Preparing Data for vContact from Proteins (Cyverse)

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

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

Generating Protein Clusters via BLASTp

Step 1.

Open 'Create BLAST Database'

Open 'Create BLAST Database' from the 'Apps' menu.

Generating Protein Clusters via BLASTp

Step 2.

Select Inputs

Select the 'Input Options' tab.

For **Input file**:

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

For **Input Sequence Format**:

- Select Protein. This is the format of the input sequences. In this case it's proteins.

For **Title for the database**:

- Title can be whatever you want. If users wish to download the database to use on their local machine, this may be useful.

All other options are irrevelent for this example.

Step 3.

Launch Analysis

Run the job!

This should only take a few minutes.

Step 4.

Results

Expect results can be found in the Create_BLAST_database 'Output' directory.

Step 5.

Open 'Blastp'

Open 'Blastp-2.2.29+' from the 'Apps' menu.
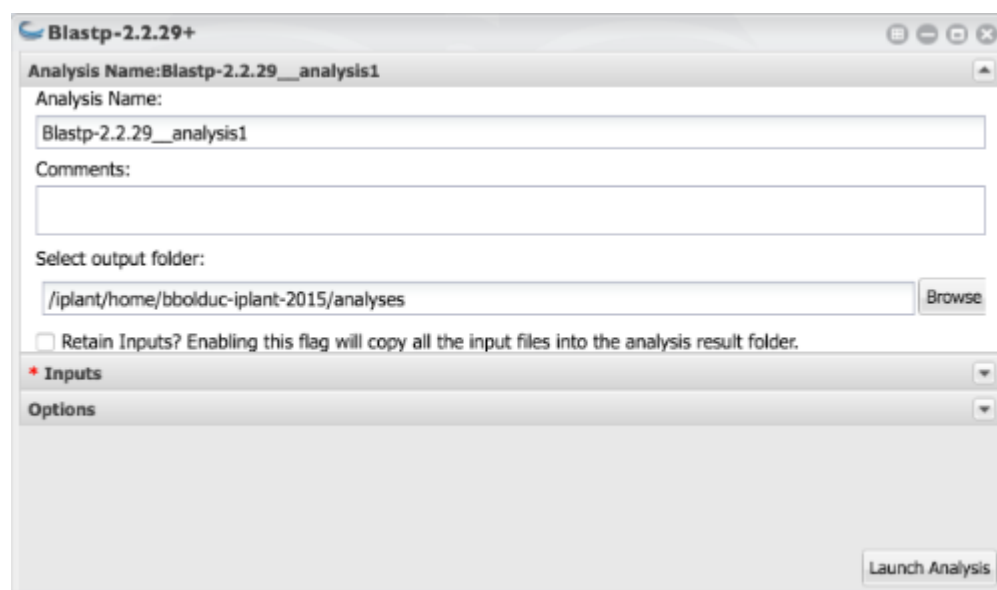
Generating Protein Clusters via BLASTp

Step 6.

Select Inputs

Select the 'Inputs' tab.

For **Query sequence**:

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

For **Database**:

- Navigate to *Community Data --> iVirus --> ExampleData --> blastp --> Inputs*. Select the *makeblastdb_dir* directory. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/blastp/Inputs into the navigation bar and select the directory.
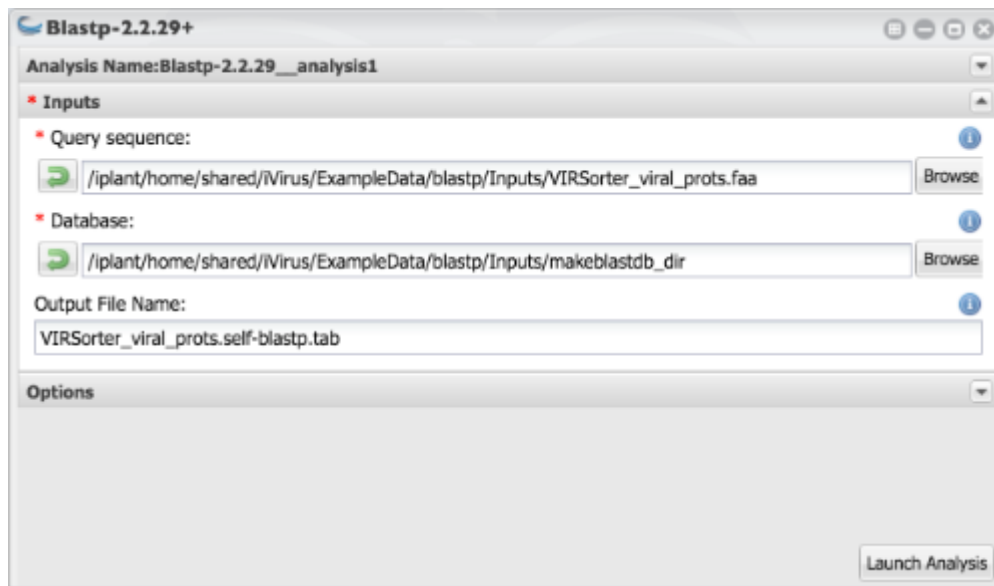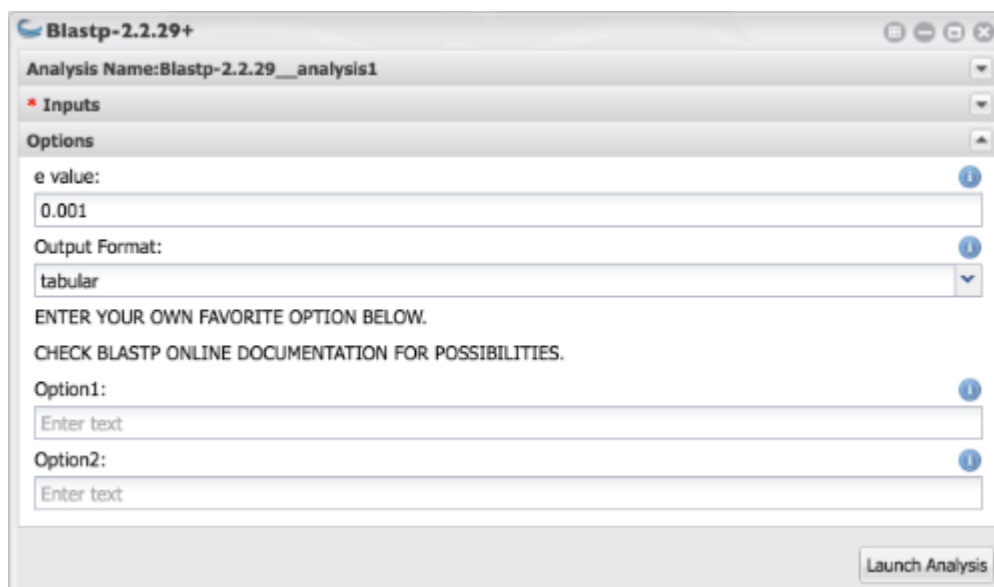
Generating Protein Clusters via BLASTp

Step 7.

Select Parameters

Under "**Output Format**" change to *tabular*. vContact PCs requires BLASTp files to be in this format.

E value can be adjusted from its default of 10.

All other options can be left as is.

Generating Protein Clusters via BLASTp

Step 8.

Launch Analysis

Run the job!

This is an all-verses-all BLASTp - that's every protein compared to all others, done for all proteins. This can concievably take many hours to days. This dataset is tiny, so it won't take more than a few minutes.

Step 9.

Results

Expect results can be found in the blastp 'Output' directory.

Step 10.

Open vContact-Gene2Contig

Open "vContact-Gene2Contig" from the "Apps" menu.
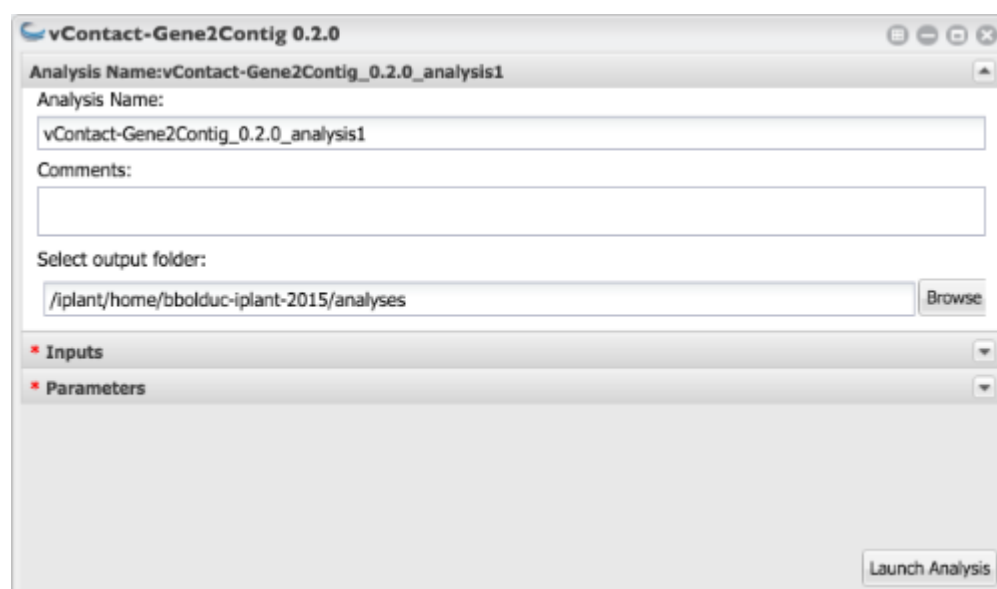
Step 11.

Select Inputs

Select the 'Inputs' tab.

For **Proteins file**:

- Navigate to *Community Data --> iVirus --> ExampleData --> vContact-Gene2Contig --> Input*. Select *VIRSorter_viral_prots.faa* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/vContact-Gene2Contig/Input into the navigation bar and select the protein fasta file.
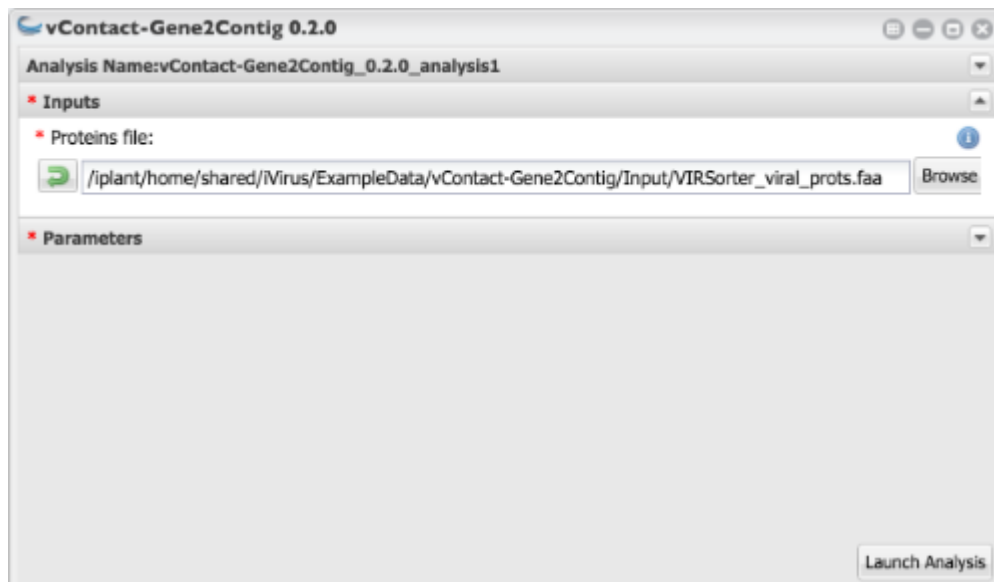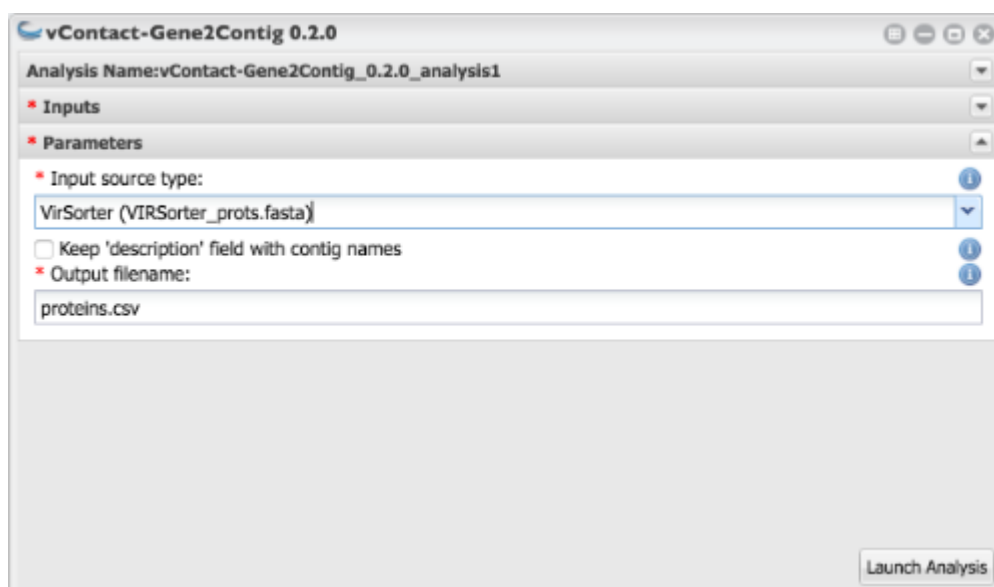
Generating Gene-to-Contig Mapping

Step 12.

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.

Generating Gene-to-Contig Mapping

Step 13.

Launch Analysis
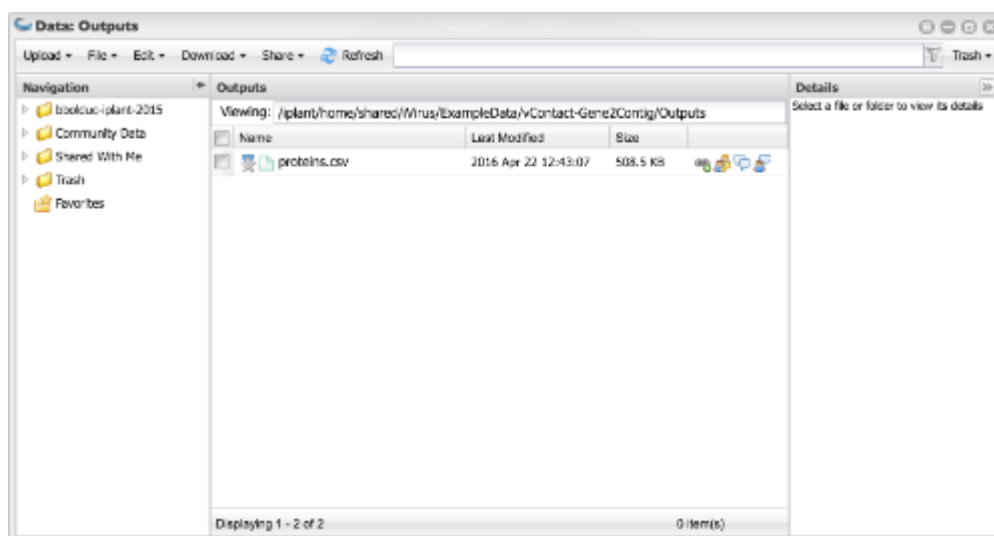
Run the job!

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!

Generating Gene-to-Contig Mapping

Step 14.

Results

Expect results can be found in the vContact-Gene2Contig 'Outputs' directory.
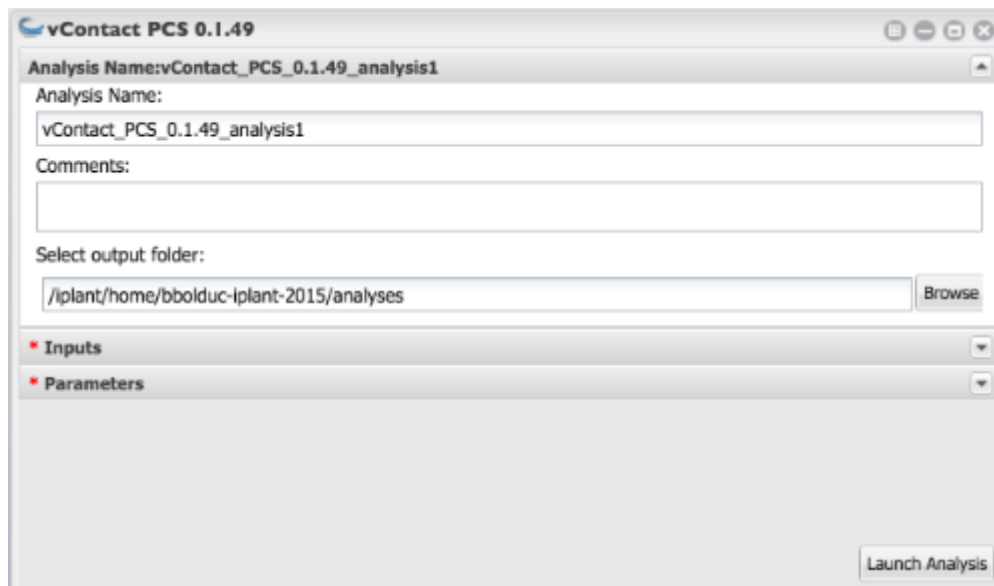

Preparing/formatting vContact inputs

Step 15.

Open vContact PCs

Open "vContact PCs" from the "Apps" menu.

Preparing/formatting vContact inputs
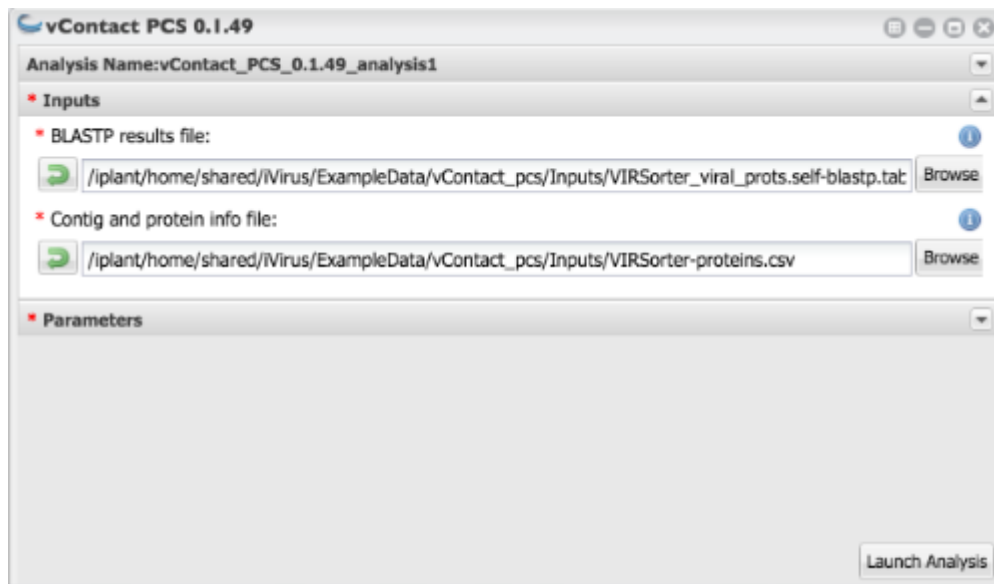
Step 16.

Select Inputs

Select the 'Inputs' tab.

For **BLASTP results file**:

- Navigate to *Community Data --> iVirus --> ExampleData --> vContact_pcs --> Inputs*. Select *VIRSorter_viral_prots.self-blastp.tab* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/SPAdes/vContact_pcs into the navigation bar and select the BLASTp file.

For the **Contig and protein info file**:

- Navigate to *Community Data --> iVirus --> ExampleData --> vContact_pcs --> Inputs*. Select VIRSorter-proteins.csv. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/SPAdes/vContact_pcs into the navigation bar and select the CSV file.

For **both of the above files**, you will be using files generated from the above TWO sections. That's *the BLASTP* file from Blastp-2.2.29+ and the *VIRSorter-proteins.csv* file from vContact-Gene2Contig.

Step 17.

Select Parameters

The only parameter is the **output file prefix**. This will be fine as is.

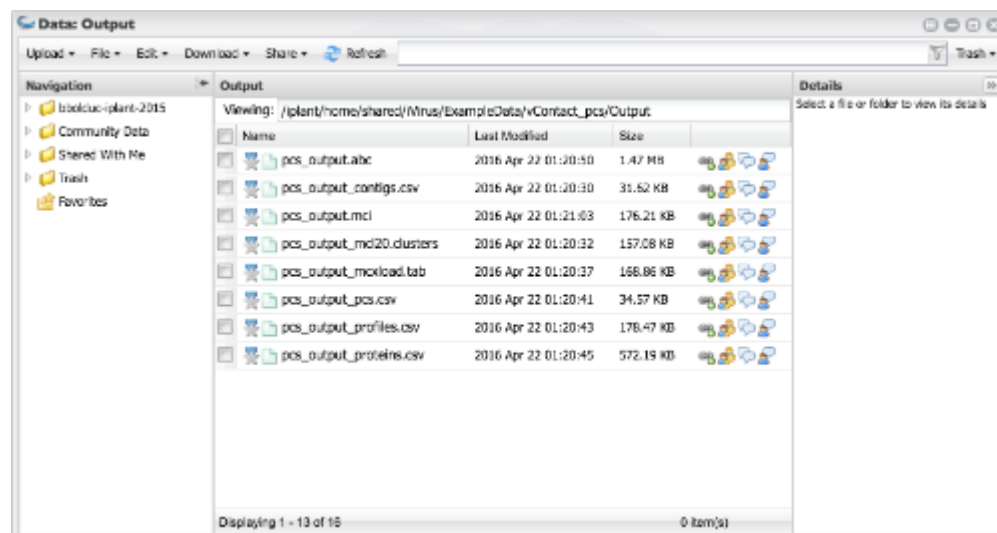**Step 18.**

Launch Analysis

Run the job! This should complete within a minute or two, depending on the size of your dataset. This example should take less than a minute *once running*.

Step 19.

Results

Expect results can be found in the vContact PCs 'Output' directory.

Results

Step 20.

Summarizing the Results

If everything above was done correctly, you should have a number of files, only THREE of which are necessary for vContact. These 3 files were generated by vContact PCs.

pcs_output_contigs.csv
pcs_output_profiles.csv
pcs_output_pcs.csv