Feb 15, 2019 Version 5

Applying vContact to Viral Sequences and Visualizing the Output (Cyverse) V.5

In 2 collections

DOI

dx.doi.org/10.17504/protocols.io.x5xfq7n

Benjamin Bolduc¹

¹The Ohio State University

VERVE Net

Sullivan Lab 1 more workspace



Benjamin Bolduc

The Ohio State University



DOI: dx.doi.org/10.17504/protocols.io.x5xfq7n

External link: https://doi.org/10.1101/533240

Protocol Citation: Benjamin Bolduc 2019. Applying vContact to Viral Sequences and Visualizing the Output (Cyverse). **protocols.io** <u>https://dx.doi.org/10.17504/protocols.io.x5xfq7n</u>

Manuscript citation:

Jang, H. Bin, Bolduc, B., Zablocki, O., Kuhn, J., Roux, S., Adriaenssens, E., ... Sullivan, M. (2019). Gene sharing networks to automate genome-based prokaryotic viral taxonomy. BioRxiv, 533240. <u>https://doi.org/10.1101/533240</u>

License: This is an open access protocol distributed under the terms of the **<u>Creative Commons Attribution License</u>**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working We use this protocol and it's working

Created: February 15, 2019

Last Modified: February 15, 2019

Protocol Integer ID: 20375

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 <u>Ocean Sampling Day (2014)</u> and processes them entirely within Cyverse, a NSF-supported cyberinfrastructure.

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 <u>register</u> 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 vContact2-0.9.5 from 'Apps'

🗲 vConTACT2 0.9.5	
Analysis Name:vConTACT2_0.9.5_analysis1	
Analysis Name:	
vConTACT2_0.9.5_analysis1	
Comments:	
Select output folder:	
/iplant/home/bbolduc-iplant-2015/analyses	Browse
Inputs	
* Parameters	•
	Launch Analysis

Starting menu for the vConTACT2 app in the CyVerse Discovery Environment

Note

vConTACT2 is constantly in a state of active development. Always check back here for newer protocols that *often* simply update the version. And always use the most recent version whenever possible.

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

/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.

vConTACT2 0.9.5		•	ε
Analysis Name:vConTACT2_0.9.5_analysis1		•	
Inputs			
FASTA-formatted amino acid protein file.:			
/iplant/home/shared/iVirus/ExampleData/vContact2/Inputs/VIRSorter_viral_prots.faa	Browse		
BLASTP/Diamond results file for creating PCs. WILL BE SKIPPED IF PROTEIN FILE IS PROVIDED.:			
Select a file or folder	Browse		
Gene-to-Genome mapping file. MUST BE PROVIDED if using raw proteins or BLASTP/Diamond.:			
/iplant/home/shared/iVirus/ExampleData/vContact2/Inputs/gene2genome_proteins.csv	Browse		
Contig info file (legacy vConTACT). MUST be used with PCs and profiles info files.:			
Select a file or folder	Browse		
PCs info file (legacy vConTACT). MUST be used with contig and profiles info files.:			
Select a file or folder	Browse		
Profiles info file (legacy vConTACT). MUST be used with contig and PCs info files.:			
Select a file or folder	Browse		
* Parameters		•	
	Launch A	nalys	is

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*.

vConTACT2 0.9.5	•	
Analysis Name:vConTACT2_0.9.5_analysis1		
Inputs		-
* Parameters		
* Protein-protein similarity method.:		
Diamond	~	- 10
* Reference database:		- 10
NCBI Bacterial and Archaeal Viral RefSeq V85 with ICTV + NCBI taxonomy	~	- 10
* PC generation method.:		- 10
MCL	~	- 10
* VC generation method.:		- 10
ClusterONE	~	- 10
Optimize hierarchical distance.		
BLASTP e-value:		- 10
0.0001	*	
Max overlap for PC clusters:		
0.1	~	
Penalty to use for PC creation.:		
2	~	
Haircut value for PCs:		
0.1	v	
Inflation value for PCs (MCL ONLY).:		
2	~	
Inflation value for VCs (MCL ONLY).:		
2	~	
Minimum Density for VCs (ClusterONE ONLY):		
0.3	~	
Minimum VC size:		
2	~	
	aunch Ai	nalysis

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 <u>Cytoscape</u> (more Initial app output directory structurebelow) to visualize the modules and the contig clusters.

Data: Outputs				0	
Jpload • File • Edit •	Download 🔹 Share 👻 Metadata 🔹 🍣 Refresh	1		•	Trash
lavigation	• Outputs			Details	
📁 bbolduc-iplant-2015	Viewing: /iplant/home/shared/iVirus/Exa	riewing: /iplant/home/shared/iVirus/ExampleData/vContact2/Outputs Select a file or folder t			
 GCommunity Data GShared With Me GTash Favorites 	Name	Last Modified	Size		
	Contact_Output	2018 Dec 19 11:44:40	:		
	UUID.err	2018 Dec 19 11:41:00	24.14 KB		
	🔲 🛅 UUID.out	2018 Dec 19 11:41:27	17.25 KB		
	VIRSorter_viral_prots.faa	2018 Dec 19 11:41:04	960.52 KB		
	gene2genome_proteins.csv	2018 Dec 19 11:41:19	508.51 KB		

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.

Seta: vContact_Output	L			8 🗆 🖨 🖸
Upload - File - Edit - Dov	vnload 🗸 Share 🖌 Metadata 🗸 🍣 Refresh			🔍 Trash 🕶
Navigation	VContact_Output			Details
bolduc-iplant-2015	Viewing: /iplant/home/shared/iVirus/ExampleData	/vContact2/Outputs/vCo	ntact_Output	Select a file or folder to view its
🕨 📁 Community Data	Name	Last Modified 🔺	Size	details
 C Shared With Me C Trash Favorites 	merged_df.csv	2018 Dec 19 11:44:50	207.15 KB	
	VConTACT_contigs.csv	2018 Dec 19 11:44:50	97.24 KB	
	merged.self-diamond.tab_mcxload.tab	2018 Dec 19 11:44:52	4.26 MB	
	VConTACT_proteins.csv	2018 Dec 19 11:44:52	16.92 MB	
	modules.ntwk	2018 Dec 19 11:45:10	61.43 MB	
	VConTACT_pcs.csv	2018 Dec 19 11:46:03	4.95 MB	
	merged.self-diamond.tab.abc	2018 Dec 19 11:46:22	112.6 MB	
	modules_mcl_5.0_pcs.pandas	2018 Dec 19 11:48:53	4.77 MB	
	merged.self-diamond.tab_mcl20.clusters	2018 Dec 19 11:49:12	3.0 MB	
	C1.ntw	2018 Dec 19 11:49:22	5.53 MB	
	sig1.0_mcl2.0_contigs.csv	2018 Dec 19 11:49:42	350.45 KB	
	modules_mcl_5.0_modules.pandas	2018 Dec 19 11:49:44	27.3 KB	
	E merged.faa	2018 Dec 19 11:49:46	62.5 MB	
	sig1.0_mcl2.0_modsig1.0_modmcl5.0_mi	2018 Dec 19 11:52:34	23.05 KB	
	Displaying 1 - 14 of 24		0 item(s)	

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.

	ownload • Share • Metadata • 🥪 Refresh			irash •
Navigation	vContact_Output	Details		
 C bbolduc-iplant-2015 C Community Data Shared With Me Trash Favorites 	Viewing: /iplant/home/shared/iVirus/ExampleData/	Select a file or folder to view its details		
	Name	Last Modified 🔺	Size	
	sig1.0_mcl2.0_contigs.csv	2018 Dec 19 11:49:42	350.45 KB	
	modules_mcl_5.0_modules.pandas	2018 Dec 19 11:49:44	27.3 KB	
	merged.faa	2018 Dec 19 11:49:46	62.5 MB	
	sig1.0_mcl2.0_modsig1.0_modmcl5.0_mi	2018 Dec 19 11:52:34	23.05 KB	
	modules_mcl_5.0.clusters	2018 Dec 19 11:52:35	143.15 KB	
	merged.dmnd	2018 Dec 19 11:52:37	65.44 MB	
	VConTACT_profiles.csv	2018 Dec 19 11:53:11	6.76 MB	
	sig1.0_mcl2.0_clusters.csv	2018 Dec 19 11:53:35	10.85 KB	
	merged.self-diamond.tab	2018 Dec 19 11:53:38	221.2 MB	
	viral_cluster_overview.csv	2018 Dec 19 11:56:19	131.09 KB	
	genome_by_genome_overview.csv	2018 Dec 19 11:56:22	302.21 KB	
	C1.clusters	2018 Dec 19 11:56:24	104.03 KB	
	sig1.0_mcl5.0_minshared3_modules.csv	2018 Dec 19 11:56:25	15.37 KB	
	merged.self-diamond.tab.mci	2018 Dec 19 11:56:27	61.99 MB	

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 \rightarrow Import \rightarrow Network \rightarrow 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:

- Inb		
Preview		
Click on a column to edit it.	Select All Select N	None
📄 Achromobacter~phage~JWX Achromobac	:ter~phag	
Achromobacter~phage~phiAxp-1 Achromob Acinetobacter~phage~IME_AB3 Achromobact Burkholderia~phage~BcepGomr Achromobacc Burkholderia~phage~KL1 Achromobacter~ph Paracoccus~phage~vB_PmaS_IMEP1 Achromo Phage~phiJL001 Achromobacter~phage~83-2 Pseudomonas~phage~73 Achromobacter~ph Pseudomonas~phage~M6 Achromobacter~ph	acter~ er~ph ter~ph age~8 ybacter 24 6.8 iage~8 hage~8	
Advanced Options		
	Cancel	C

- 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.

Column 1	•	⊙ Column 2	 Column 3
Achromobacter~phage~JWX		Achromobacter~phage~83-24	137.5156184166
Achromobacter~phage~phiAxp-1		Achromobacter~phage~83-24	20.0508646777
Acinetobacter~phage~IME_AB3		Achromobacter~phage~83-24	8.4953442133
Burkholderia~phage~BcepGomr		Achromobacter~phage~83-24	16.64788164436
Burkholderia~phage~KL1		Achromobacter~phage~83-24	10.98340786382
Paracoccus~phage~vB_PmaS_IMEP1		Achromobacter~phage~83-24	8.83530649290
Phage~phiJL001		Achromobacter~phage~83-24	6.83481558194

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...

	Remove Duplicate	ed Edges	
Remove duplicated	dges from the fo	llowing network	s:
c1.ntw			
🗹 Ignore edge dire	ction		
Create an edge	able column with	number of dup	licated edges
Note: This operation ca	not be undone.		
		Cancel	ОК

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!