Clustering Viral Genomes in iVirus

Benjamin Bolduc, Simon Roux

Abstract

Cluster genomes is script that clusters genomes at a set nucleotide identity and coverage length. Additonally, it offers the ability to cluster sequences whose ends may not align correspondingly, i.e. the special cases of assembled, circular viral genomes that are treated as 'linear' by other sequence-clustering software (that can "miss" the ends).

This is 'beta-like' software that has been vetted, though has not been as thoroughly tested as other widely-recognized clustering tools, such as CD-HIT, UCLUST, etc... As with all software, please examine the final results to see if they make sense.

Citation: Benjamin Bolduc, Simon Roux Clustering Viral Genomes in iVirus. **protocols.io** dx.doi.org/10.17504/protocols.io.gwebxbe
Published: 06 Jan 2017

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/

All source code is available at the Sullivan lab bitbucket repository, MAVERICLab - at stampede-clustergenomes.

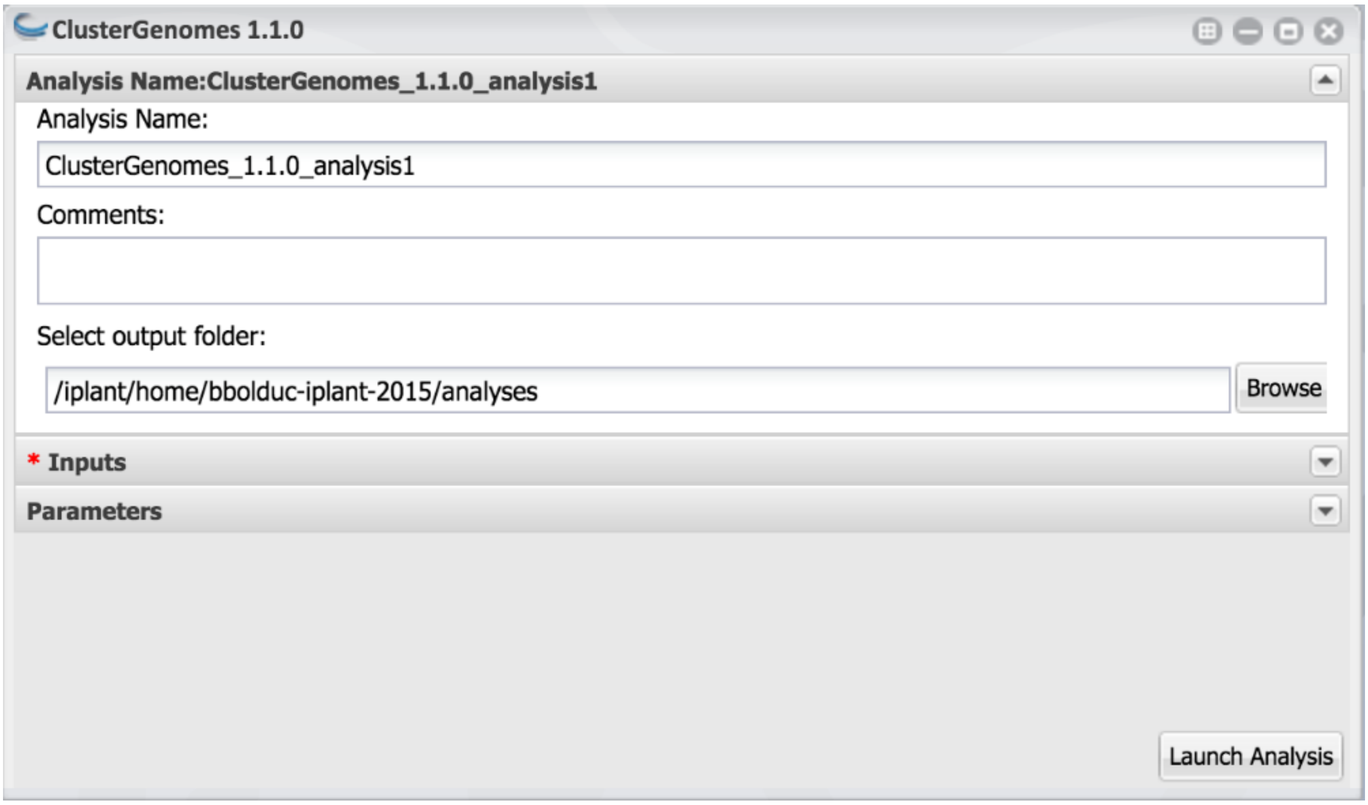
Protocol

Cluster Genomes

Step 1.

Open ClusterGenomes

Open ClusterGenomes-1.1.0 (beta from 'Apps'



ClusterGenomes 1.1.0

Analysis Name:ClusterGenomes_1.1.0_analysis1

Analysis Name: ClusterGenomes_1.1.0_analysis1

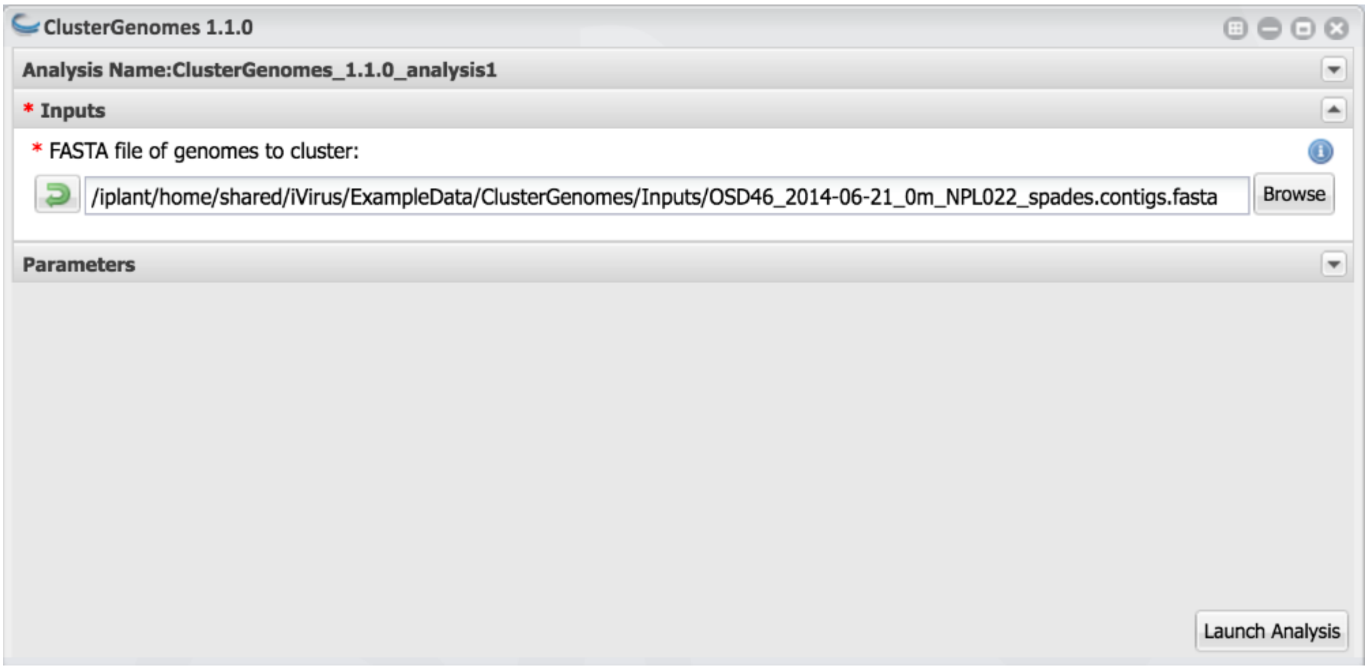
Comments:

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

Cluster Genomes

Step 2.

Select Input



ClusterGenomes 1.1.0

Analysis Name:ClusterGenomes_1.1.0_analysis1

* Inputs

* FASTA file of genomes to cluster: /iplant/home/shared/iVirus/ExampleData/ClusterGenomes/Inputs/OSD46_2014-06-21_0m_NPL022_spades.contigs.fasta

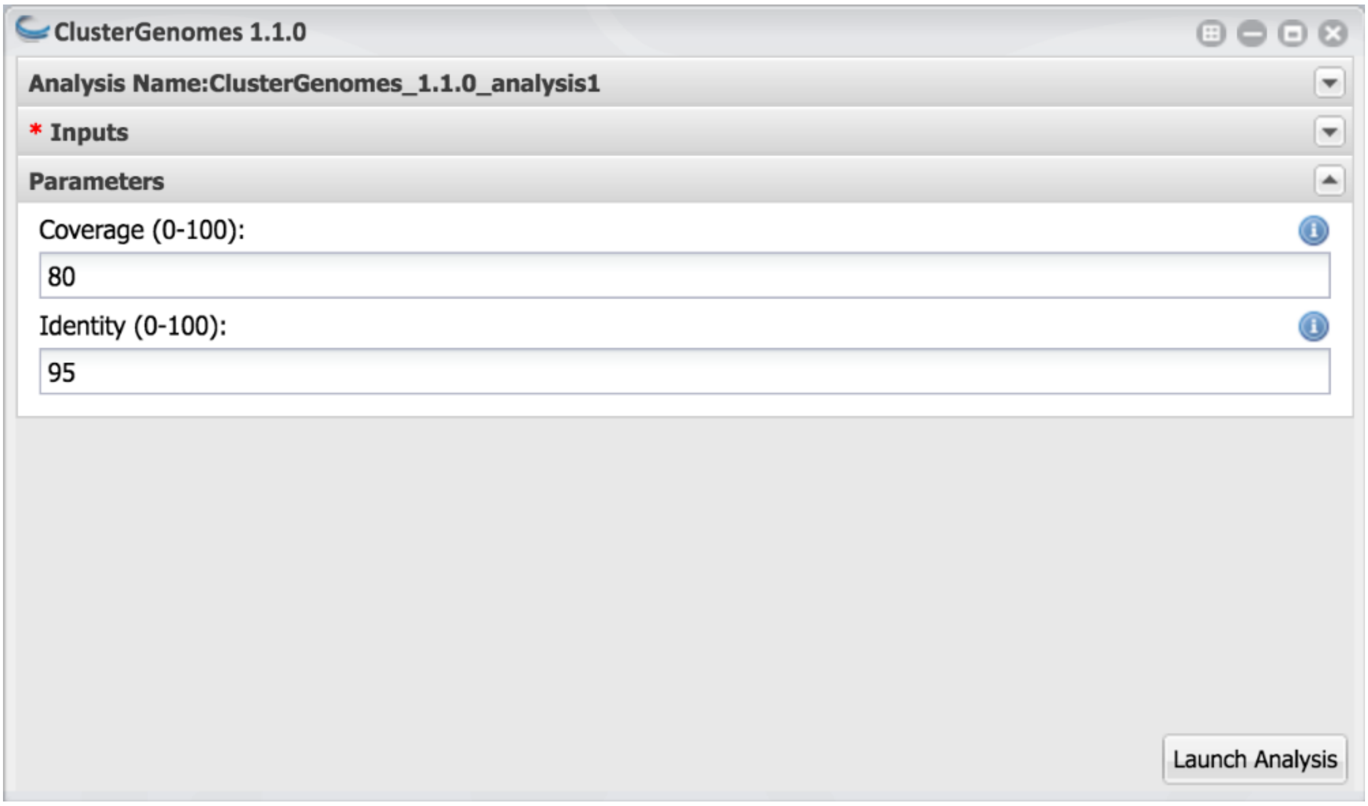
Cluster Genomes

Step 3.

Select Parameters

Coverage: The length one sequence needs to "cover" across another sequence for it to be considered within the same cluster.

Identity: The nucleotide identity required across the coverage region for one sequence to be considered aligned with another.



ClusterGenomes 1.1.0

Analysis Name:ClusterGenomes_1.1.0_analysis1

Coverage (0-100): 80

Identity (0-100): 95

Cluster Genomes

Step 4.

Launch Analysis

Run the job!

Depending on the number of sequences in the data file, the job could take minutes to several hours.

Cluster Genomes

Step 5.

Results

Expected results can be found in the 'Output' directory.

*_95-80.clstr: contains a list of each cluster, its members and the identify found to group the sequences, with the seed sequence denoted

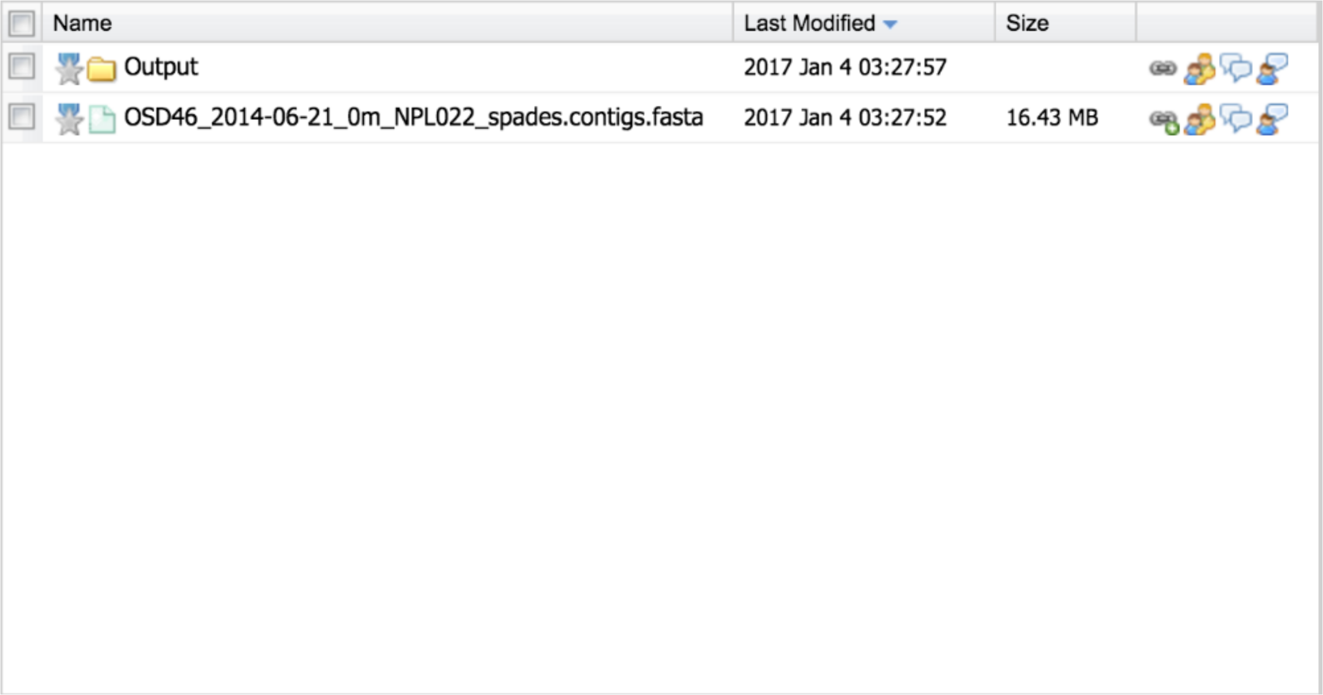
*_95-80_seeds.fna: contains the nucleotide sequence for the "seed" of each cluster

-nucmer.out.: both coords and delta files are generated by nucmer, the underlying tool responsible for identifying overlapping regions between sequences

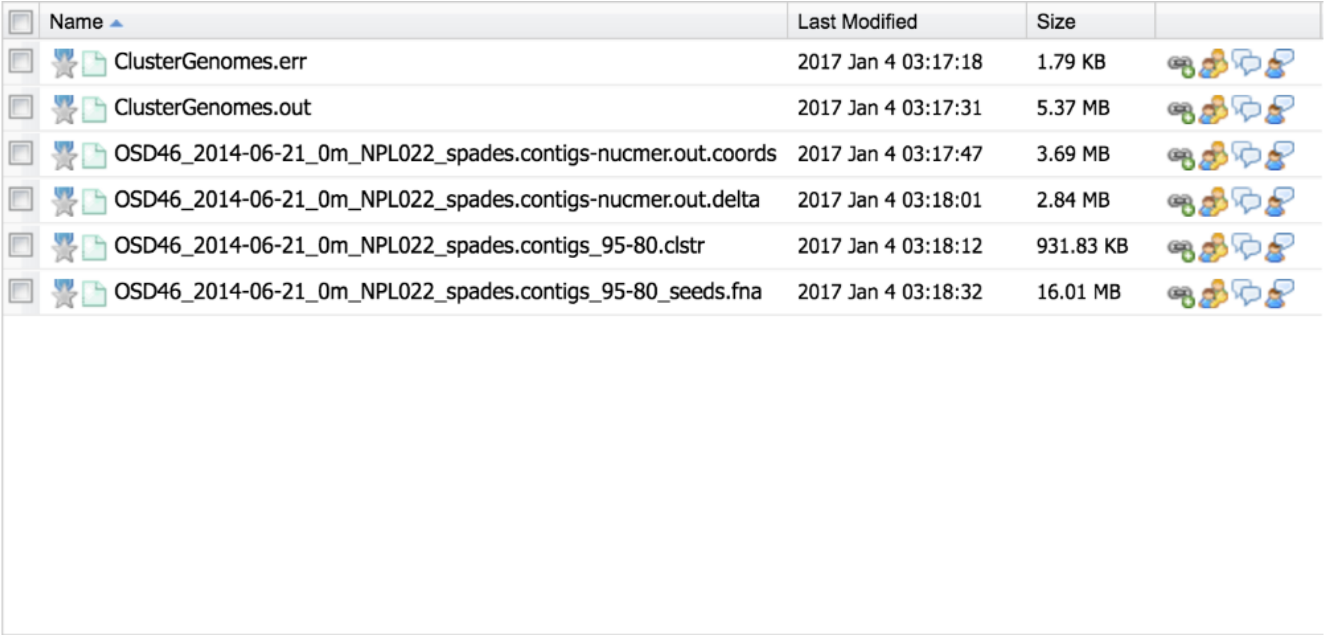
ClusterGenomes.out: list of matches found between sequences

ClusterGenomes.err: general output from the run

EXPECTED RESULTS



□	Name	Last Modified	Size
□	Output	2017 Jan 4 03:27:57	
□	OSD46_2014-06-21_0m_NPL022_spades.contigs.fasta	2017 Jan 4 03:27:52	16.43 MB



□	Name ▲	Last Modified	Size
□	ClusterGenomes.err	2017 Jan 4 03:17:18	1.79 KB
□	ClusterGenomes.out	2017 Jan 4 03:17:31	5.37 MB
□	OSD46_2014-06-21_0m_NPL022_spades.contigs-nucmer.out.coords	2017 Jan 4 03:17:47	3.69 MB
□	OSD46_2014-06-21_0m_NPL022_spades.contigs-nucmer.out.delta	2017 Jan 4 03:18:01	2.84 MB
□	OSD46_2014-06-21_0m_NPL022_spades.contigs_95-80.clstr	2017 Jan 4 03:18:12	931.83 KB
□	OSD46_2014-06-21_0m_NPL022_spades.contigs_95-80_seeds.fna	2017 Jan 4 03:18:32	16.01 MB

NOTES

Benjamin Bolduc 06 Jan 2017

TIP: Depending on the sequences in the dataset, a few or a lot of sequences could cluster. Just because only a few sequences are found to cluster with others does not indicate there was a problem with the data or with the software.