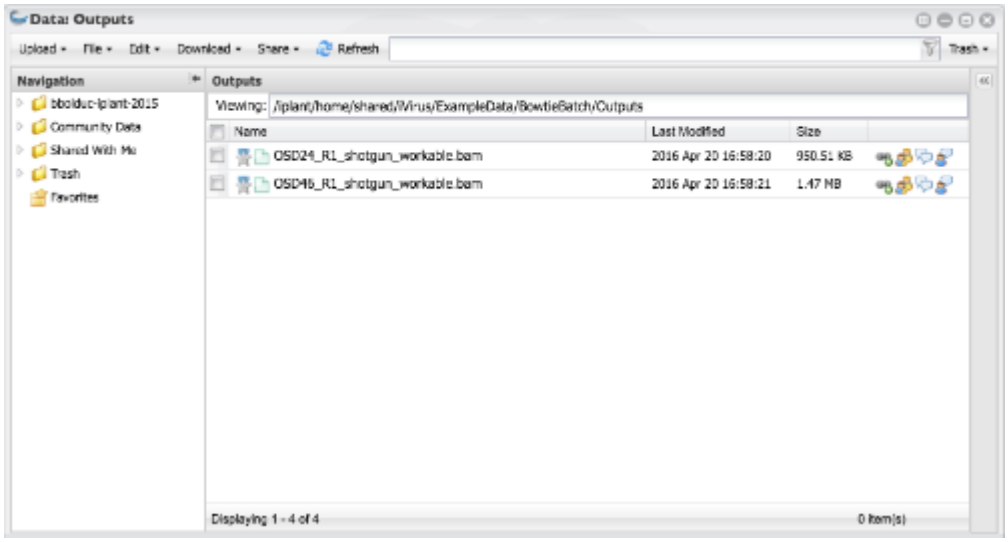
BowtieBatch

Step 5.

Results

Expected results can be found from the 'Outputs' directory of BowtieBatch. They'll consist of a directory containing BAM files. If **Keep Sam files** was selecting, they'll be SAM files there as well.

Navigate to *Community Data --> iVirus --> ExampleData --> BowtieBatch --> Outputs*. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/BowtieBatch/Outputs into the navigation bar.

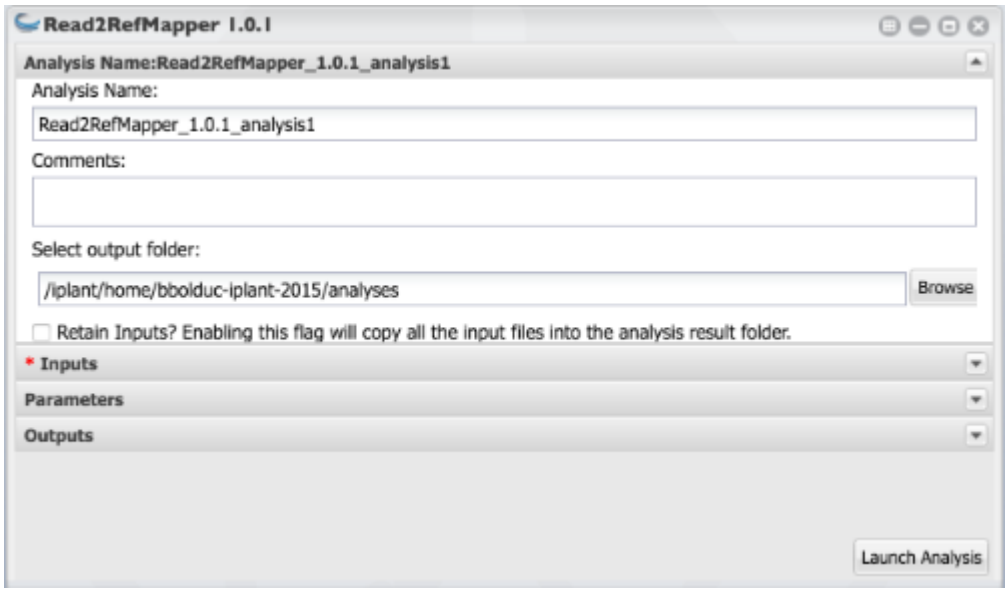


Read2RefMapper

Step 6.

Open Read2RefMapper

Open Read2RefMapper from "Apps"



Read2RefMapper

Step 7.

Select Inputs

Select the 'Inputs' tab.

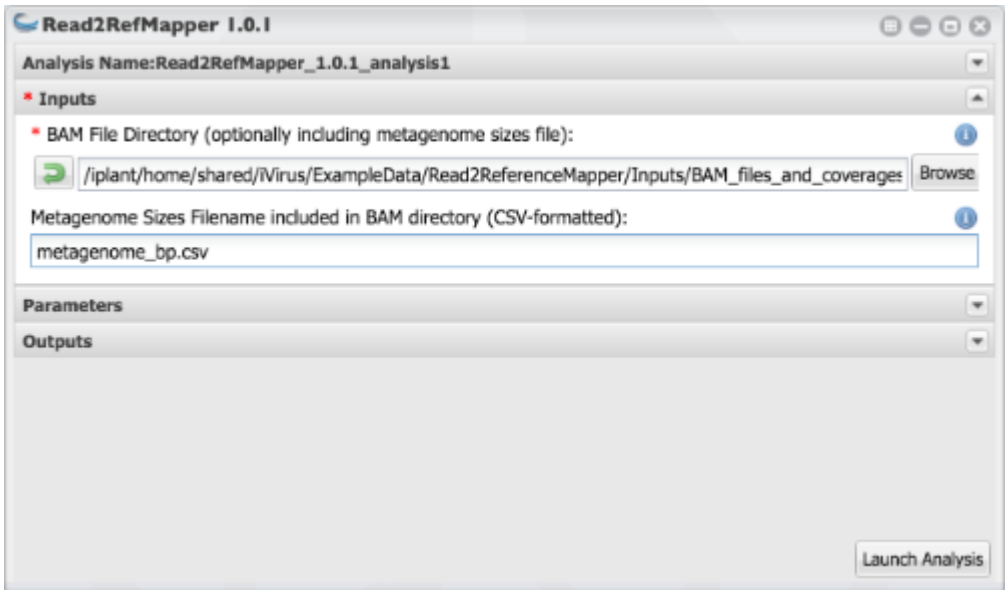
For **BAM File Directory (optionally including metagenome sizes file)**:

- Navigate to *Community Data* --> *iVirus* --> *ExampleData* --> Read2ReferenceMapper --> *Inputs*. Select the *BAM_files_and_coverages_file* folder. This will contain BAM files and a coverage file. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/Read2ReferenceMapper/Inputs into the navigation bar

and select the folder.

For **Metagenome Sizes Filename included in BAM directory (CSV-formatted)**:

- Type the filename of the metagenome sizes file, for this example, it's *metagenome_bp.csv.* When provided, the metagenome sizes are used to adjust for differential coverage resulting from differentially sized metagenomes. For non-example data, this file needs to be created by the user - as only they know the total size of their metagenomes.



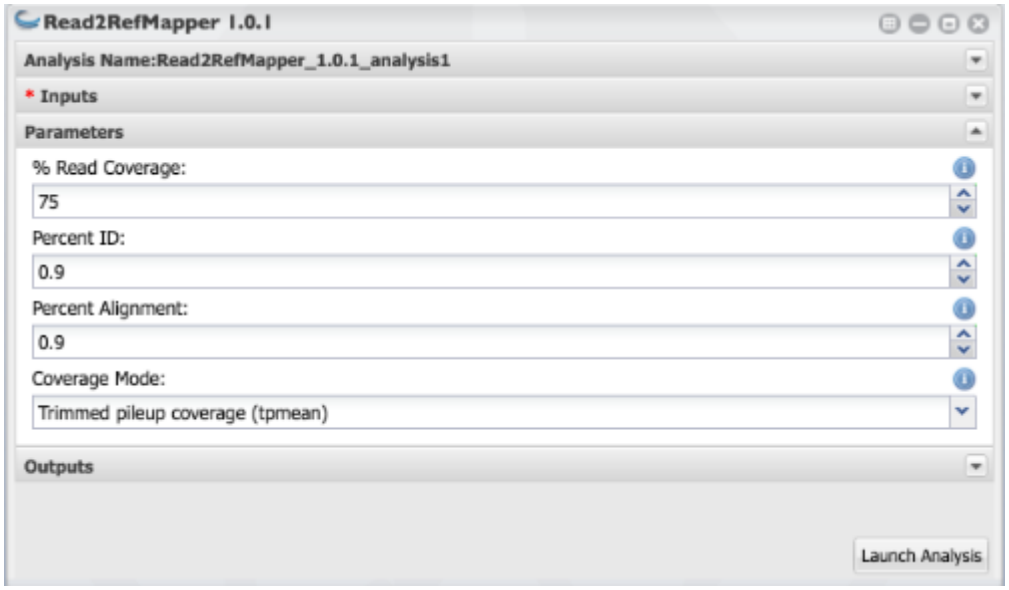
Read2RefMapper

Step 8.

Select Parameters

For this example, the defaults are sufficient.

% Read Coverage: Percent of the reference sequence that must be covered by reads. For example, if 75% of a reference must be covered by reads to be considered "present" in that sample.
Percent ID: Minimum identity a read must be against the reference sequence. For example, 0.90 is 90% identity between the read and reference.
Percent Alignment: What percent of the read length must be aligned to be considered matching. For example, 0.90 is 90% of the *read length* must align to the reference. This allows users to excludes reads where only a small local alignment is present.
Coverage Mode: How coverage should be calculated.



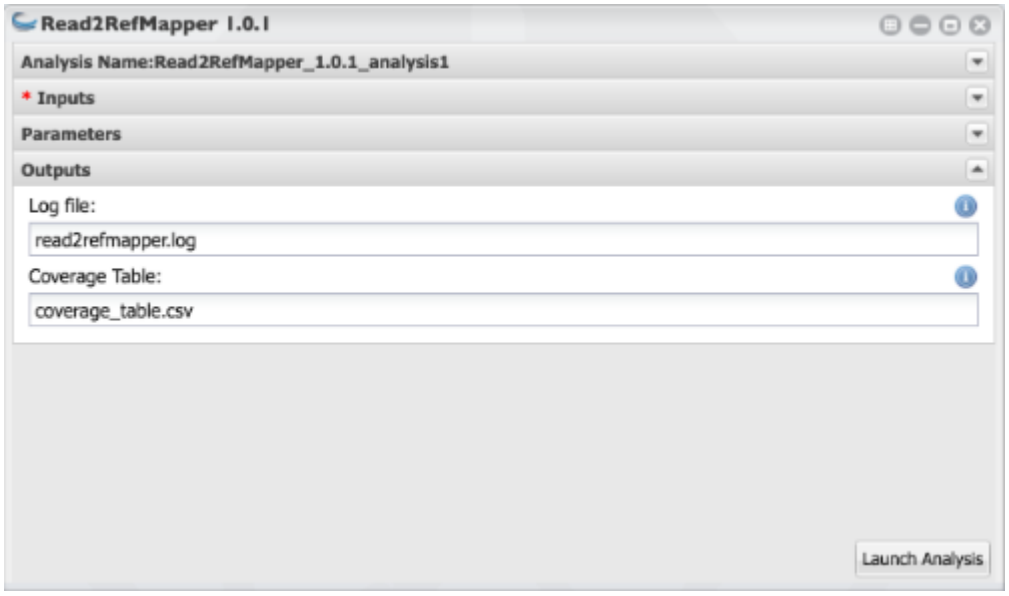
Read2RefMapper

Step 9.

Specify Outputs

Defaults will be sufficient for this example.

Log file: Filename to store logging information.
Coverage table: Name of the coverage table.



Read2RefMapper

Step 10.

Launch Analysis

Run the job!

Read2RefMapper

Step 11.

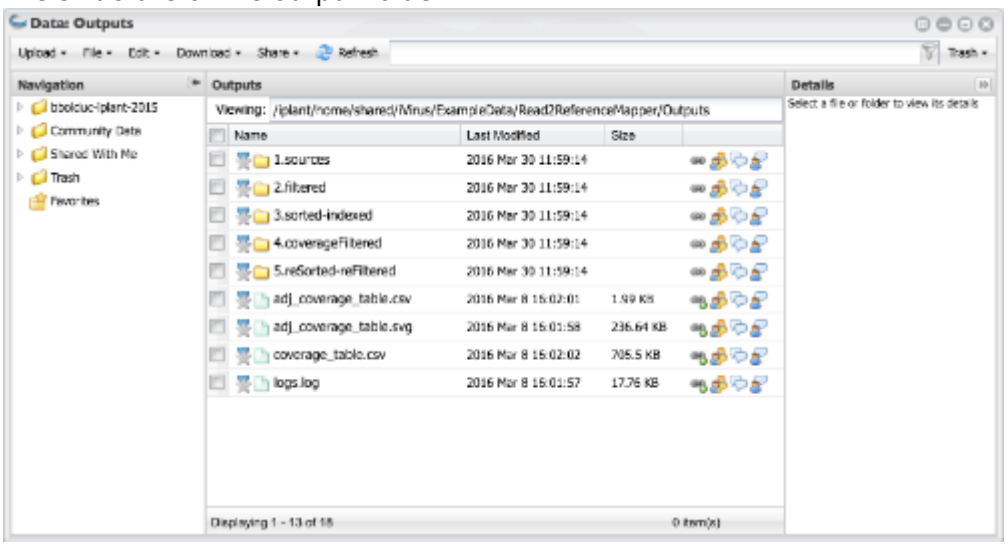
Results

Expect results can be found in the associated app's "Output" directory.

There will be several directories, corresponding to each step along the pipeline. Notably are the *coverage_table.csv*, which will have raw coverages for each contig across all metagenomes, and the *adj_coverage_table.csv*. In this example, we specified a metagenome bp file, meaning Read2RefMapper will adjust/normalize the coverage table according to the metagenome sizes. If no metagenome bp file is found, then the adjusted/normalized coverage table will not be generated!

EXPECTED RESULTS

File structure of the output folder.



A quickly drawn heatmap of the adjusted/normalized coverage table. **This is not suitable for publication-quality images**. This is only for a quick look.