

Version 5

Apr 26, 2021

Preparing Data for vContact from Proteins (Cyverse) V.5

In 1 collection

Benjamin Bolduc¹¹The Ohio State University*In Development* dx.doi.org/10.17504/protocols.io.buimnuc6

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.

EXTERNAL LINK

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

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Bolduc B, Jang HB, Doulier 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)

DOI

dx.doi.org/10.17504/protocols.io.buimnuc6

EXTERNAL LINK

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

PROTOCOL CITATION

Benjamin Bolduc 2021. Preparing Data for vContact from Proteins (Cyverse). **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.buimnuc6>
Version created by Benjamin Bolduc

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Bolduc B, Jang HB, Doulier 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)

COLLECTIONS ⓘ

 **Processing a Viral Metagenome Using iVirus**

WHAT'S NEW

Updated to reflect changes in methodology and in response user feedback (mainly, to avoid confusion stemming from the same file name(s) referring to different data products).

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

CREATED

Apr 26, 2021

LAST MODIFIED

Apr 26, 2021

PROTOCOL INTEGER ID

49453

PARENT PROTOCOLS

Part of collection

[Processing a Viral Metagenome Using iVirus](#)

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 STARTING

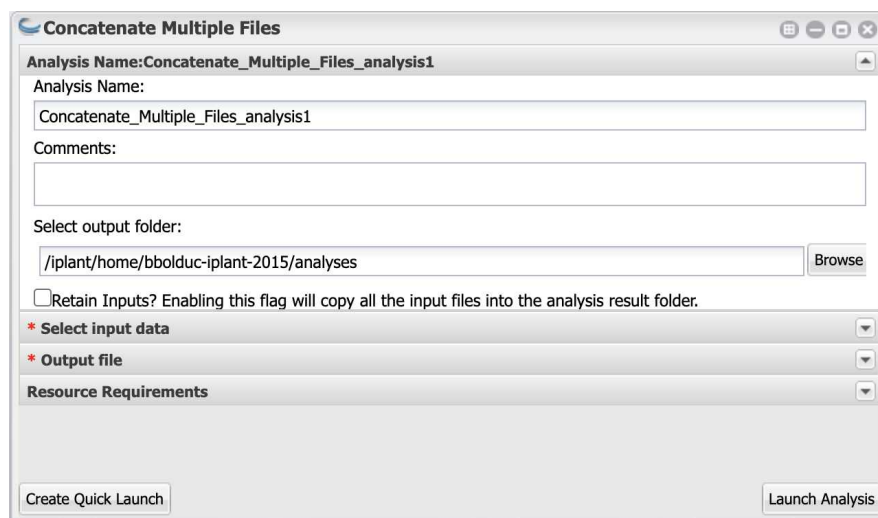
To run this protocol, users must first [register](#) for Cyverse account. All data (both inputs and outputs) are available within Cyverse's data store at `/iplant/home/shared/iVirus/ExampleData/`

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.

Preparing Files for Gene2Genome

- 1 After VirSorter finishes, you'll want to combine the FASTA-formatted nucleotide files into a single file. This makes the next few steps easier.

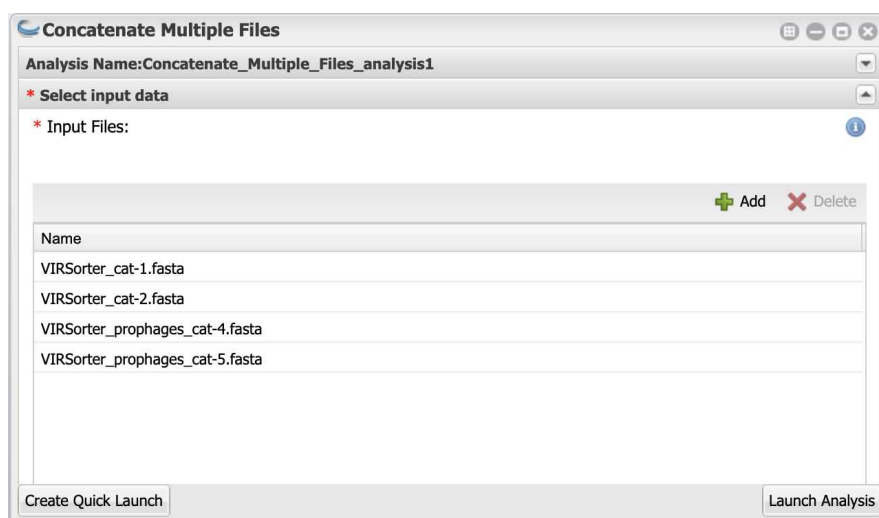
Open "Concatenate Multiple Files" from the "Apps" menu.



- 2 Select the input data:

Navigate to Community Data --> iVirus --> ExampleData --> iVirus_Ecogenomics_Pipeline. From the

VirSorter/Output/Fasta_files, add "VirSorter_cat-1.fasta" and "VirSorter_cat-2.fasta". However, in this example, "VirSorter_prophages_cat-4.fasta" and "VirSorter_prophages_cat-5.fasta" are *empty*, meaning that VirSorter did not predict any prophage. That's normal for this data.

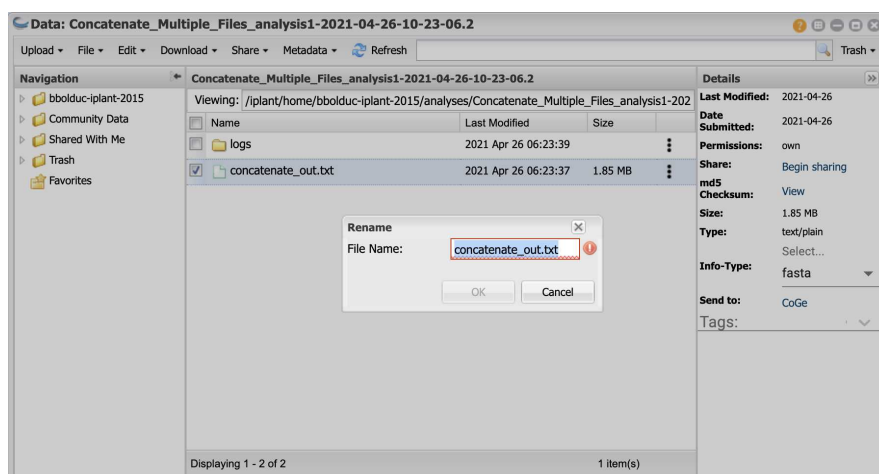


"High confidence" contigs usually derive from categories 1-2, 4-5. While it is possible some false positives are in categories 2 and 5, and some false negatives are in 3 and 6, it is generally believed that this is a fair compromise between balancing FPs with FNs. If you want to get 100% of all possible viruses potentially identified through VirSorter, you will need to *manually inspect* genomes in categories 3 and 6.

3 Launch analysis!

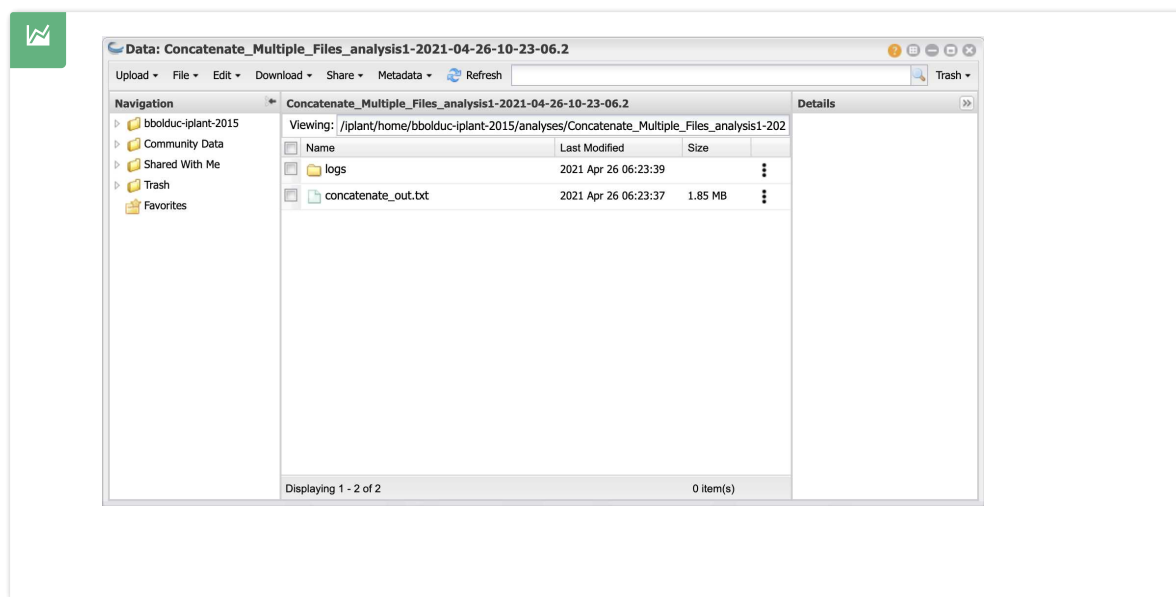
4 Since I don't like filenames that don't describe their content, we'll be renaming this to something more sensible.

Select the "concatenate_out.txt" file, and then from the "Edit" menu, select "Rename"



Rename it something that makes sense. For me, that'd be VirSorter_cat1245.fasta. That's because it represents VirSorter categories 1-2, 4-5 (even though 4-5 are empty). That should only take a second, and the result is

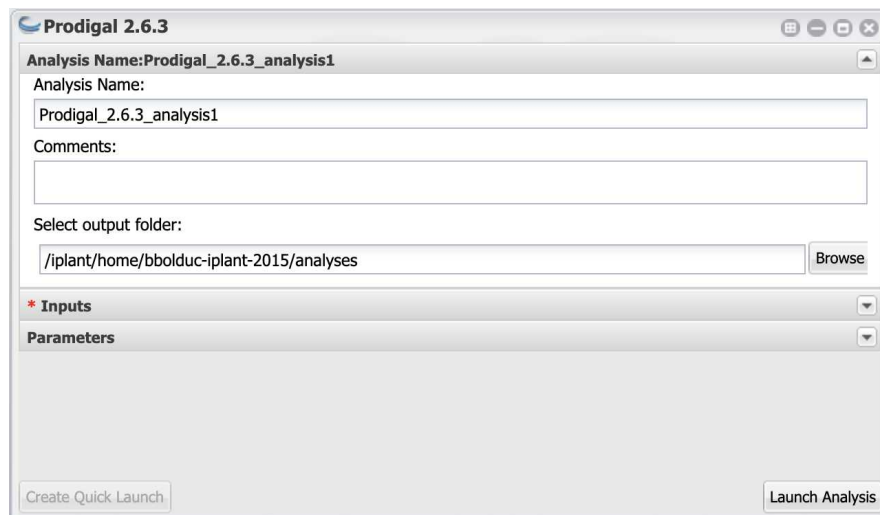
- 5 In the output directory, they'll be a single file, "concatenate_out.txt" and a "logs" directory.



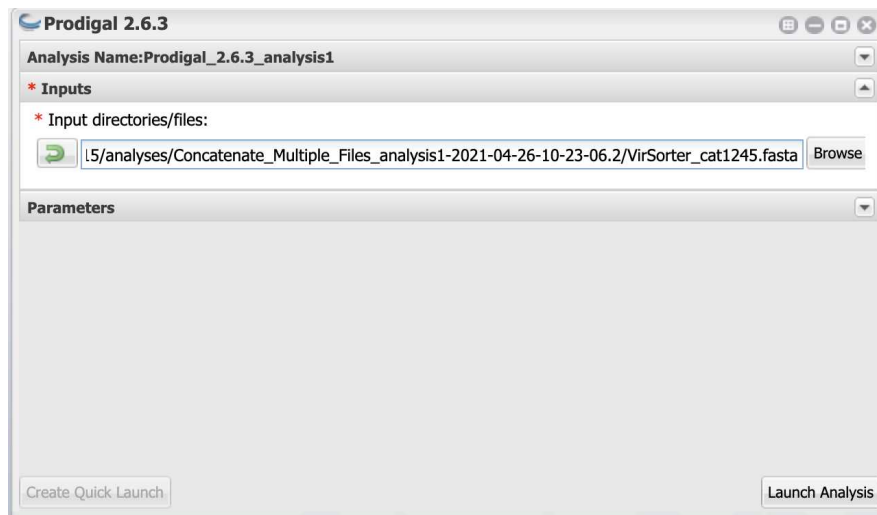
Run Prodigal

- 6 The next tool will be Prodigal. We're using Prodigal because the next app, vContact2-Gene2Genome, can take Prodigal output and convert it for use in vContact2.

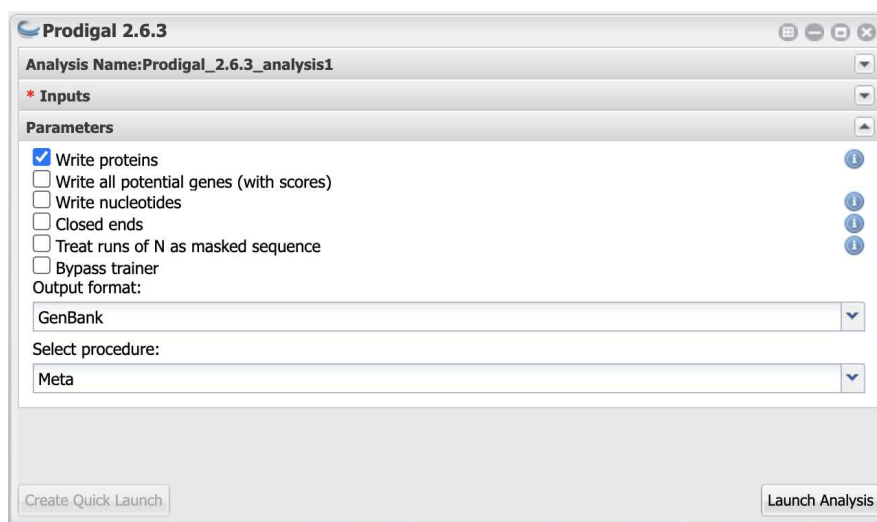
Open "Prodigal 2.6.3" from the "Apps" menu.



- 7 Under Inputs, select the concatenated file we just created.

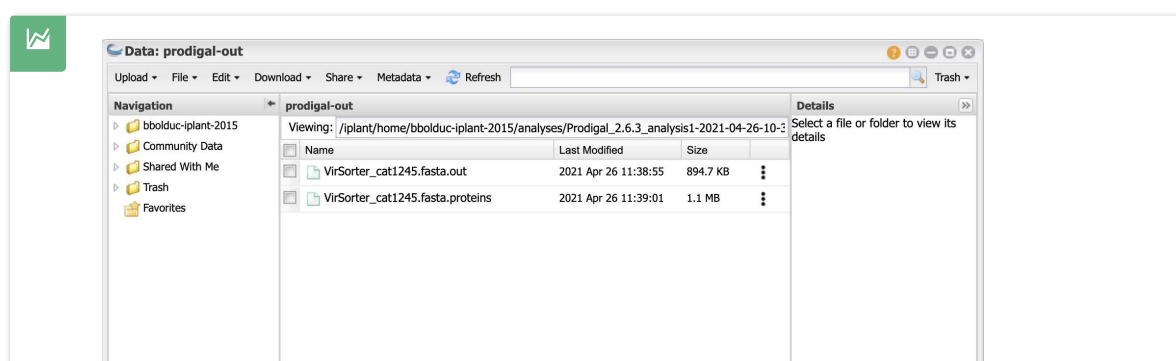


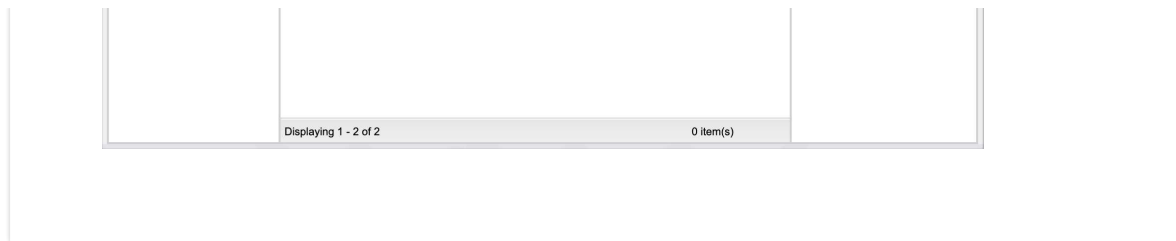
and under "Parameters", ensure that "Write Proteins" is selected and "Select procedure" is "Meta"



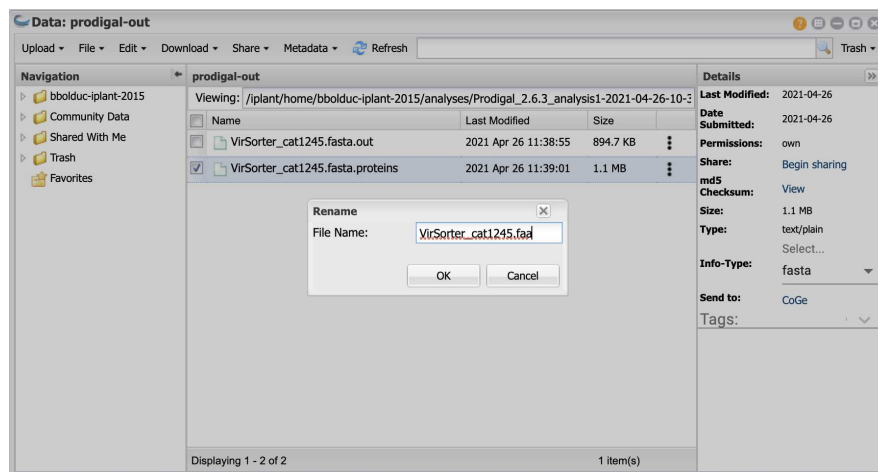
8 Launch analysis!

9 The output directory will contain the original file, and a prodigal output directory.





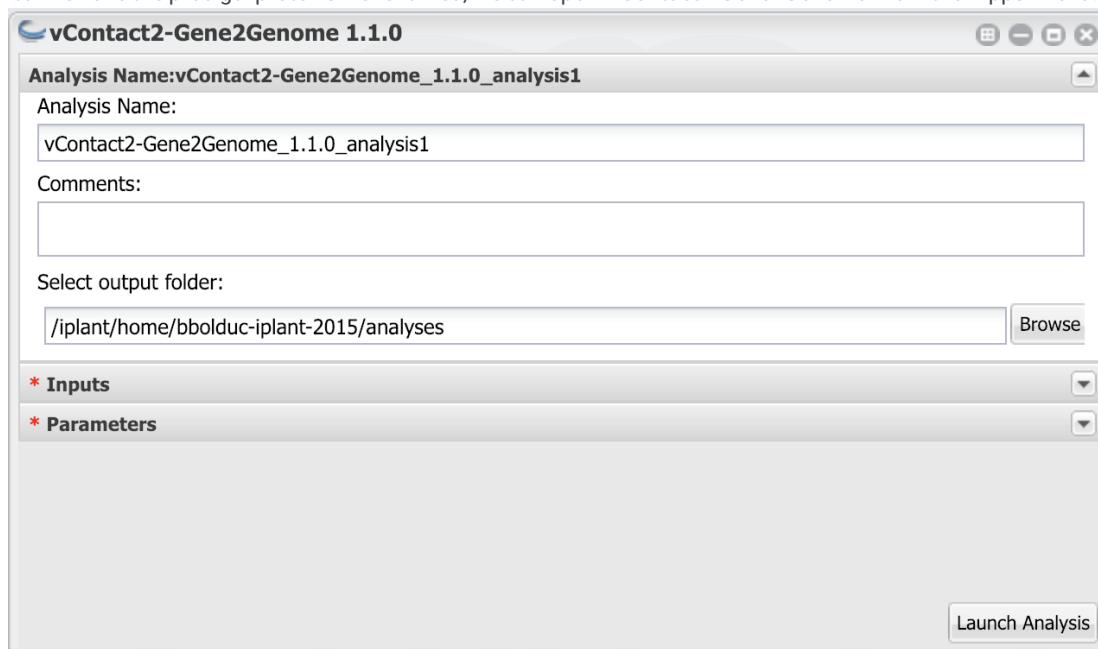
10 Rename that file to "VirSorter_cat1245.faa"



Generating Gene-to-Genome Mapping

11 Open vContact2-Gene2Genome

After we have the prodigal proteins file renamed, we can open "vContact2-Gene2Genome" from the "Apps" menu.



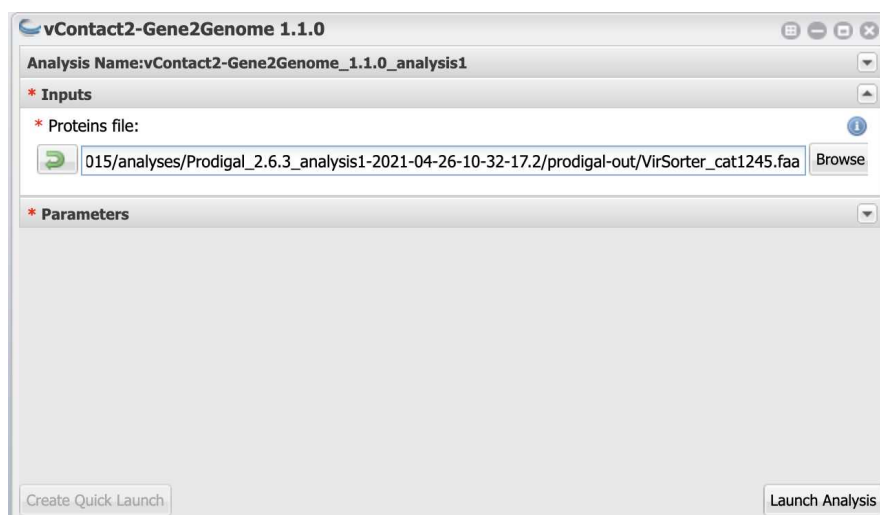
Starting menu for the Gene2Genome app in the CyVerse Discovery Environment.

12 Select Inputs

Under the 'Inputs' tab, select that renamed file

For **Proteins file**:

- Navigate to *Community Data* --> *iVirus* --> *ExampleData* --> *iVirus_Ecogenomics_Pipeline* --> *vContact2-Gene2Genome* --> *Input*. Select *VirSorter_cat1245.faa* Alternatively, copy-and-paste the location: `/iplant/home/shared/iVirus/ExampleData/iVirus_Ecogenomics_Pipeline/Prodigal_2.6.3/prodigal-out` into the navigation bar and select the protein fasta file.



13 Select Parameters

Under "**Input source type**" change to *Prodigal amino acid proteins*. 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 Prodigal, so we'll use Prodigal'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.



14 Launch Analysis

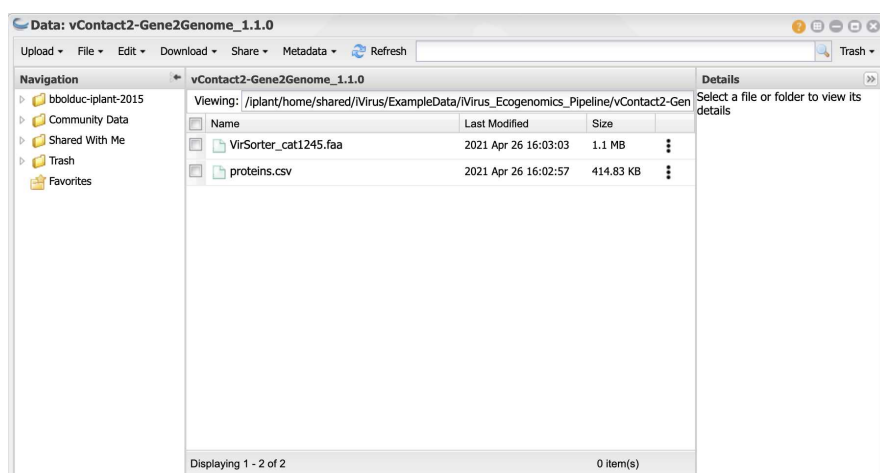
Run the job!

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!

15 Results



- Expect results can be found in the vContact2-Gene2Genome 'Outputs' directory. They'll be 4 files: 2 with agave messages (the errors and outputs, deleted in the screenshot), the original proteins file (VirSorter_cat1245.faa) and the gene-to-genome mapping file, "proteins.csv"



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

