ECOGEO 'Omics Training: 4.3 Binning Module - Anvi'o Version 2

A. Murat Eren

Abstract

This is the hands-on interactive component to the ECOGEO Workshop Module on Binning.

Citation: A. Murat Eren ECOGEO 'Omics Training: 4.3 Binning Module - Anvi'o. **protocols.io** dx.doi.org/10.17504/protocols.io.fi7bkhn
Published: 18 Aug 2016

Before start

Before starting, please visit the ECOGEO website for more information on this "Introduction to Environmental 'Omics" training series. The site contains a pre-packaged virtual machine that can be downloaded and used to run all of the protocols in this protocols.io collection. In addition to the VM, the website contains video and presentations from our initial "Intro to Env 'Omics" workshop held at the Univ. of Hawai'i at Manoa on 25-26 Jul 2016.

Please email ‘ecogeo-join@earthcube.org’ to join the ECOGEO listserv for future updates.

This tutorial is tailored for Anvi’o v2.0.1 (installed on the VM). The purpose of this tutorial is to demonstrate some of the anvi’o capabilities using the Infant Gut dataset, which was generated, analyzed, and published by Sharon et al. (2013), and was re-analyzed in the anvi’o methods paper.

The tutorial used during the actual workshop is no longer active. A very similar tutorial is available online (http://merenlab.org/tutorials/infant-gut/) and should be used to explore the entire hands-on demonstration in the video, as well as reading up on Meren's $0.02!

Protocol

Step 1.

BAM files and contigs FASTA

In the directory 00_BAM_FILES_AND_CONTIGS you will find the the co-assembly of the time series data, contigs.fa, but you do NOT have an indexed BAM file for each sampling day. These BAM files

were generated by mapping metagenomic short reads from each sample to contigs.fa. These files are too large for the VM.

Anvi'o metagenomic workflow starts with BAM files and a FASTA file for your contigs. There are many ways to get your contigs and BAM files for your metagenomes. But we have started implementing a tutorial that describes the workflow we use to generate these files regularly: " A tutorial on assembly-based metagenomics ". Please feel free to take a look at that one, as well.

cmd COMMAND
```
$ cd /home/c-debi/ecogeo/binning/00_BAM_FILES_AND_CONTIGS
```
Step 2.

Contigs database & anvi'o merged profile

The directory 01_ANVIO_MERGED_PROFILE contains the anvi'o contigs database that is generated from contigs.fa, and the anvi'o merged profile, which is generated by merging individual anvi'o profiles for each of the BAM file in 00_BAM_FILES_AND_CONTIG.

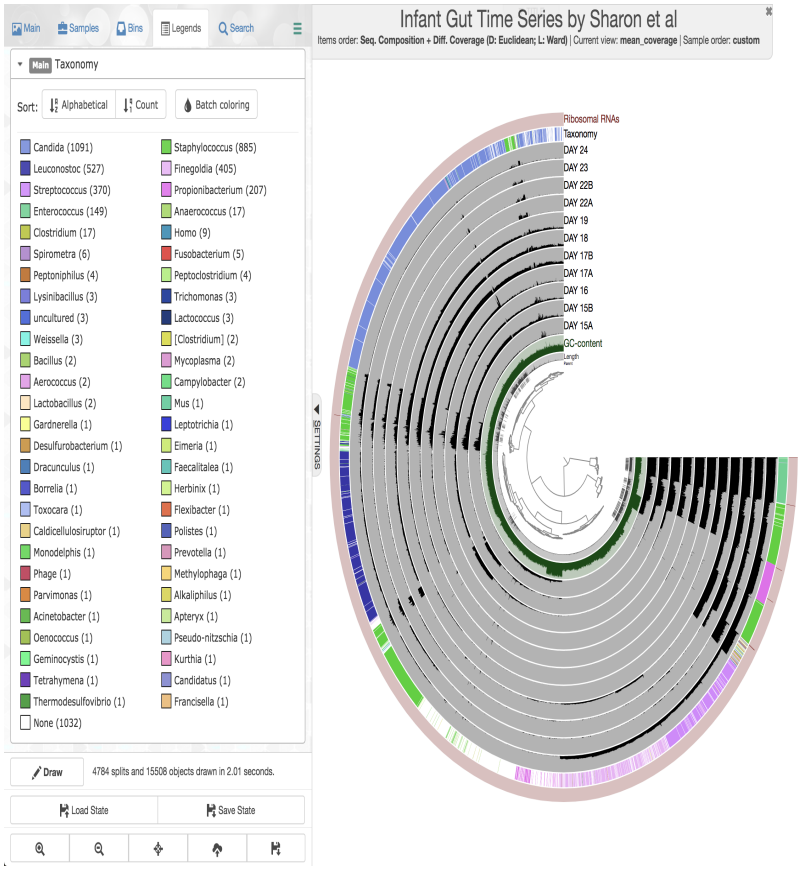
An anvi'o contigs database keeps all the information related to your contigs: positions of open reading frames, kmer frequencies for each contig, where splits start and end, functional and taxonomic annotation of genes, etc. The contigs database is an essential component of everything related to anvi'o metagenomic workflow. In contrast to the contigs database, an anvi'o profile database stores sample-specific information about contigs. Profiling a BAM file with anvi'o creates a single profile that reports properties for each contig in a single sample based on mapping results. Each profile database automatically links to a contigs database, and anvi'o can merge single profiles that link to the same contigs database into anvi'o merged profiles (which is what you have in this directory). If you would like to learn more about these, here are some direct links: creating an anvi'o contigs database , creating single anvi'o profiles, and merging anvi'o profiles.

cmd COMMAND
```
$ cd ../01_ANVIO_MERGED_PROFILE/
$ anvi-interactive -p PROFILE.db -c CONTIGS.db
```
Execute anvi-interactive interface. To kill the interactive interface, press Cntl + C in the terminal window.

Step 3.

Importing taxonomy

Centrifuge (code, pre-print) is one of the options to import taxonomic annotations into an anvi'o contigs database. Centrifuge files for the infant gut data are already in the directory 02_CENTRIFUGE_FILES of your data pack.



cmd COMMAND

```
$ anvi-import-taxonomy -c CONTIGS.db -i ../02_CENTRIFUGE_FILES/*.tsv -p centrifuge
$ anvi-interactive -p PROFILE.db -c CONTIGS.db
```

Adds an additional layer with taxonomy.

NOTES

Elisha Wood-Charlson 10 Aug 2016

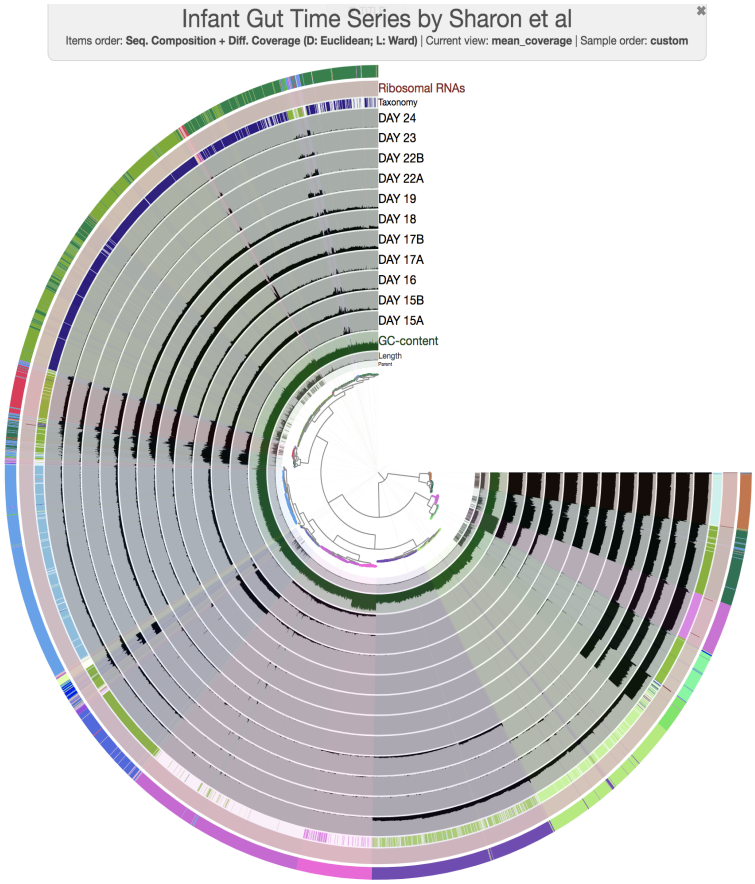
In the Layers tab find the *Taxonomy* layer, set its height to 200, and click *Draw* again. Then click *Save State* button, and overwrite the default state. This will make sure anvi'o remembers to make the height of that layer 200px the next time you run the interactive interface!

Step 4.

Importing external binning results

The directory 03_EXTERNAL_BINNING_RESULTS contains number of files describe the binning of contigs in the infant gut data based on various binning approaches (see annotation).

There is a SHARON_et_al.txt files with BLAST searched sequences in bins identified by the authors of the study so our contigs will have matching names.



cmd COMMAND

```
$ anvi-import-collection -c CONTIGS.db -p PROFILE.db -C CONCOCT --contigs-
mode ../03_EXTERNAL_BINNING_RESULTS/CONCOCT.txt
$ anvi-interactive -p PROFILE.db -c CONTIGS.db
```

Imports CONCOCT file that describes the results of an external binning effort into your profile database. Once the display is ready, click Bins > Load bin collection > CONCOCT > Load.

NOTES

Elisha Wood-Charlson 10 Aug 2016

Three files are courtesy of Elaina Graham, who used GroopM (v0.3.5), MetaBat (v0.26.3), and MaxBin (v2.1.1) to bin contigs using BAM files and the contigs FASTA in 01_ANVIO_MERGED_PROFILE.

For future references, here are the parameters:
# GroopM v0.3.5 (followed the general workflow on their manual)
$ groopm parse groopm.db contigs.fa [list of bam files]
$ groopm core groopm.db -c 1000 -s 10 -b 1000000
$ groopm recruit groopm.db -c 500 -s 200
$ groopm extract groopm.db contigs.fa
# MetaBat v0.26.3 (used jgi_summarize_bam_contig_depths to get a depth file from BAM files).
$ metabat -i contigs.fa -a depth.txt -o bin
# MaxBin v2.1.1
$ run_MaxBin.pl -contig contigs.fa -out maxbin_IGM -abund_list [list of

all coverage files in associated format]

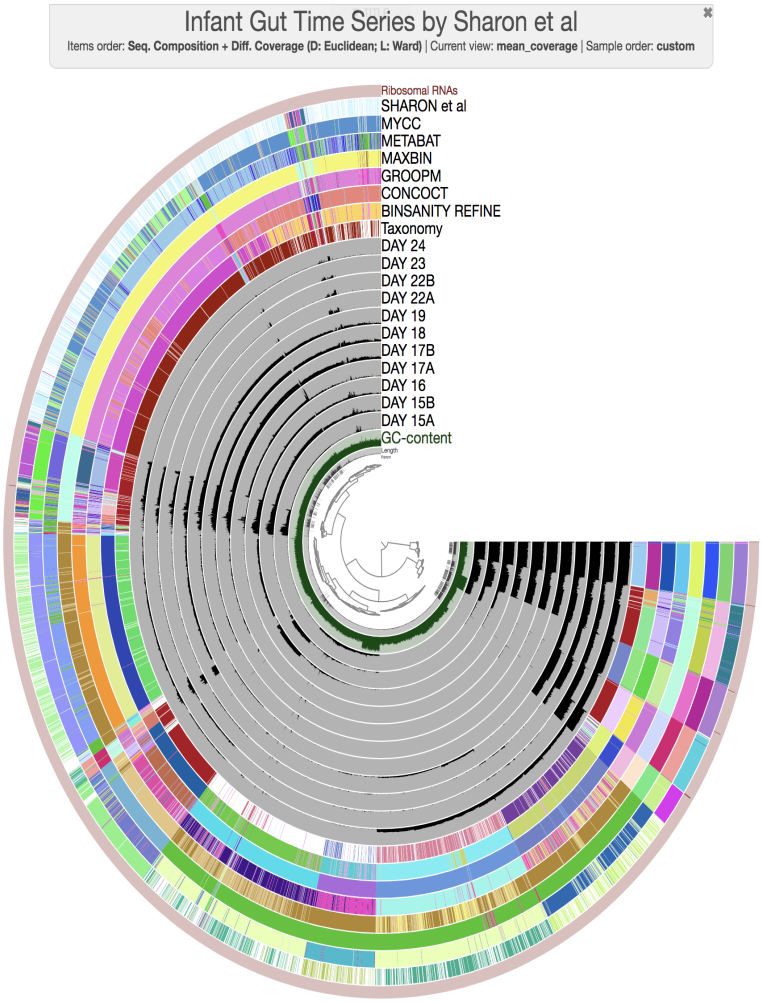
CONCOCT results come from the CONCOCT module embedded within anvi'o.

Step 5.

Comparing multiple binning approaches

Here we have the organization of contigs based on hierarchical clustering analysis, taxonomy from Centrifuge per contig (which is independent of the organization of contigs so it is a very good validation to see whether the organization makes sense), and results from the original publication from Sharon et al., in which authors did a very careful job to identify every genome in the dataset, including resolving the *Staphylococcus pangenome.* So these are the things we will assume "true enough" to build upon.

To compare binning results, we could import each external binning result into the profile database the way we imported CONCOCT. Unfortunately, Anvi'o Can only display one bin collection. But we have a workaround - merge all binning results into a single file (anvi-script-merge-collections) and use that file as an 'additional data file' to visualize them in the interactive interface.



cmd COMMAND

```
$ python anvi-script-merge-collections -c CONTIGS.db -
i ../03_EXTERNAL_BINNING_RESULTS/*.txt -o collections.tsv
$ head ../03_EXTERNAL_BINNING_RESULTS/collections.tsv | column -t
$ anvi-interactive -p PROFILE.db -c CONTIGS.db -A collections.tsv
```

When you look at collections.tsv, it has a very simple format.

Step 6.

This should give you a good start with using Anvi'o to visualize your bins. If you would like to continue with this tutorial online, please visit the Meren Lab tutorial on infant gut. The workshop organizers highly recommend reading through the full tutorial, even if you don't run the codes, to get "**Meren's two cents on binning**".