VERSION 4

DEC 19, 2018

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.wigfcbw

External link:
<https://dx.doi.org/10.7717/peerj.3243>

Protocol Citation: Benjamin Bolduc 2018. Preparing Data for vContact from Proteins (Cyverse). **protocols.io** <https://dx.doi.org/10.17504/protocols.io.wigfcbw>

MANUSCRIPT CITATION:
 Bolduc B, Jang HB, Doulcier G, You Z, Roux S, Sullivan MB, vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect and . PeerJ doi: [10.7717/peerj.3243](https://doi.org/10.7717/peerj.3243)

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
 We use this protocol and it's working

🌐 Preparing Data for vContact from Proteins (Cyverse) V.4

📁 In 1 collection

Benjamin Bolduc¹

¹The Ohio State University

Sullivan Lab

iVirus

Benjamin Bolduc

ABSTRACT

Preparing data for use in vContact by using VirSorted [Ocean Sampling Day \(2014\)](#) contigs, using tools available in [Cyverse](#). This protocol creates a BLAST DB, BLASTs sequences, and creates a gene-to-contig mapping file. Results from this protocol are suitable for vContact-PCs.

GUIDELINES

This is part of a larger protocol *Collection* that involves the end-to-end processing of raw viral metagenomic reads obtained from a sequencing facility to assembly and analysis using Apps (i.e. tools) developed by iVirus and implemented within the Cyverse cyberinfrastructure.

BEFORE START INSTRUCTIONS

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/`

Starting at version 4, this method is dramatically simplified, as nearly all steps were integrated into vConTACT2's functionality. This results in a faster, easier-to-use, and less complicated method. Any user wishing to repeat experiments *exactly* as described in the original iVirus manuscript should run version 3 or earlier. However, except for very minor spelling changes, the results files are nearly identical, and the content 100% identical.

Created: Dec 19, 2018

Last Modified: Dec 19, 2018

PROTOCOL integer ID:
18728

Generating Gene-to-Genome Mapping

1 Open vContact2-Gene2Genome

Open "vContact2-Gene2Genome" from the "Apps" menu.

Starting menu for the Gene2Genome app in the CyVerse Discovery Environment.

2 Select Inputs

Select the 'Inputs' tab.

For **Proteins file**:

- Navigate to *Community Data* --> *iVirus* --> *ExampleData* --> *vContact2-Gene2Genome* --> *Input*. Select *VIRSorter_viral_prots.faa* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/vContact2-Gene2Genome/Input into the navigation

bar and select the protein fasta file.

Inputs section of Gene2Genome

3 Select Parameters

Under "**Input source type**" change to *VirSorter*. Users can select a number of different parsing formats depending on the ORF caller they used to generate their proteins. For this example, everything passed through VirSorter, so we'll use VirSorter's formatting convention to extract the contigs each ORF/gene derives.

Keep 'description' field with contig names: Some formats have descriptions in their fasta files. Flagging this option keeps those descriptions.

Enable vContact1 headers: vContact1 and vContact2 fundamentally use the same input information, but are formatted a little differently. *If you use vContact 1* you must select this box. Failure to do so will result in vContact 1, well, failing. If you are using vContact2, then keep this *unchecked*.

Output filename: Anything useful or descriptive.

Parameters section of Gene2Genome

4 Launch Analysis

Run the job!

This should take minutes. Depending on the queue in Cyverse, it will likely take longer to submit and start the job than it does to run it!

5 Results

Expected result

Expect results can be found in the vContact2-Gene2Genome 'Outputs' directory. They'll be 4 files: 2 with agave messages (the errors and outputs), the original proteins file (VIRSorter_viral_protos.faa) and the gene-to-genome mapping file.

Output files produced by the vContact2-Gene2Genome app

From this point, it's off to vConTACT2! The *gene2genome_proteins.csv* file and the *original proteins file* (used as input for this app) are the only hard requirements for vConTACT2.