Preparing Data for vContact from Proteins (Cyverse) Version 3

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.

Citation: Benjamin Bolduc Preparing Data for vContact from Proteins (Cyverse). **protocols.io**
dx.doi.org/10.17504/protocols.io.gwdbxa6
Published: 05 Jan 2017

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/

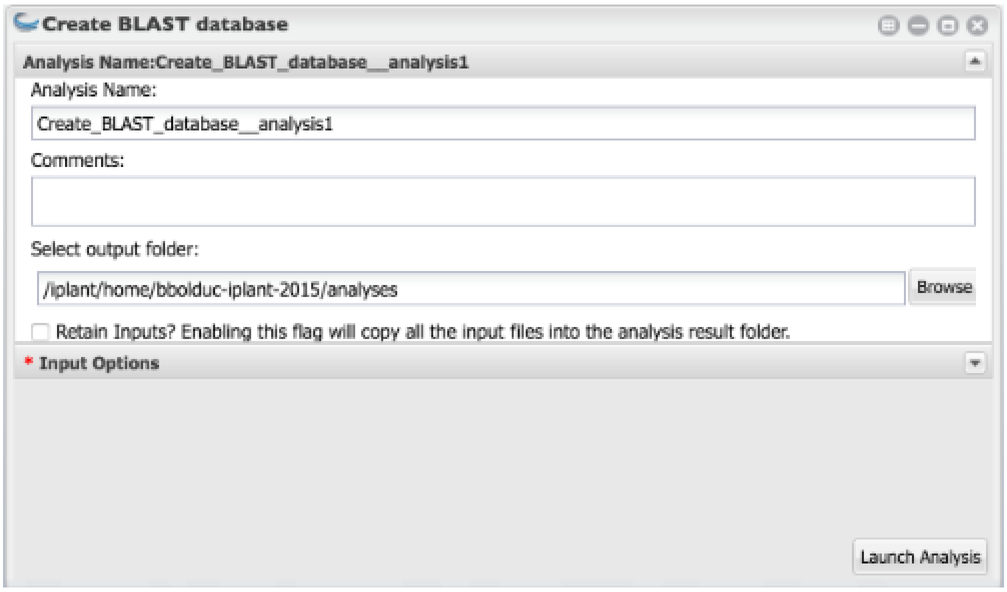
Protocol

Generating Protein Clusters via BLASTp

Step 1.

Open 'Create BLAST Database'

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



Create BLAST database

Analysis Name:Create_BLAST_database__analysis1

Analysis Name: Create_BLAST_database__analysis1

Comments:

Select output folder: /iplant/home/bbolduc-iplant-2015/analyses

□ Retain Inputs? Enabling this flag will copy all the input files into the analysis result folder.

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.

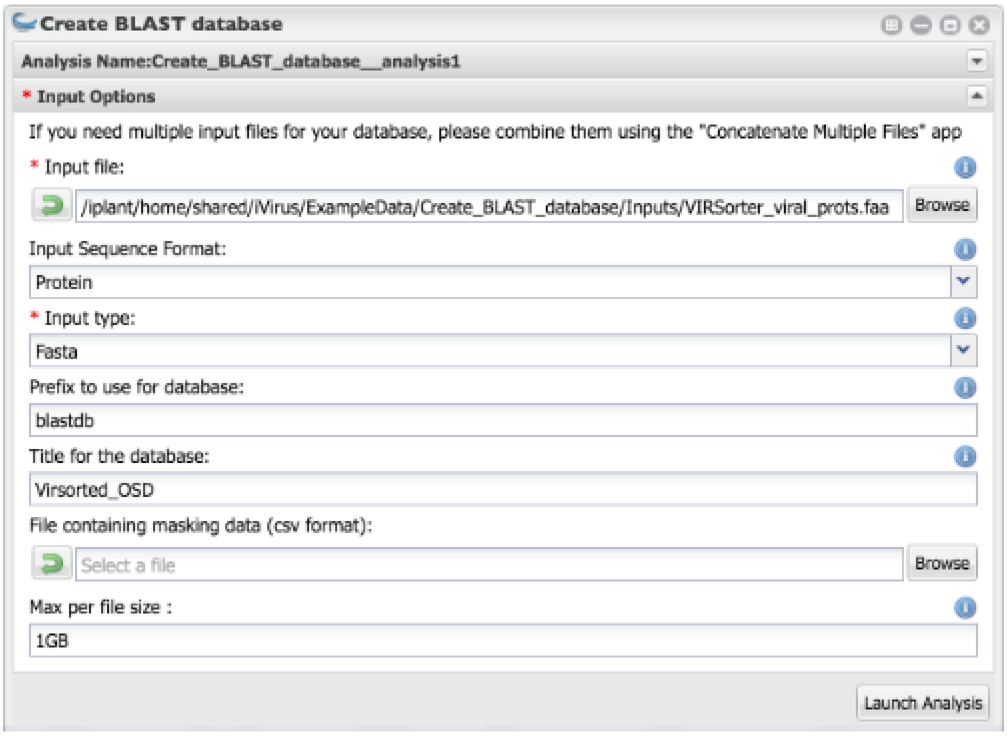
For **Input Sequence Format**:

- 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.



Create BLAST database

Analysis Name:Create_BLAST_database__analysis1

* Input Options

If you need multiple input files for your database, please combine them using the "Concatenate Multiple Files" app

* Input file: /iplant/home/shared/iVirus/ExampleData/Create_BLAST_database/Inputs/VIRSorter_viral_prots.faa

Input Sequence Format: Protein

* Input type: Fasta

Prefix to use for database: blastdb

Title for the database: Virsorted_OSD

File containing masking data (csv format): Select a file

Max per file size : 1GB

Generating Protein Clusters via BLASTp

Step 3.

Launch Analysis

Run the job!

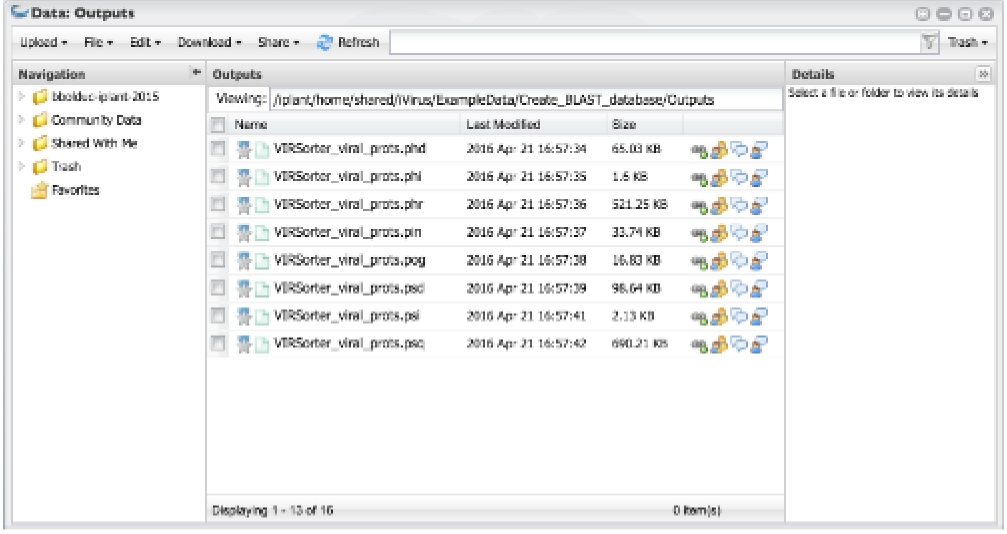
This should only take a few minutes.

Generating Protein Clusters via BLASTp

Step 4.

Results

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



Data: Outputs

Select a file or folder to view its details

Viewing: /iplant/home/shared/iVirus/ExampleData/Create_BLAST_database/Outputs

Name	Last Modified	Size
VIRSorter_viral_prots.phd	2016 Apr 21 16:57:34	65.03 KB
VIRSorter_viral_prots.phi	2016 Apr 21 16:57:35	1.6 KB
VIRSorter_viral_prots.phr	2016 Apr 21 16:57:36	521.25 KB
VIRSorter_viral_prots.pin	2016 Apr 21 16:57:37	33.74 KB
VIRSorter_viral_prots.pog	2016 Apr 21 16:57:38	16.83 KB
VIRSorter_viral_prots.psd	2016 Apr 21 16:57:39	98.64 KB
VIRSorter_viral_prots.psi	2016 Apr 21 16:57:41	2.13 KB
VIRSorter_viral_prots.psq	2016 Apr 21 16:57:42	690.21 KB

Displaying 1 - 13 of 16

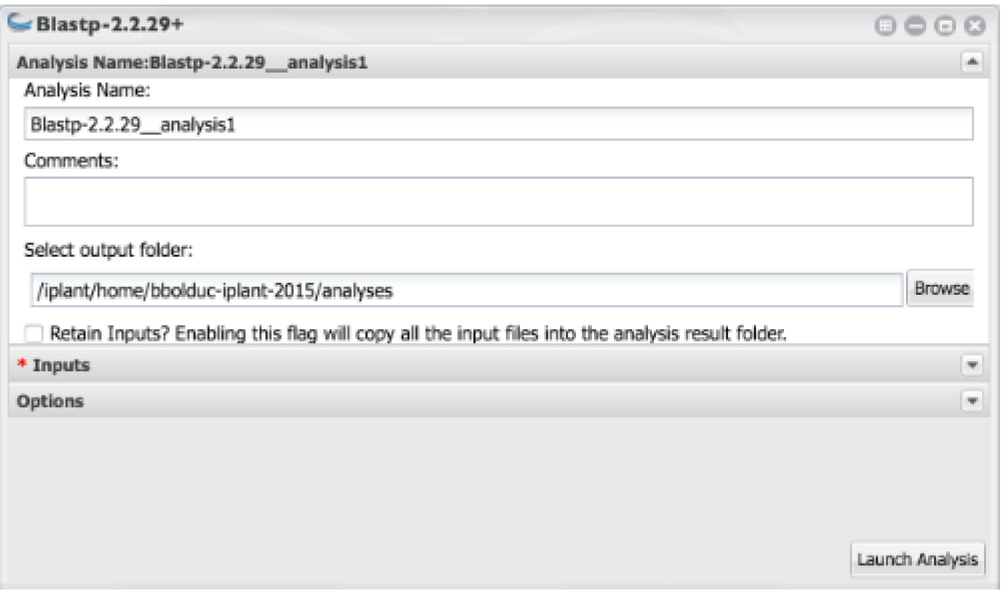
0 item(s)

Generating Protein Clusters via BLASTp

Step 5.

Open 'Blastp'

Open 'Blastp-2.2.29+' from the 'Apps' menu.



Analysis Name: Blastp-2.2.29__analysis1

Comments:

Select output folder: /iplant/home/bbolduc-iplant-2015/analyses

□ Retain Inputs? Enabling this flag will copy all the input files into the analysis result folder.

Generating Protein Clusters via BLASTp

Step 6.

Select Inputs

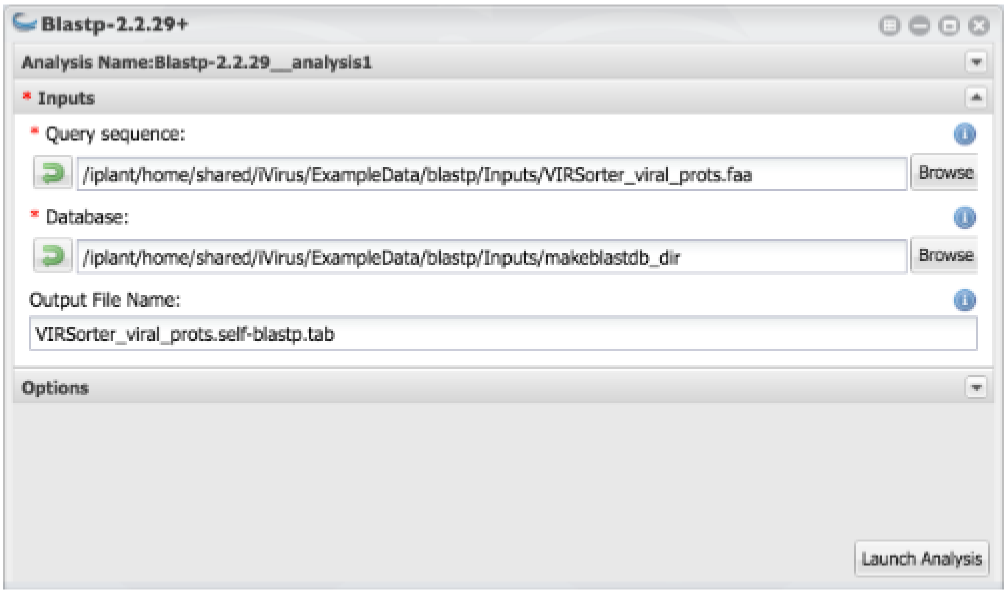
Select the 'Inputs' tab.

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.


Blastp-2.2.29+

Analysis Name:Blastp-2.2.29__analysis1

* Inputs

* Query sequence: /iplant/home/shared/iVirus/ExampleData/blastp/Inputs/VIRSorter_viral_prots.faa

* Database: /iplant/home/shared/iVirus/ExampleData/blastp/Inputs/makeblastdb_dir

Output File Name: VIRSorter_viral_prots.self-blastp.tab

Generating Protein Clusters via BLASTp

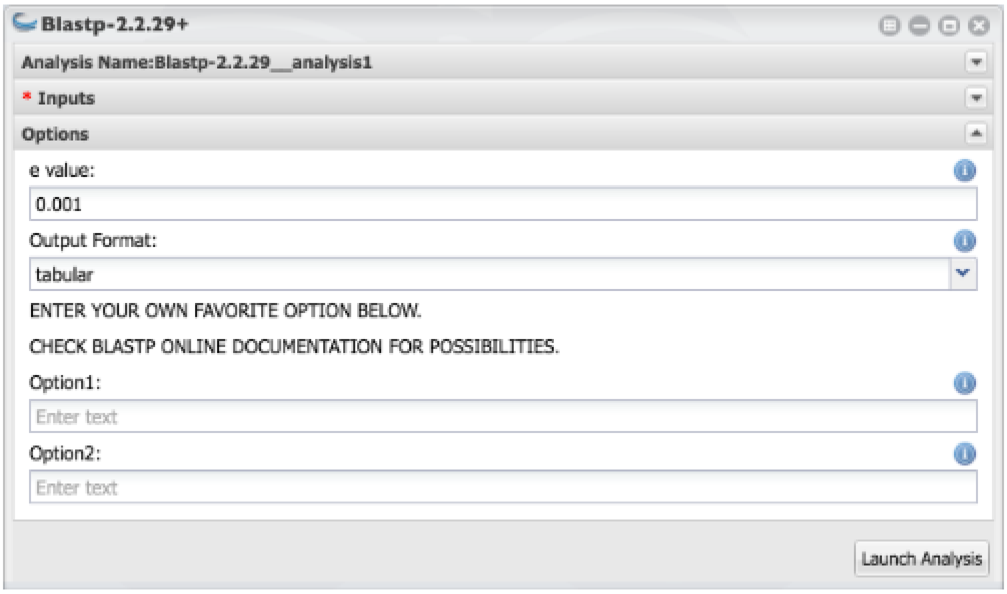
Step 7.

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.


Blastp-2.2.29+

Analysis Name:Blastp-2.2.29__analysis1

* Inputs

Options

e value: 0.001

Output Format: tabular

ENTER YOUR OWN FAVORITE OPTION BELOW.

CHECK BLASTP ONLINE DOCUMENTATION FOR POSSIBILITIES.

Option1: Enter text

Option2: Enter text

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.

Generating Protein Clusters via BLASTp

Step 9.

Results

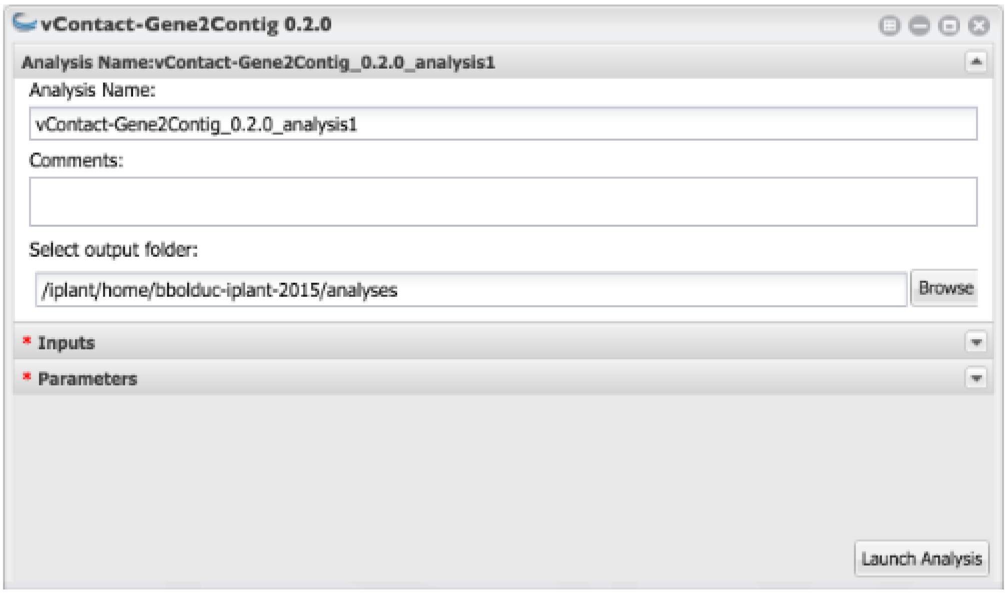
Expect results can be found in the blastp 'Output' directory.

Generating Gene-to-Contig Mapping

Step 10.

Open vContact-Gene2Contig

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



vContact-Gene2Contig 0.2.0

Analysis Name:vContact-Gene2Contig_0.2.0_analysis1

Analysis Name: vContact-Gene2Contig_0.2.0_analysis1

Comments:

Select output folder: /iplant/home/bbolduc-iplant-2015/analyses

Generating Gene-to-Contig Mapping

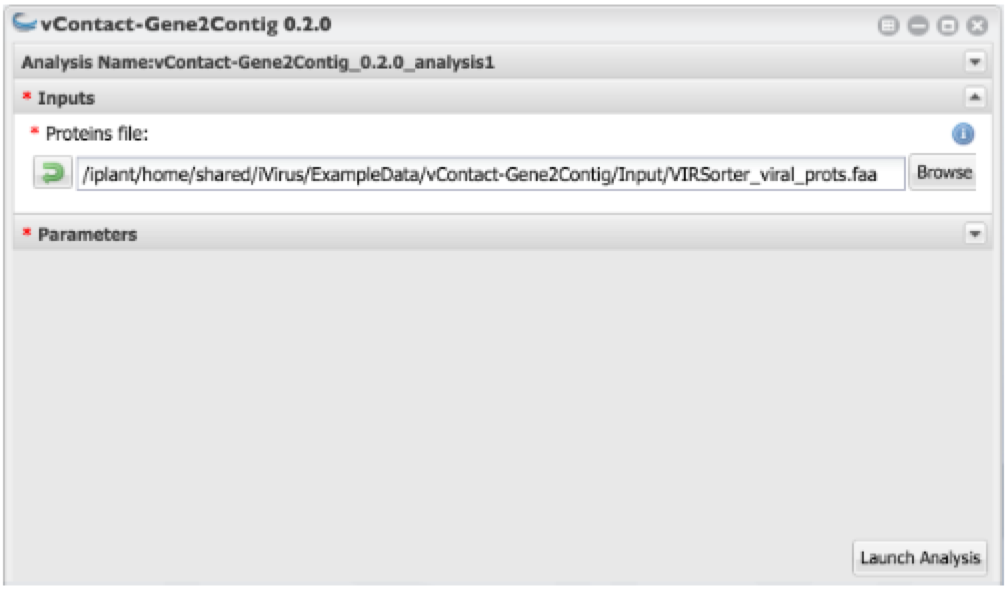
Step 11.

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.



vContact-Gene2Contig 0.2.0

Analysis Name:vContact-Gene2Contig_0.2.0_analysis1

* Inputs

* Proteins file:

/iplant/home/shared/iVirus/ExampleData/vContact-Gene2Contig/Input/VIRSorter_viral_prots.faa

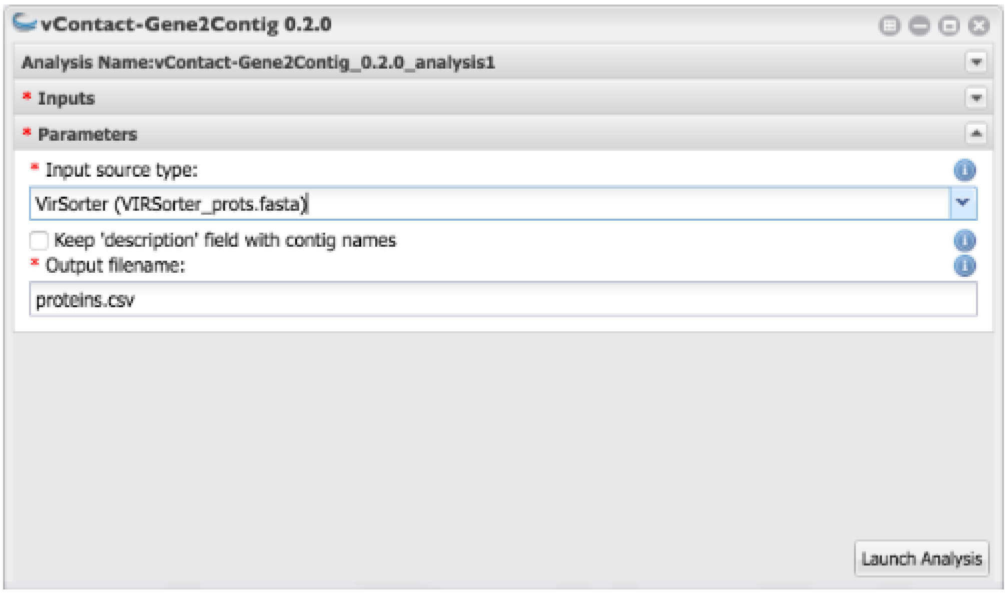
Generating Gene-to-Contig Mapping

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.



vContact-Gene2Contig 0.2.0

Analysis Name:vContact-Gene2Contig_0.2.0_analysis1

* Inputs

* Parameters

* Input source type: VirSorter (VIRSorter_prots.fasta)

□ Keep 'description' field with contig names

* Output filename: proteins.csv

Generating Gene-to-Contig Mapping

Step 13.

Launch Analysis

Run the job!

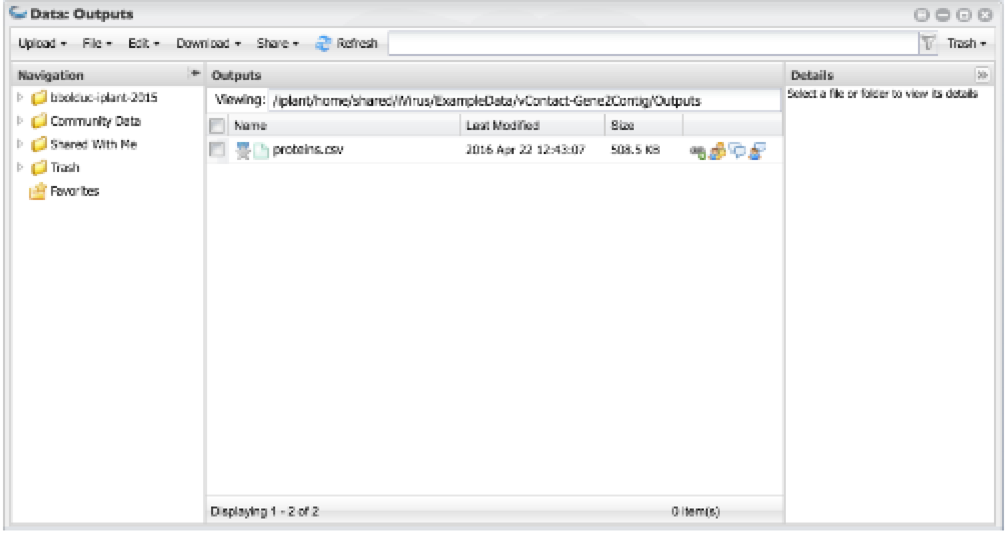
This should take minutes. Depending on the queue in Cyverse, it will likely take longer to submit and start the job than it does to run it!

Generating Gene-to-Contig Mapping

Step 14.

Results

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



Viewing: /iplant/home/shared/iVirus/ExampleData/vContact-Gene2Contig/Outputs

Name	Last Modified	Size
proteins.csv	2016 Apr 22 12:43:07	508.5 KB

Displaying 1 - 2 of 2

0 item(s)

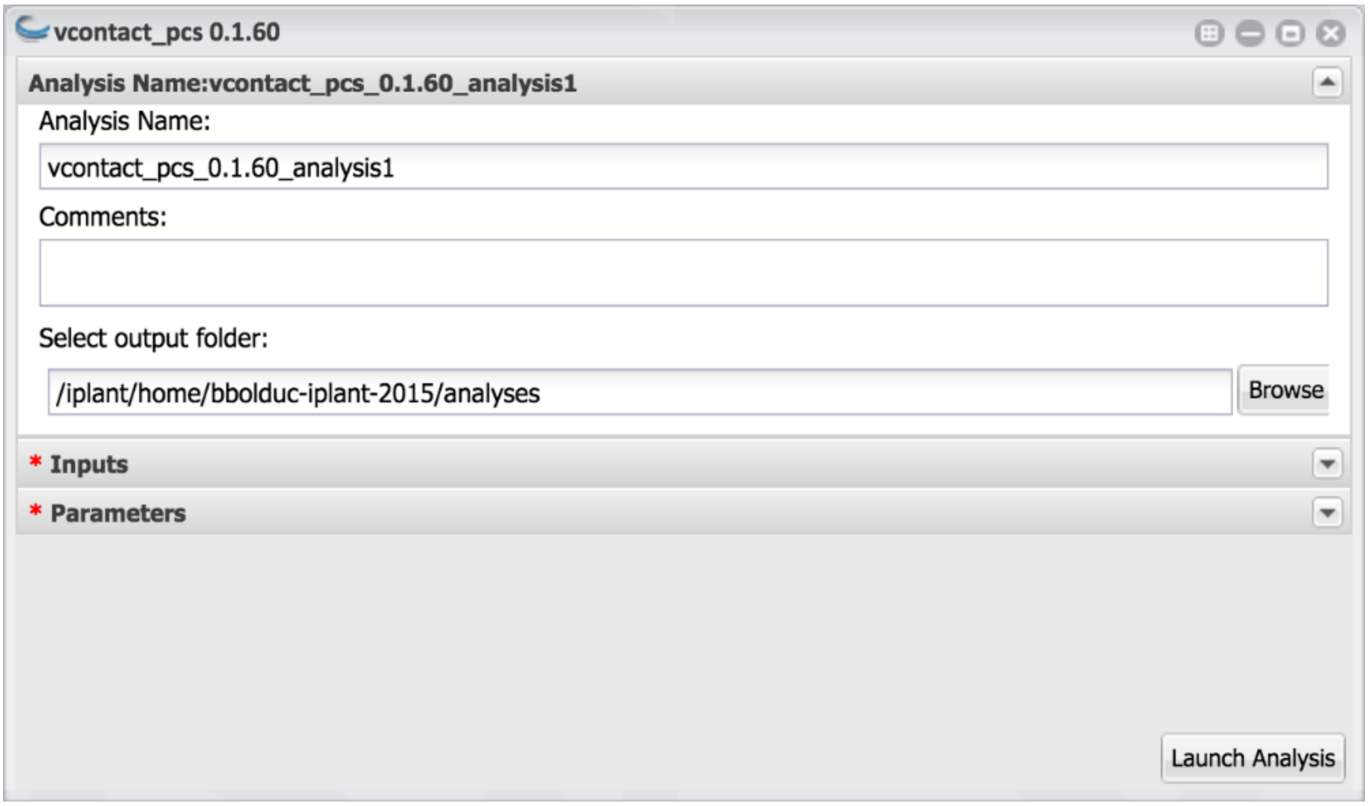
Select a file or folder to view its details

Preparing/formatting vContact inputs

Step 15.

Open vContact PCs

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



vcontact_pcs 0.1.60

Analysis Name:vcontact_pcs_0.1.60_analysis1

Analysis Name:

vcontact_pcs_0.1.60_analysis1

Comments:

Select output folder:

/iplant/home/bbolduc-iplant-2015/analyses

Preparing/formatting vContact inputs

Step 16.

Select Inputs

Select the 'Inputs' tab.

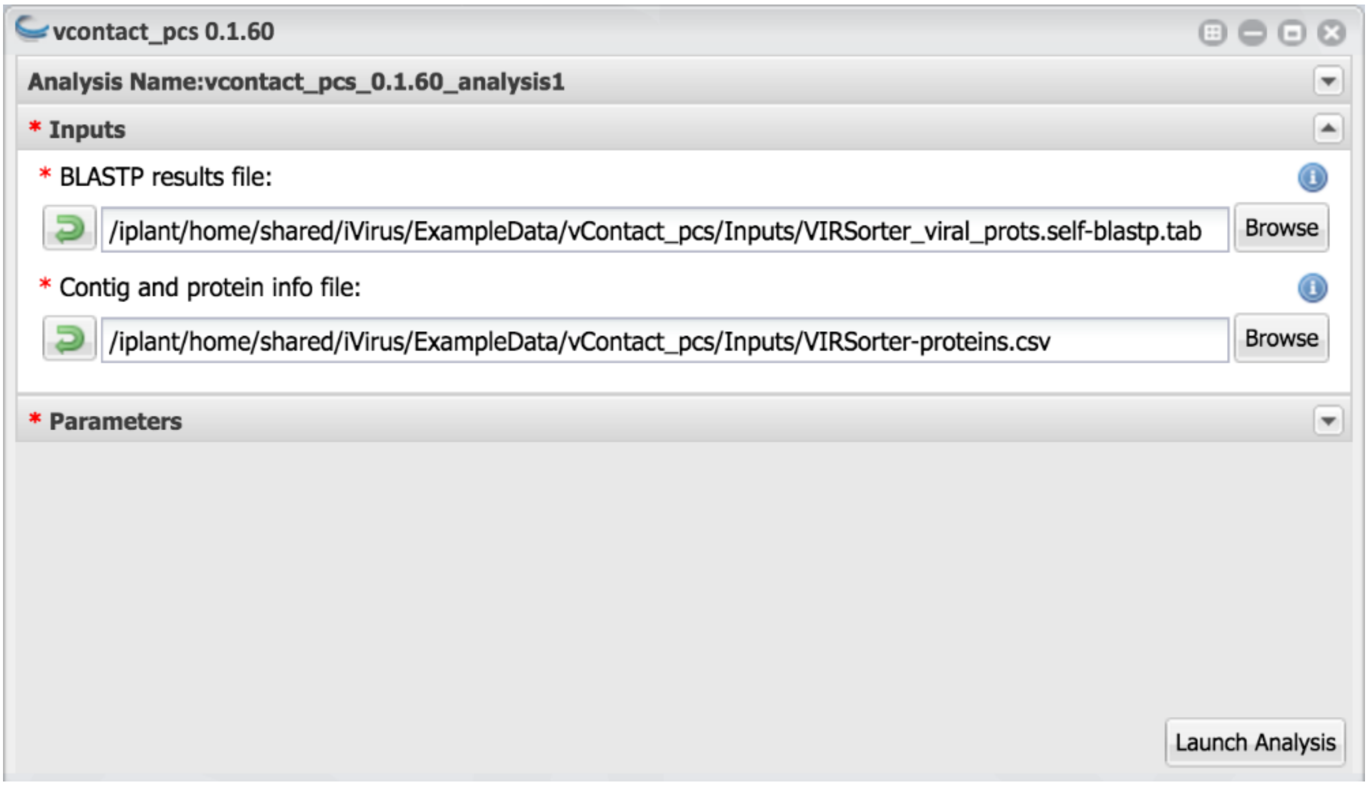
For **BLASTP results file**:

- 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.


vcontact_pcs 0.1.60

Analysis Name:vcontact_pcs_0.1.60_analysis1

* Inputs

* BLASTP results file:

/iplant/home/shared/iVirus/ExampleData/vContact_pcs/Inputs/VIRSorter_viral_prots.self-blastp.tab

* Contig and protein info file:

/iplant/home/shared/iVirus/ExampleData/vContact_pcs/Inputs/VIRSorter-proteins.csv

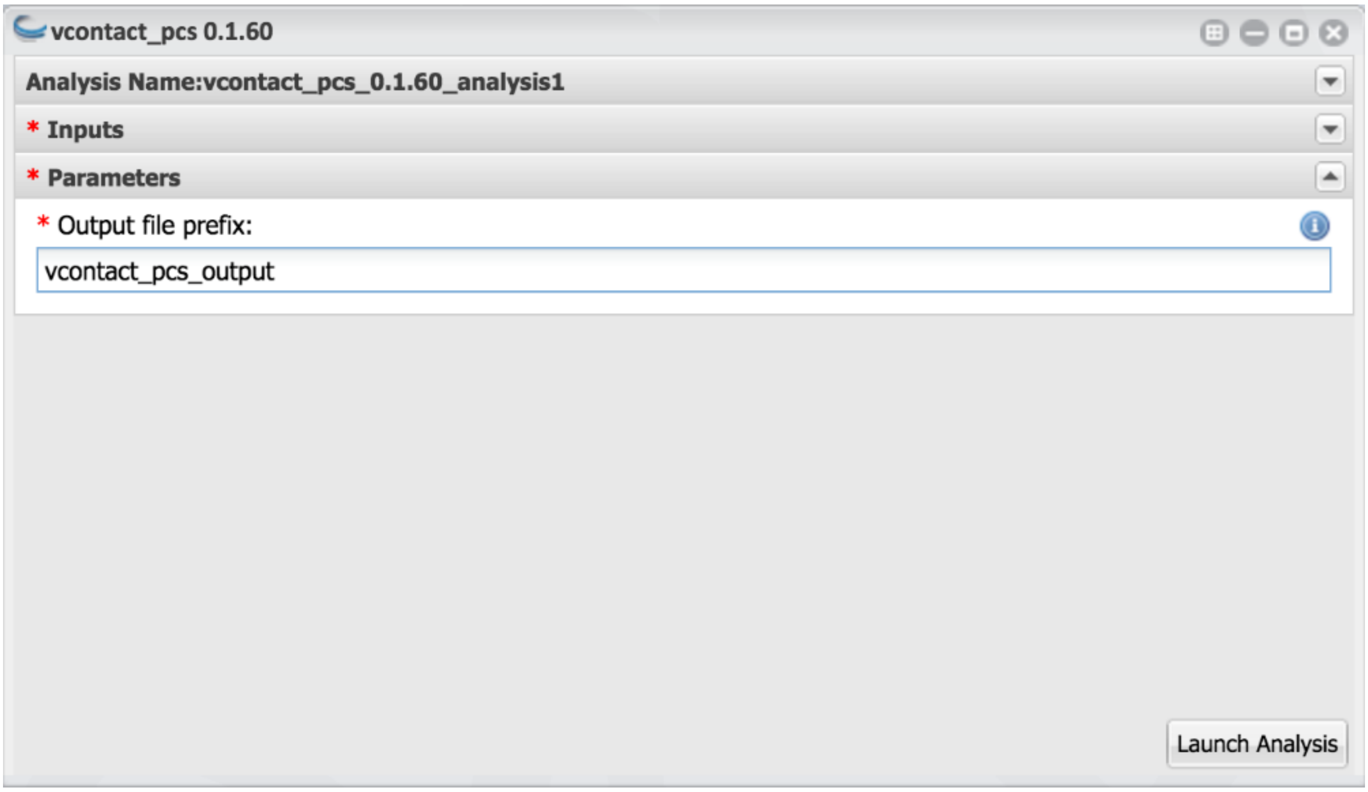
* Parameters

Preparing/formatting vContact inputs

Step 17.

Select Parameters

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


vcontact_pcs 0.1.60

Analysis Name:vcontact_pcs_0.1.60_analysis1

* Inputs

* Parameters

* Output file prefix:

vcontact_pcs_output

NOTES

Benjamin Bolduc 05 Jan 2017

TIP: When handling multiple vContact datasets, it's easier to name this something more convenient, or else they'll be 5x "pcs_output_contigs.csv" everywhere.

Preparing/formatting vContact inputs

Step 18.

Launch Analysis

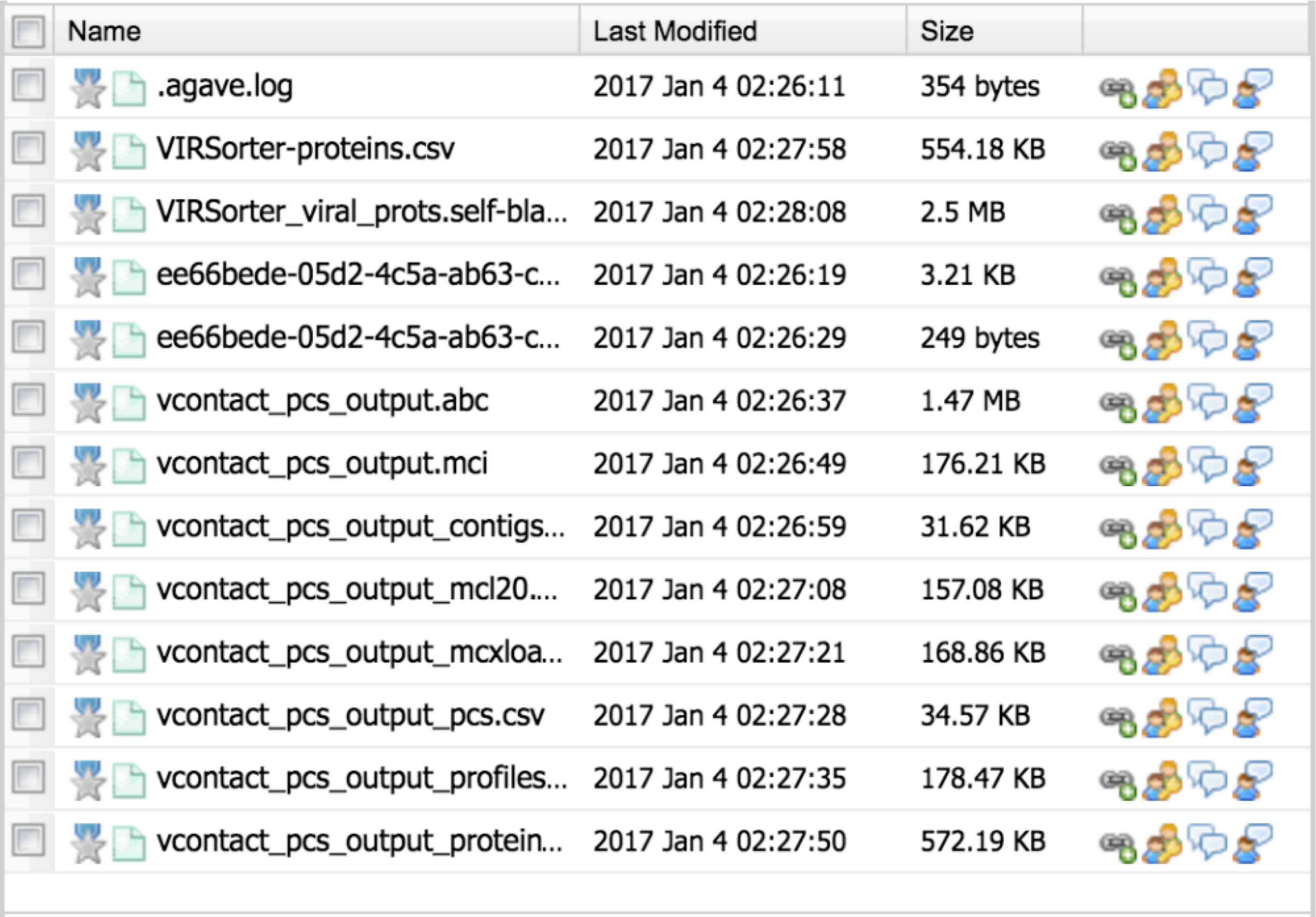
Run the job! This should complete within a minute or two, depending on the size of your dataset. This example should take less than a minute *once running*.

Preparing/formatting vContact inputs

Step 19.

Results

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



□	Name	Last Modified	Size	
□	.agave.log	2017 Jan 4 02:26:11	354 bytes	
□	VIRSorter-proteins.csv	2017 Jan 4 02:27:58	554.18 KB	
□	VIRSorter_viral_prots.self-bla...	2017 Jan 4 02:28:08	2.5 MB	
□	ee66bede-05d2-4c5a-ab63-c...	2017 Jan 4 02:26:19	3.21 KB	
□	ee66bede-05d2-4c5a-ab63-c...	2017 Jan 4 02:26:29	249 bytes	
□	vcontact_pcs_output.abc	2017 Jan 4 02:26:37	1.47 MB	
□	vcontact_pcs_output.mci	2017 Jan 4 02:26:49	176.21 KB	
□	vcontact_pcs_output_contigs...	2017 Jan 4 02:26:59	31.62 KB	
□	vcontact_pcs_output_mcl20....	2017 Jan 4 02:27:08	157.08 KB	
□	vcontact_pcs_output_mcxloa...	2017 Jan 4 02:27:21	168.86 KB	
□	vcontact_pcs_output_pcs.csv	2017 Jan 4 02:27:28	34.57 KB	
□	vcontact_pcs_output_profiles...	2017 Jan 4 02:27:35	178.47 KB	
□	vcontact_pcs_output_protein...	2017 Jan 4 02:27:50	572.19 KB	

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.

vcontact_pcs_output_contigs.csv
vcontact_pcs_output_profiles.csv
vcontact_pcs_output_pcs.csv