Applying vContact to Viral Sequences and Visualizing the Output (Cyverse)

Version 4

Benjamin Bolduc[1]

[1]The Ohio State University

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

Sullivan Lab iVirus

Benjamin Bolduc

ABSTRACT

A collection of protocols designed to guide the user in processing a viral metagenome from raw sequence data to assembly, and subsequent analysis. The user uses *actua*/reads from Ocean Sampling Day (2014) and processes them entirely within Cyverse, a NSF-supported cyberinfrastructure.

EXTERNAL LINK

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

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

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

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

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

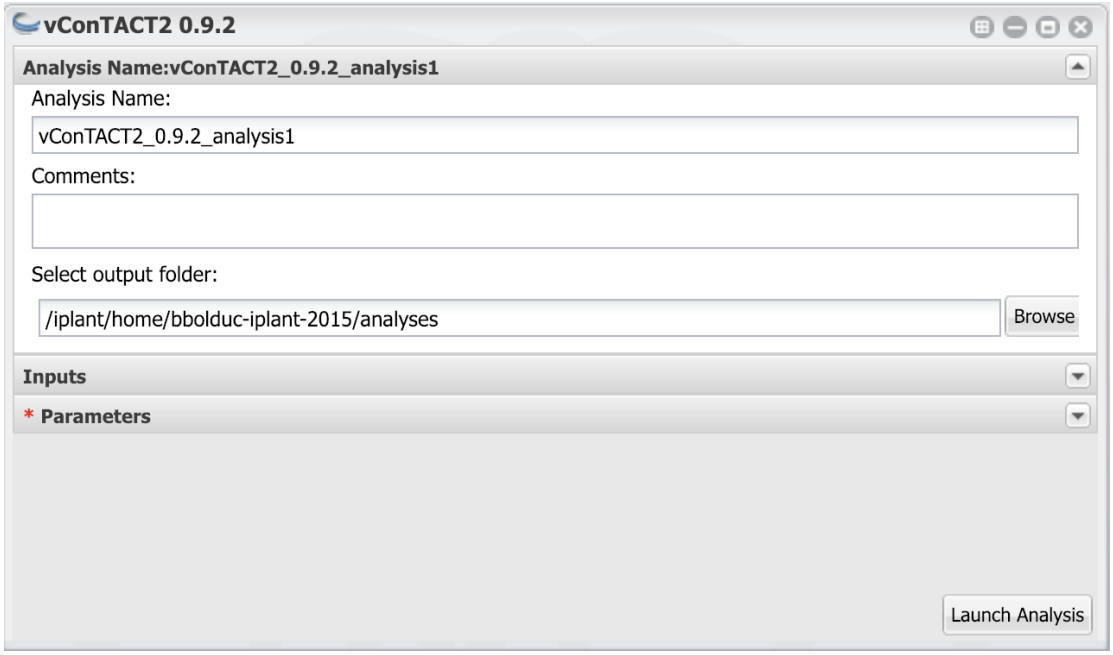
To run this protocol, users must first register for Cyverse account. All data (both inputs and outputs) are available within Cyverse's data store at /iplant/home/shared/iVirus/ExampleData/

1. Download and install Java JDK 8
2. Download and install Cytoscape 3.x

Affiliating contigs through their shared proteins

1 Open vConTACT2

Open vContact2-0.9.2 from 'Apps'



Starting menu for the vConTACT2 app in the CyVerse Discovery Environment

2 Select Inputs

Select the 'Inputs tab.

There are 3 main ways to provide input files to vConTACT2:

1) Provide a **FASTA-formatted amino acid proteins file** and **gene-to-genome mapping file.** This is the easiest, simpliest and most straight-forward

For the **FASTA-formatted amino acid protein file**

- This file is straightforward, a standard fasta-formatted file (each protein id/name starting with ">", with the following line IUPAC amino acid codes). The sequences should be derived from a virus-identification tool, such as VirSorter or VirFinder.
- Navigate to *Community Data --> iVirus --> ExampleData* --> vContact2 *--> Inputs.* Select *VIRSorter_viral_prots.faa* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/vContact2/Inputs into the navigation bar and select the faa file.

For the **gene-to-genome mapping file**

- This file is generated from the vConTACT2-Gene2Genome app. This file *must contain the headers* "protein_id", "contig_id" and "keywords." protein_id is the gene name, which must match the name from the amino acid file. contig_id is the name of the genome associated with that gene/protein. keywords can be a single element describing the gene. Examples include "dna_pol" or "helicase." Multiple keywords need to be separated by a semi colon (";"), for example "dna_pol; helicase; podoviridae; experimental."

2) **Provide the old "vConTACT1" input files**. This option is mainly provided for existing users of vConTACT1 who want to compare results from the old method to the new.

For **Protein clusters info file**:

- This file contains the "id", "size", "annotated" and "keys" for each PC in the dataset, with id (PC ID), size (number of genes within the PC), annotated (number of genes including annotation) and keys (;-separated list of key terms extracted from gene annotations).
- Navigate to *Community Data --> iVirus --> ExampleData* --> vContact *--> Inputs --> vcontact_pcs_0.1.60*. Select *vcontact_pcs_output_pcs.csv* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/vContact/vcontact_pcs_0.1.60 into the navigation bar and select the csv file.

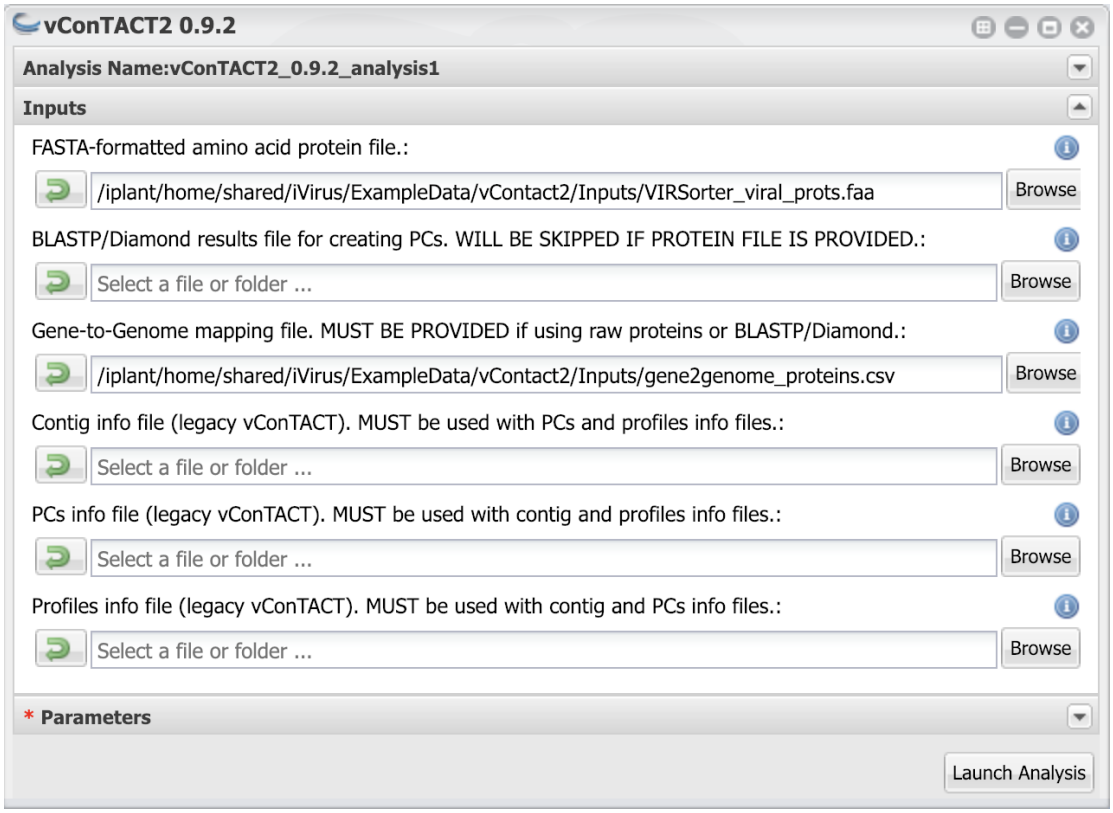
For the **Contig info file**:

- This file contains the 'id' and 'proteins' in the dataset, with id corresponding to the contig and proteins the number of proteins identified for each contig.
- Navigate to *Community Data --> iVirus --> ExampleData* --> vContact *--> Inputs --> vcontact_pcs_0.1.60*. Select *vcontact_pcs_output_contigs.csv* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/vContact/vcontact_pcs_0.1.60 into the navigation bar and select the csv file.

For **Protein cluster profiles**:

- This file contains the 'contig_id' and 'pc_id' between contigs and PCs in the dataset. Essentially a list of the membership of each gene within a contig to its affiliated PC.
- Navigate to *Community Data* --> *iVirus* --> *ExampleData* --> vContact --> *Inputs* --> *vcontact_pcs_0.1.60*. Select *vcontact_pcs_output_profiles.csv* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/vContact/vcontact_pcs_0.1.60 into the navigation bar and select the csv file.

3) **Provide a BLASTP/Diamond file and a gene-to-genome mapping file**. This is for users who want/need to run the protein search locally. This is often for those who have many sequences and require more than 2 days of processing time. In such a case, a user would run blastp (or diamond) on their local compute (local machine, server, HPC) for days to weeks to months, and then upload the results to CyVerse. *This* is that file.



Input section of the vConTACT2 app in the CyVerse Discovery

NOTE

The number of available input options here can be overwhelming. You *nearly always* need the Genome-to-Genome mapping file AND either a FASTA-formatted amino acid file *or* a Blastp/Diamond results file. The contig info, pc info, profiles info files are not really necessary and generally only serve to confuse people.

3 Select Parameters

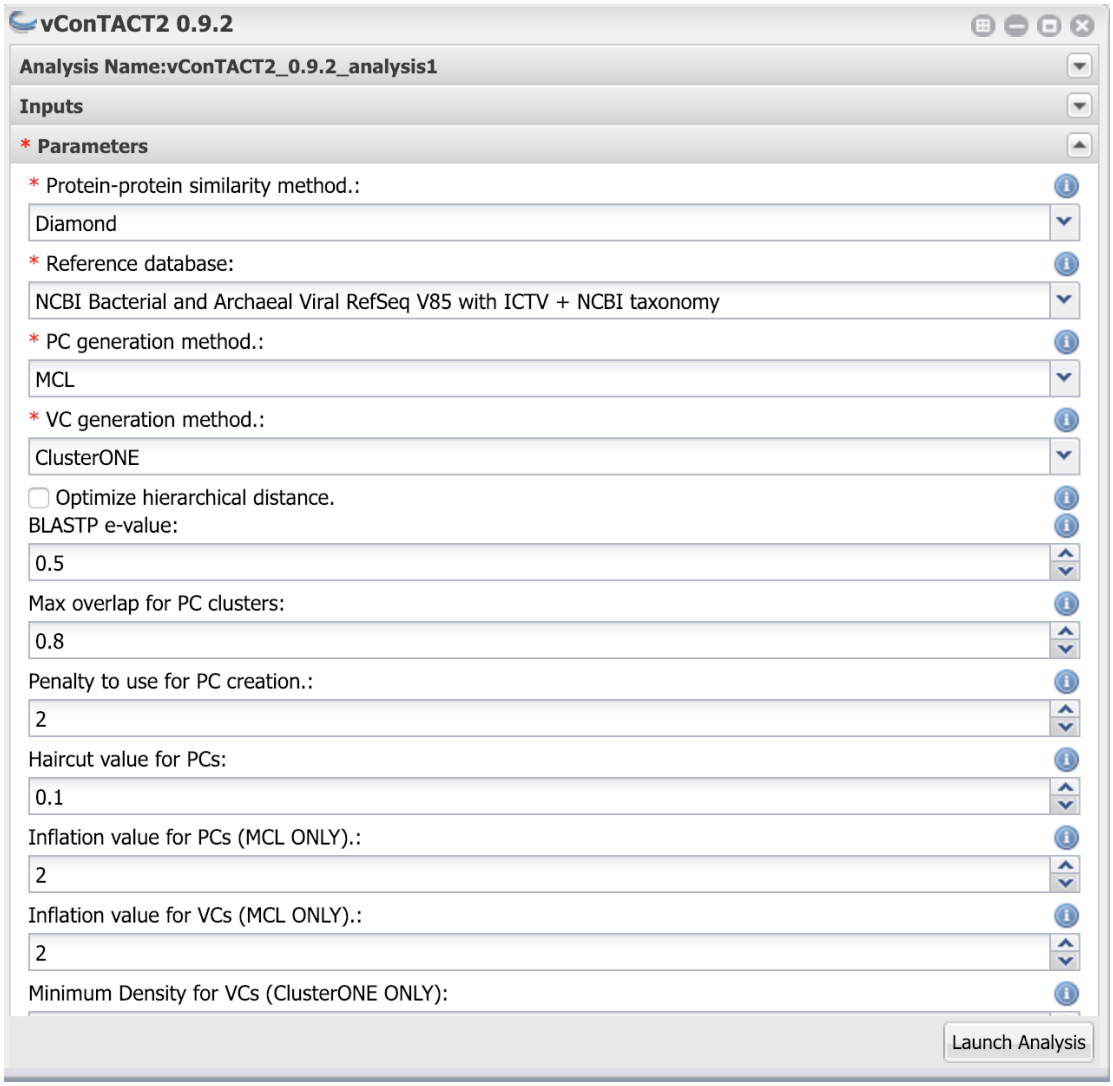
Select the 'Parameters' tab.

The default options will suffice for this example. Consult the relevant documentation for what each of these options mean. *Briefly though*, the only options that most will change are the **Reference database** and **Protein-protein similarity method**.

Reference database: A selection between bacterial and archaeal viral refseq ("prokaryotic") or just archaeal ("archaeal"), using either the ICTV taxonomy (more "accurate", but accounts for a small portion of the RefSeq genomes) or the ICTV + NCBI taxomy (supplements the ICTV taxonomy with NCBI).

Protein-protein similarity method: A selection between BLASTP and Diamond. BLASTP is what vConTACT1 originally used, and results in *arguably* a more accurate PC clustering result. *However*, the final viral clusters (those that we have confidence in) are often indistinguishable between the two. Diamond is much faster, and a little more stringent (by default). Faster often means ~5 mins vs 2 hours.

There are *many, many* parameters available for vConTACT2. This is because each stage in the processing has a tool or function with its own sets of arguments. Our own lab's work (and accompanying manuscript) has revealed optimal values for each argument, but each person's dataset is different. While it is possible that a specific dataset will require drastically different defaults, *they should work for the vast majority (if not all) users*.



NOTE

Technically, vConTACT1 used MCL for the **VC Generation Method**. Our lab's research (and publication) strongly support ClusterONE as being superior to MCL in terms of separation, sensitivity, and accuracy. The one downside of ClusterONE is that highly overlapping genomes (those present in 2 or more viral clusters) are excluded from the analysis.

4 Launch Analysis

Run the job!

vContact2 can take minutes to hours to the better part of a day to complete.

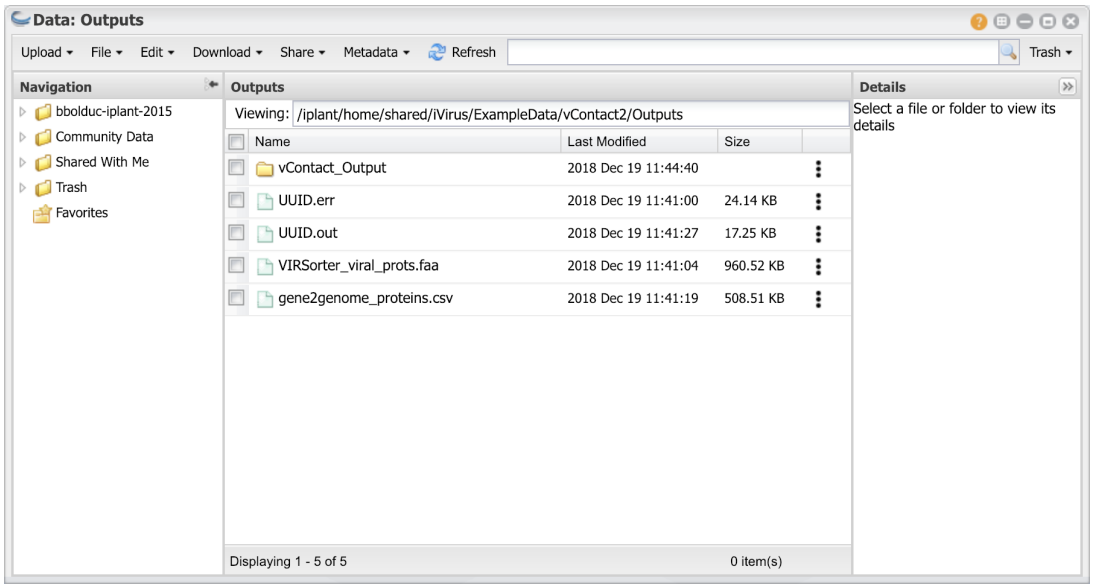
NOTE

Jobs run using Diamond should take anywhere from 5 minutes to a few hours. Using BLASTP can take a couple hours to 2 days. If jobs on CyVerse are being cancelled after 2 days due to time limits, you may need to run vConTACT2 on local compute.

5 Results

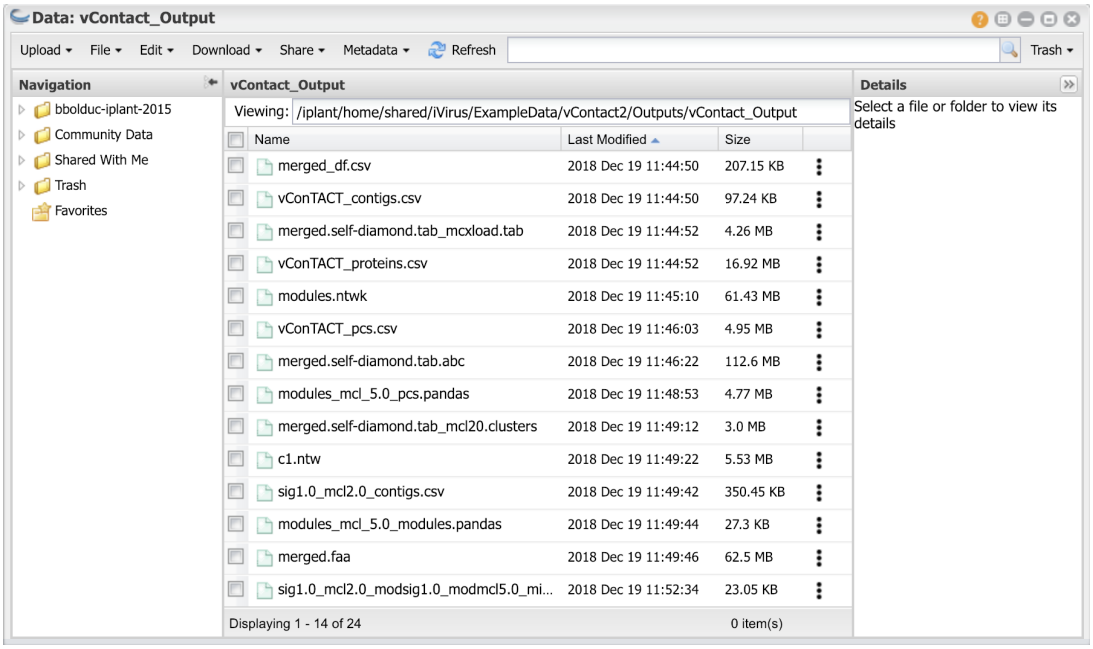
vConTACT2 will generate *a lot* of files. The output directory will consist of any input files that were used (in this example, we used a proteins amino acid file and the gene2genomes file), the CyVerse output and error logs, and the actual results from the vConTACT2 run. The network files can be imported into Cytoscape (more Initial app output directory structurebelow) to visualize the modules and the contig clusters.

EXPECTED RESULT

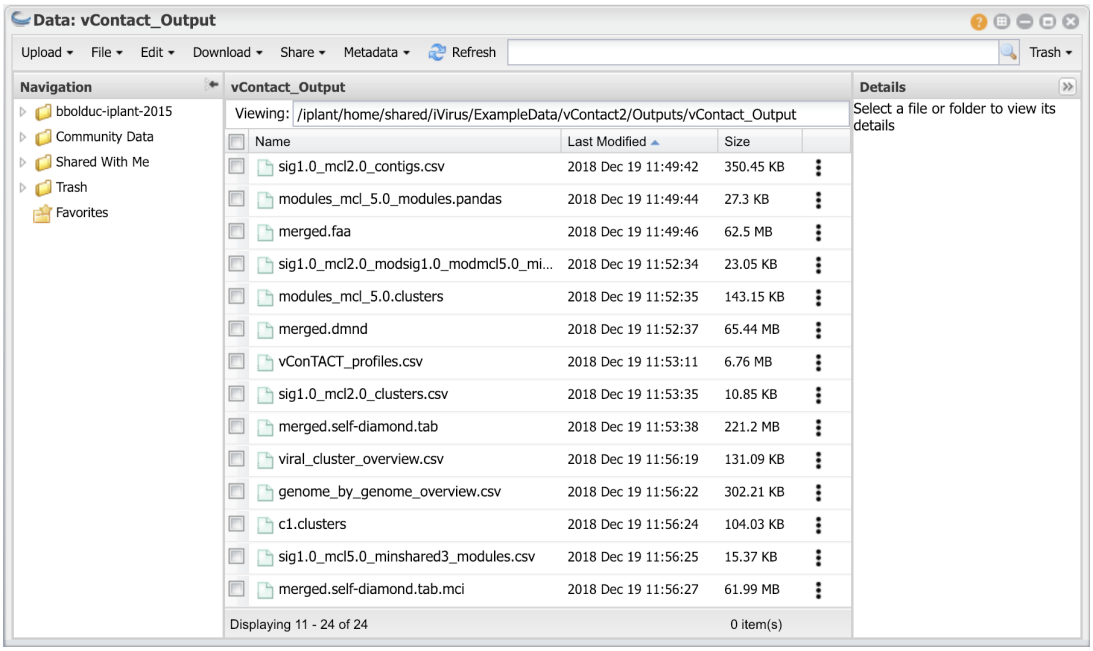


Initial app output directory structure for vConTACT2.

In the screenshot below, the notable files are **c1.ntw**. This is the network file that needs to be sent to Cytoscape. Other mentions are **vConTACT_contigs.csv**, ***_proteins.csv** and ***_pcs.csv**. For users of vConTACT1, you'll remember that those 3 files were the original inputs. Since vConTACT2 now handles the generation of those files internally, there's no need for users to do it themselves. However, this files are important for restarting failed runs and/or troubleshooting any issues during vConTACT2 runs.



In the figure below, the two notable files are **viral_cluster_overview.csv** and **genome_by_genome_overview.csv**. They contain information regarding the membership, confidence levels, taxonomy, and clustering of the virus clusters and individual genomes, respectively.



EXPECTED RESULT

Cluster Visualization

6 Open Cytoscape

Open Cytoscape *on your local machine.*

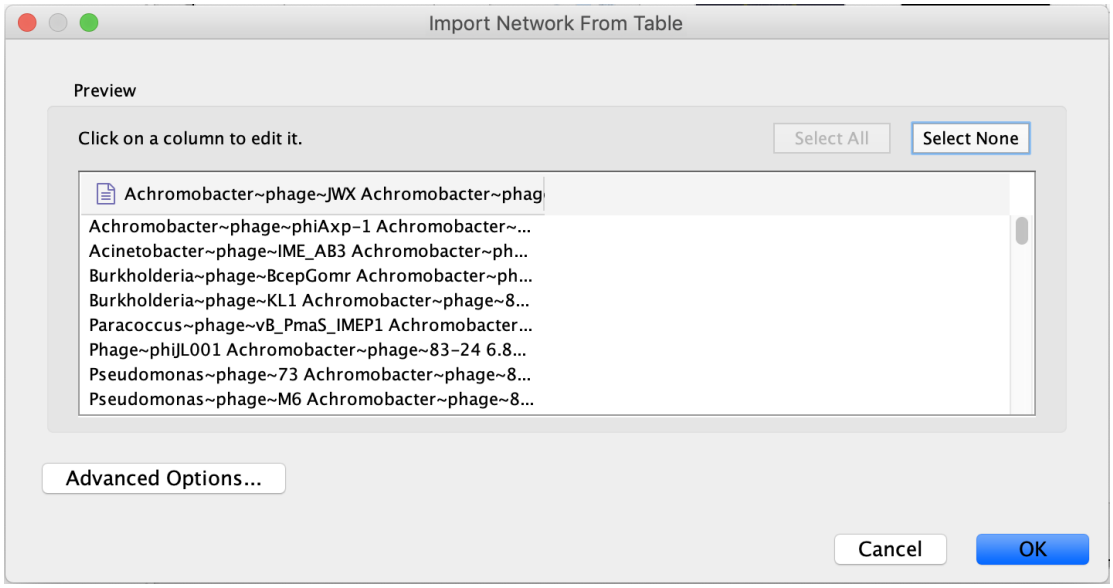
7 Locate and Select Network File

- If a 'splash window' appears, select 'Start New Session - From Network File...'
- If the window doesn't appear, go to File -> Import -> Network -> File...

Select the contig *.ntw (typically, **c1.ntw** as in the example data above, but can also be cc_sig1.0_mcl2.0.ntw).

8 Import Network File

When you import the datafile, you'll be presented with a data table:

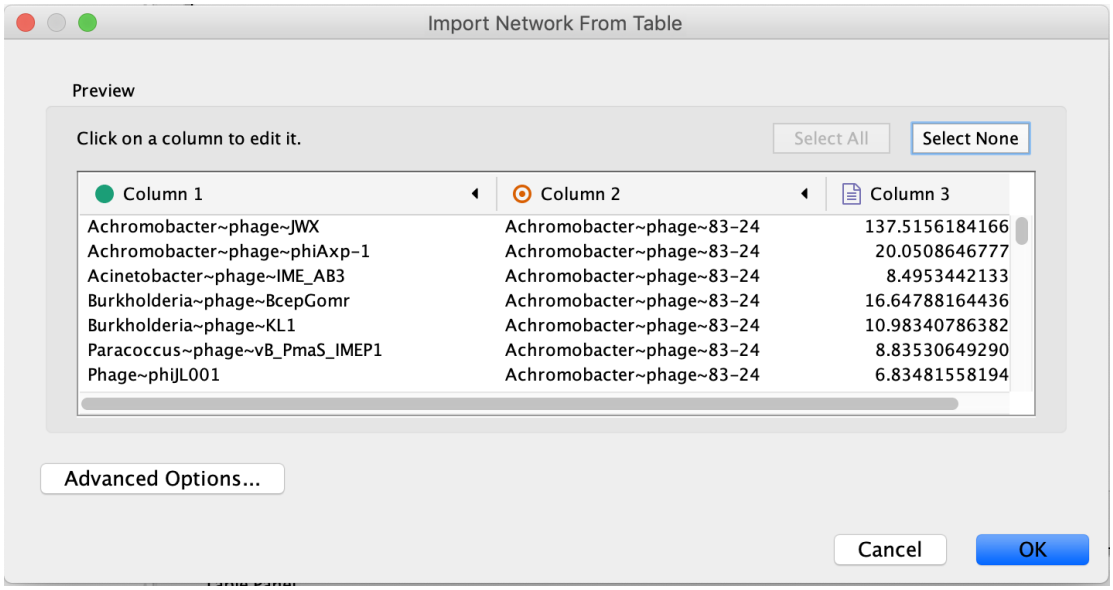


1. 'Select 'Advanced Options' and select the appropriate Delimiter, in this case 'SPACE.' and click 'OK.'

- At this point you can change the 'Default Interaction' to something more meaningful, or keep as is.
- This changes the single column import into 3 (there might be one hiding on the right)

1. Click on 'Column 1' and under *Meaning*, select *Source Node* (little green button).

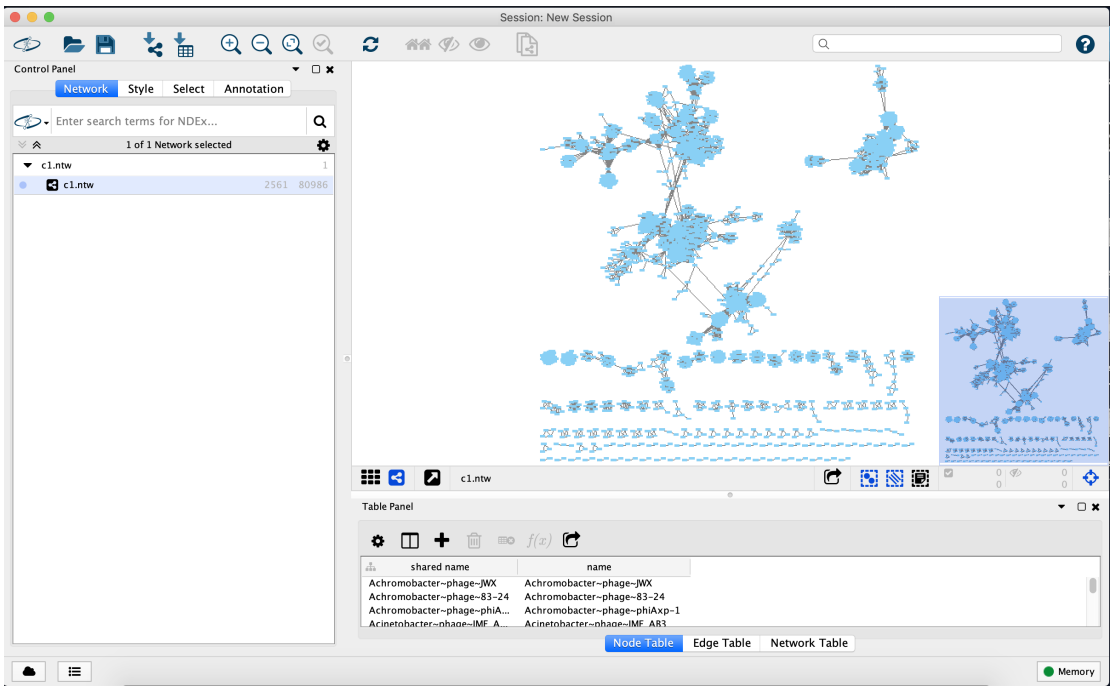
2. Click on 'Column 2' and under *Meaning*, select *Target Node* (red bullseye).
3. Click on 'Column 3' and under *Meaning*, select *Edge Attribute* (purple file).
4. Select 'Ok.' One this happens, it might take a while to load the network.



9 Results

Depending on the size of your network, Cytoscape might not automatically create a *View* for the network. Our example case is small enough so it should automatically create one. However, real data often has 100s, 1000s, 10s of 1000s of nodes and can be memory intensive.

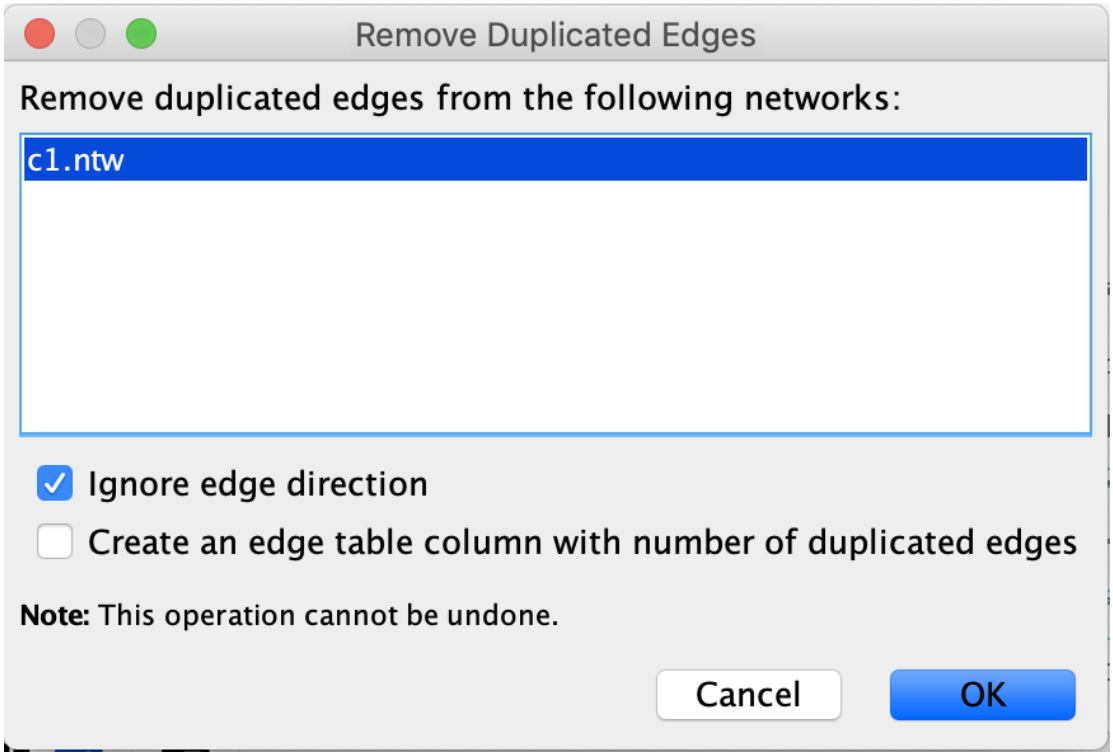
If your data is large, you can still visualize the network. A popup will appear, "Create Network Views?" Select "Ok." Once finished, the network view will be *roughly* ordered by cluster size!



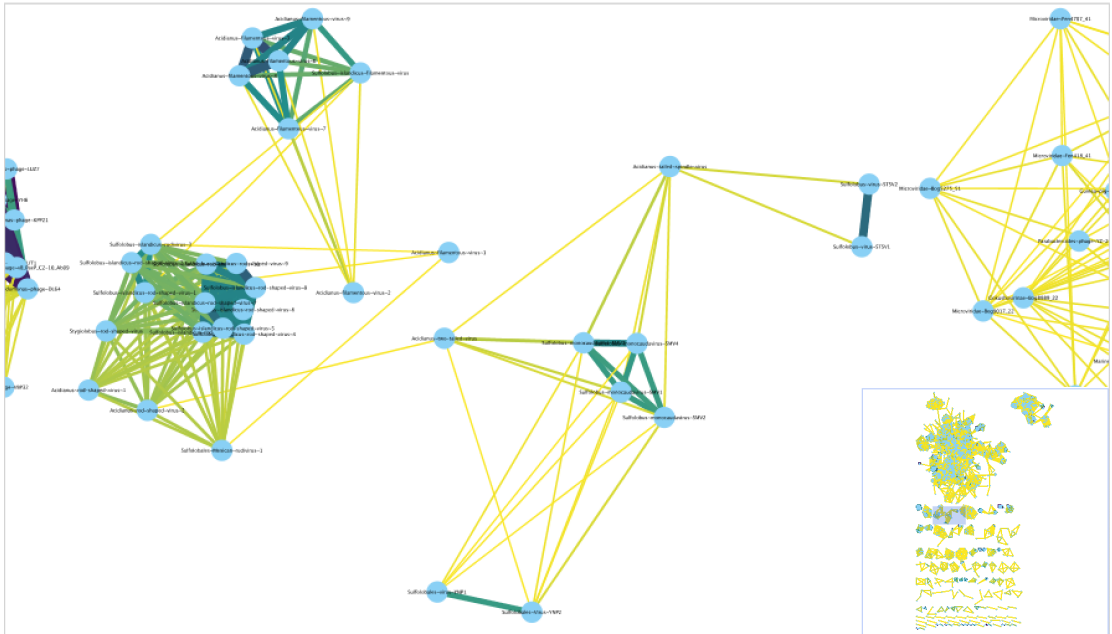
10 Cleaning Up

There's *a lot* of options in Cytoscape - far more than can be elborated here. Play around and try different things. Although to make this look a bit more presentable you'll want to remove duplicated edges and apply a visual style.

Remove duplicate edges...



Apply a visual style....



There's a lot more that can be done with vConTACT2 outputs and Cytoscape, but are a little lengthy to detail in a protocol. Experiment!