# Applying vContact to Viral Sequences and Visualizing the Output (Cyverse) Version 3

**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 *actual* reads from Ocean Sampling Day (2014) and processes them entirely within Cyverse, a NSF-supported cyberinfrastructure.

**Citation:** Benjamin Bolduc Applying vContact to Viral Sequences and Visualizing the Output (Cyverse). **protocols.io** dx.doi.org/10.17504/protocols.io.gwcbxaw

**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/

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

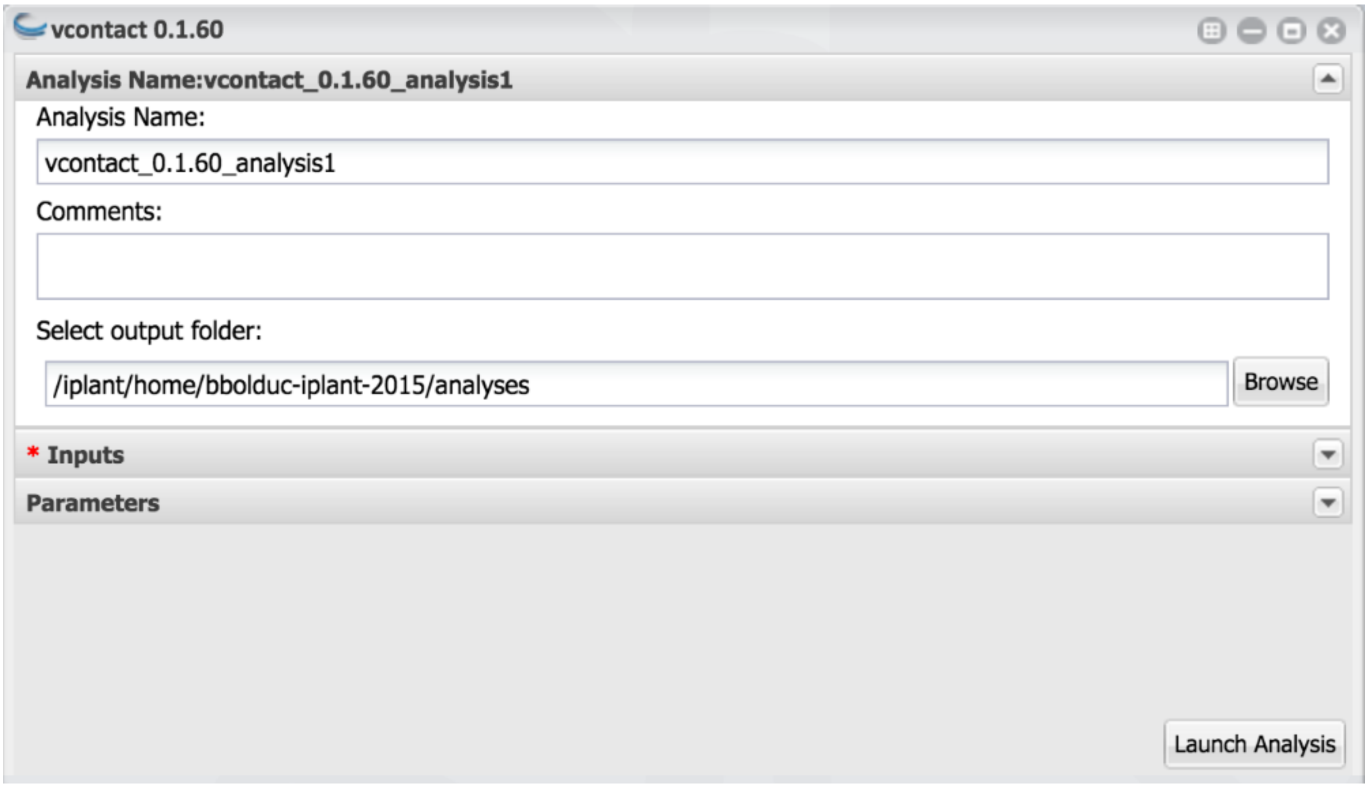
## Protocol

Affiliating contigs through their shared proteins

**Step 1.**

# Open vContact

Open vContact-0.1.60 from 'Apps'



vcontact 0.1.60

Analysis Name:vcontact_0.1.60_analysis1

Analysis Name:

vcontact_0.1.60_analysis1

Comments:

Select output folder:

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

Affiliating contigs through their shared proteins

**Step 2.**

## Select Inputs

Select the 'Inputs tab.

For **Protein clusters info file**:

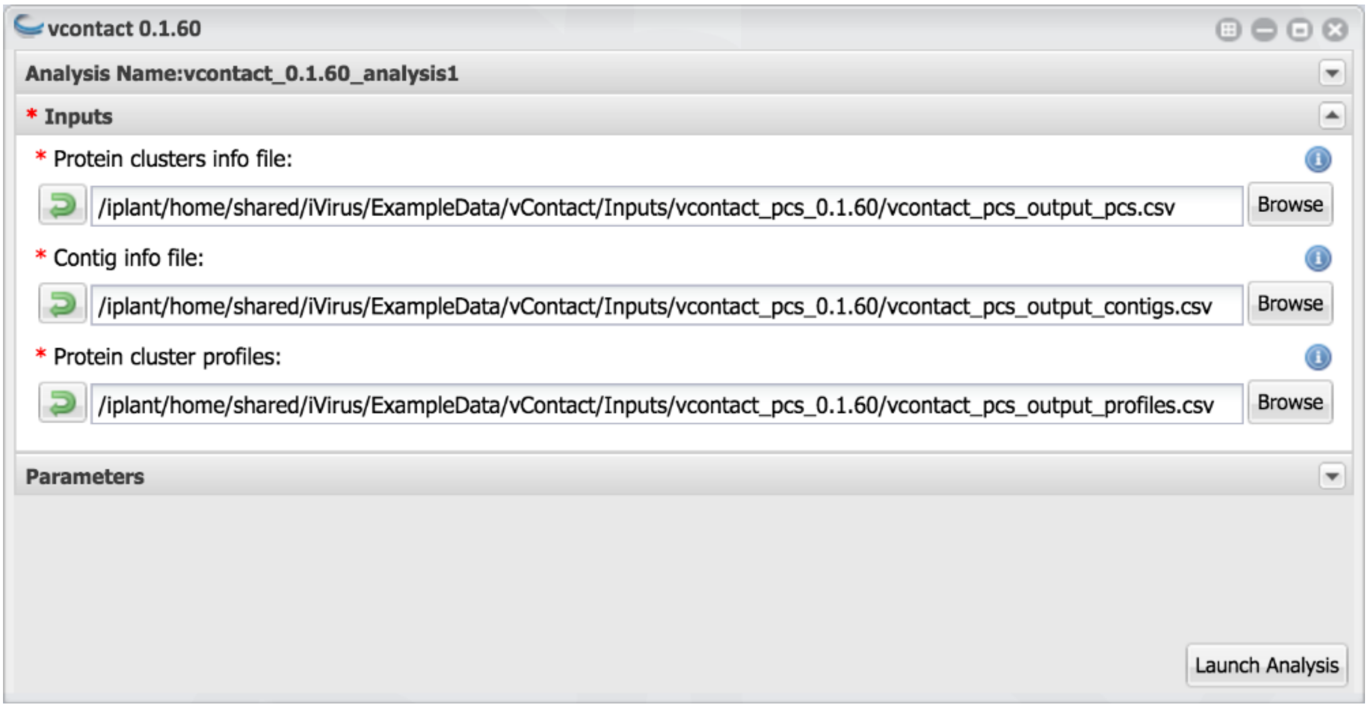
- This file contains the "id", "size", "annotated" and "keys" for each PC in the dataset, with id (PC ID), size (number of genes within the PC), annotated (number of genes including annotation) and keys (;-separated list of key terms extracted from gene annotations).
- Navigate to *Community Data* --> *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.



* Protein clusters info file: /iplant/home/shared/iVirus/ExampleData/vContact/Inputs/vcontact_pcs_0.1.60/vcontact_pcs_output_pcs.csv Browse

* Contig info file: /iplant/home/shared/iVirus/ExampleData/vContact/Inputs/vcontact_pcs_0.1.60/vcontact_pcs_output_contigs.csv Browse

* Protein cluster profiles: /iplant/home/shared/iVirus/ExampleData/vContact/Inputs/vcontact_pcs_0.1.60/vcontact_pcs_output_profiles.csv Browse

NOTES

**Benjamin Bolduc** 05 Jan 2017

The inputs for this step were generated by vContact-PCs-0.1.60 using a tab-formatted BLASTP file and a gene-to-contig mapping file.

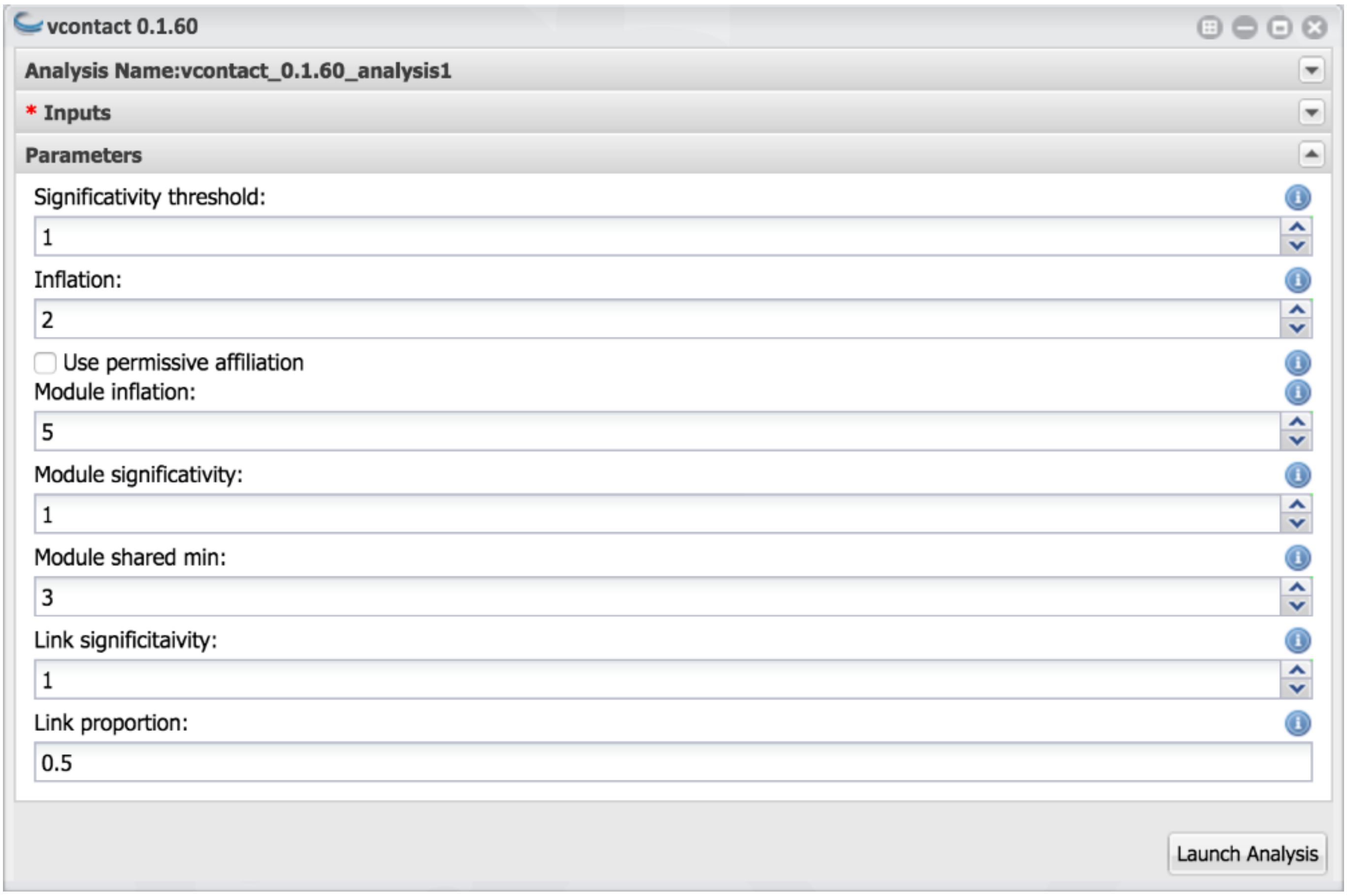
Affiliating contigs through their shared proteins

**Step 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.



Significativity threshold: 1

Inflation: 2

□ Use permissive affiliation

Module inflation: 5

Module significativity: 1

Module shared min: 3

Link significitaivity: 1

Link proportion: 0.5

Affiliating contigs through their shared proteins

**Step 4.**

# Launch Analysis

Run the job!

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

Affiliating contigs through their shared proteins

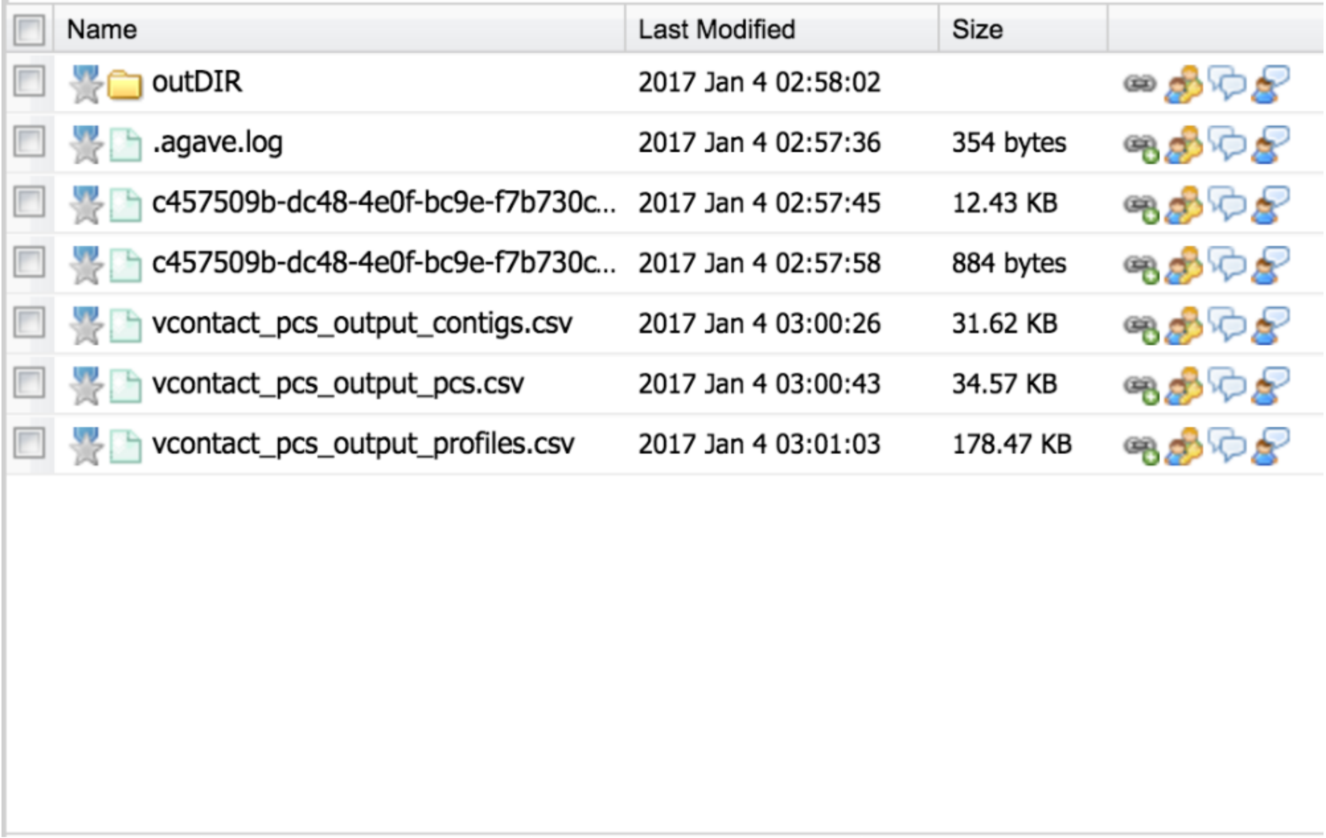
**Step 5.**

# Results

vContact will generate a results folder (outDIR in this example) with network files (*.ntw), contig clustering information (cc_*) and modules (mod_*). The network files can be imported into Cytoscape (below) to visualize the modules and the contig clusters.

Expected results can be found from the 'Outputs' directory of vContact.

EXPECTED RESULTS



| Name | Last Modified | Size |
|---|---|---|
| outDIR | 2017 Jan 4 02:58:02 | |
| .agave.log | 2017 Jan 4 02:57:36 | 354 bytes |
| c457509b-dc48-4e0f-bc9e-f7b730c... | 2017 Jan 4 02:57:45 | 12.43 KB |
| c457509b-dc48-4e0f-bc9e-f7b730c... | 2017 Jan 4 02:57:58 | 884 bytes |
| vcontact_pcs_output_contigs.csv | 2017 Jan 4 03:00:26 | 31.62 KB |
| vcontact_pcs_output_pcs.csv | 2017 Jan 4 03:00:43 | 34.57 KB |
| vcontact_pcs_output_profiles.csv | 2017 Jan 4 03:01:03 | 178.47 KB |

Notable files:

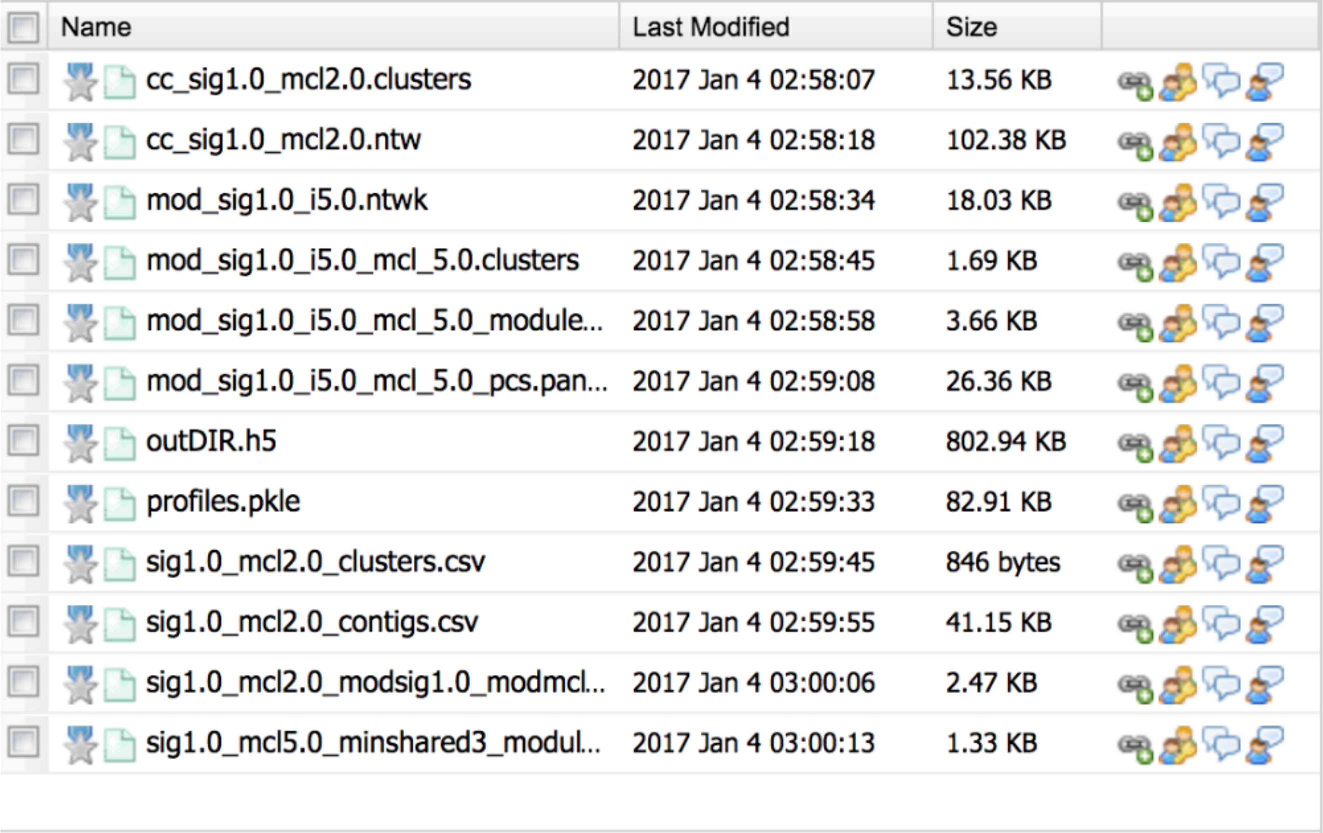
**cc_sig1.0_mcl2.0.clusters**: Contains each contig cluster's members, one line per cluster, with each member tab separated

**cc_sig1.0_mcl2.0.ntw:** Contig clusters network edge file, with source member, target member, and significance score strength

**mod_sig1.0_i5.0_mcl_5.0.clusters**: Contains each module's members, one line per module, with each member tab separated

**mod_sig1.0_i5.0.ntwk**: Module network edge file, with source PC, target PC, and significance score strength

*.pandas: These files are generated by python-pandas to store a raw copy of the data parsed above



| Name | Last Modified | Size |
|---|---|---|
| cc_sig1.0_mcl2.0.clusters | 2017 Jan 4 02:58:07 | 13.56 KB |
| cc_sig1.0_mcl2.0.ntw | 2017 Jan 4 02:58:18 | 102.38 KB |
| mod_sig1.0_i5.0.ntwk | 2017 Jan 4 02:58:34 | 18.03 KB |
| mod_sig1.0_i5.0_mcl_5.0.clusters | 2017 Jan 4 02:58:45 | 1.69 KB |
| mod_sig1.0_i5.0_mcl_5.0_module... | 2017 Jan 4 02:58:58 | 3.66 KB |
| mod_sig1.0_i5.0_mcl_5.0_pcs.pan... | 2017 Jan 4 02:59:08 | 26.36 KB |
| outDIR.h5 | 2017 Jan 4 02:59:18 | 802.94 KB |
| profiles.pkle | 2017 Jan 4 02:59:33 | 82.91 KB |
| sig1.0_mcl2.0_clusters.csv | 2017 Jan 4 02:59:45 | 846 bytes |
| sig1.0_mcl2.0_contigs.csv | 2017 Jan 4 02:59:55 | 41.15 KB |
| sig1.0_mcl2.0_modsig1.0_modmcl... | 2017 Jan 4 03:00:06 | 2.47 KB |
| sig1.0_mcl5.0_minshared3_modul... | 2017 Jan 4 03:00:13 | 1.33 KB |

Cluster Visualization

**Step 6.**

## Open Cytoscape

Open Cytoscape *on your local machine*.

Cluster Visualization

**Step 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...

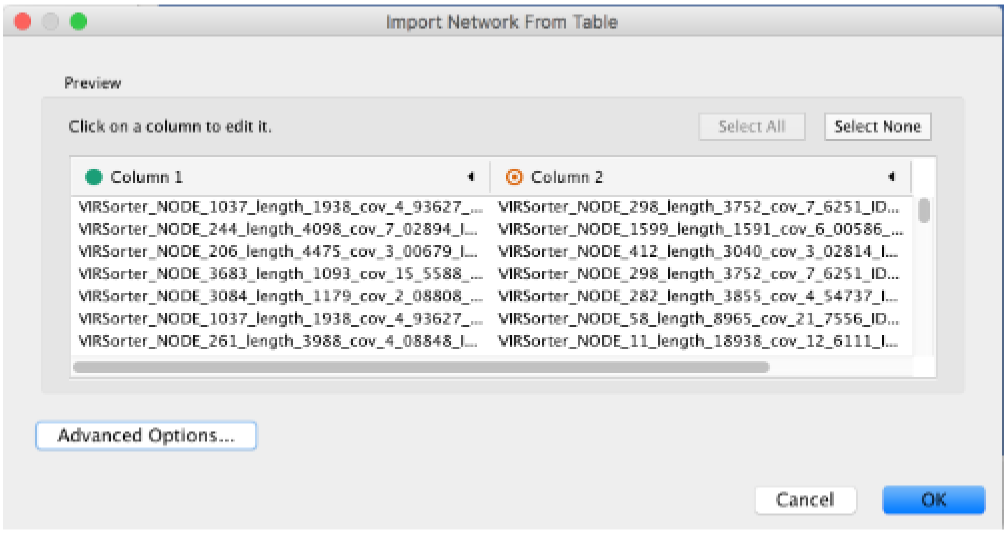
Select the contig *.ntw (typically *cc_sig1.0_mcl2.0.ntw).

Cluster Visualization

**Step 8.**

# Import Network File

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)
2. Click on 'Column 1' and under *Meaning*, select *Source Node* (little green button).
3. Click on 'Column 2' and under *Meaning*, select *Target Node* (red bullseye).
4. Click on 'Column 3' and under *Meaning*, select *Edge Attribute* (purple file).
5. Select 'Ok.' One this happens, it might take a while to load the network.



| ● Column 1 | ⊙ Column 2 |
|---|---|
| VIRSorter_NODE_1037_length_1938_cov_4_93627_... | VIRSorter_NODE_298_length_3752_cov_7_6251_ID... |
| VIRSorter_NODE_244_length_4098_cov_7_02894_l... | VIRSorter_NODE_1599_length_1591_cov_6_00586_... |
| VIRSorter_NODE_206_length_4475_cov_3_00679_l... | VIRSorter_NODE_412_length_3040_cov_3_02814_l... |
| VIRSorter_NODE_3683_length_1093_cov_15_5588_... | VIRSorter_NODE_298_length_3752_cov_7_6251_ID... |
| VIRSorter_NODE_3084_length_1179_cov_2_08808_... | VIRSorter_NODE_282_length_3855_cov_4_54737_l... |
| VIRSorter_NODE_1037_length_1938_cov_4_93627_... | VIRSorter_NODE_58_length_8965_cov_21_7556_ID... |
| VIRSorter_NODE_261_length_3988_cov_4_08848_l... | VIRSorter_NODE_11_length_18938_cov_12_6111_l... |

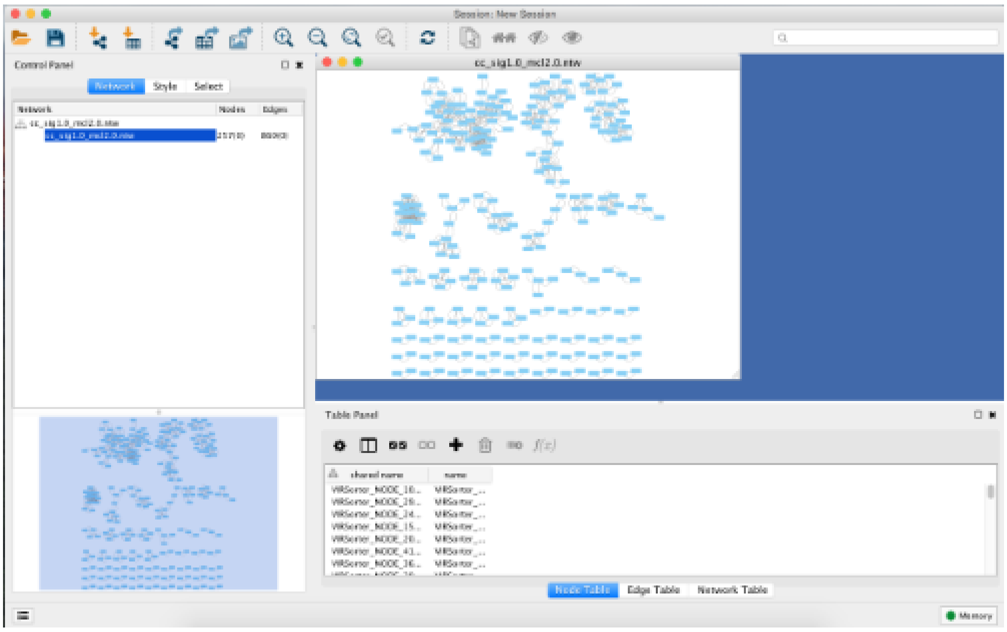
Cluster Visualization

**Step 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!



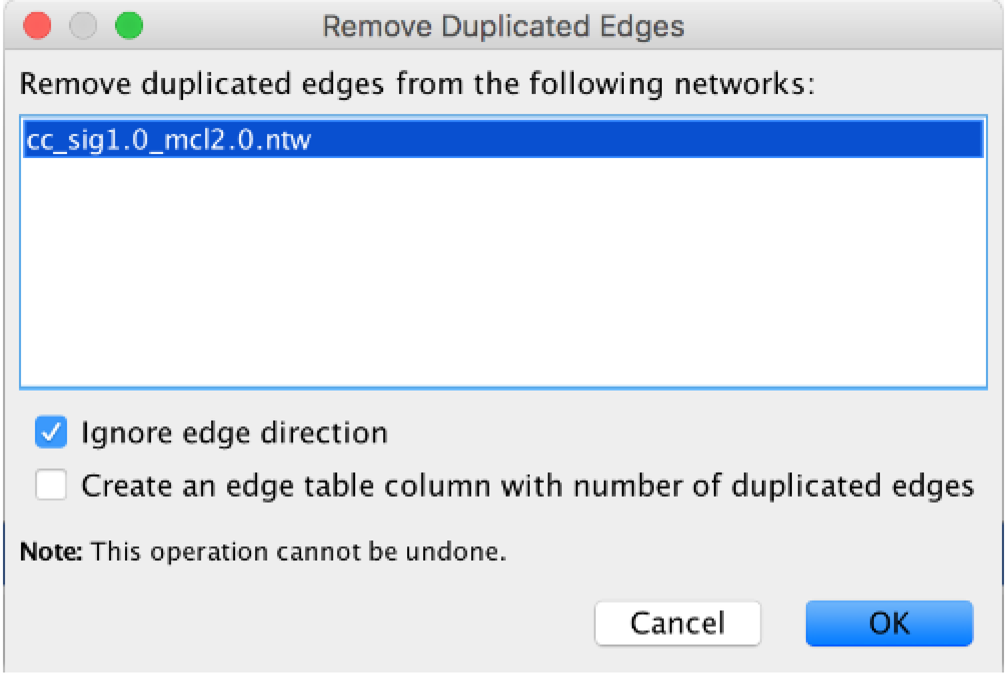
Cluster Visualization

**Step 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...



Remove duplicated edges from the following networks:

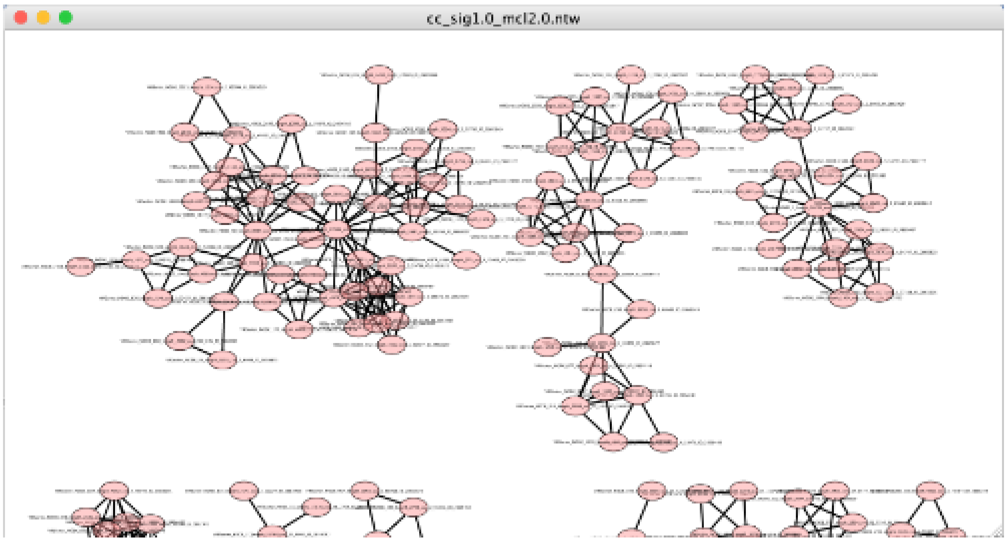
cc_sig1.0_mcl2.0.ntw

☑ Ignore edge direction

□ Create an edge table column with number of duplicated edges

**Note:** This operation cannot be undone.

Apply a visual style....



cc_sig1.0_mcl2.0.ntw