Week 7: Binning Genomes with Anvi'o Version 2

Rika Anderson

Abstract

Citation: Rika Anderson Week 7: Binning Genomes with Anvi'o. **protocols.io**
dx.doi.org/10.17504/protocols.io.g6tbzen
Published: 04 May 2017

Protocol

Intro

Step 1.

This week in lab we'll learn how to visualize your metagenomes and pull out individual genome bins. We aren't going to use a toy dataset this week—we're going straight into analysis with your metagenome datasets for your projects.

Genomes are disentangled from metagenomes by clustering reads together according to two properties: **coverage** and **tetranucleotide frequency**. Basically, if contigs have similar coverage patterns between datasets, they are clustered together; and if they have similar kmers appear over and over again, they will cluster together. When we cluster contigs together like this, we get a collection of contigs that are thought to represent a reconstruction of a genome from your metagenomic sample. We call these 'genome bins,' or 'metagenome-asembled genomes (MAGs).'

There is a lot of discussion in the field about which software packages are the best for making these genome bins. And of course, the one you choose will depend a lot on your dataset, what you're trying to accomplish, and personal preference. I chose anvi'o because it is a nice visualization tool that builds in many handy features.

I am drawing a lot of information for this tutorial from the anvi'o website. If you'd like to learn more, see the link below.

LINK:
http://merenlab.org/2016/06/22/anvio-tutorial-v2/

Preparing your contigs database for anvi'

Step 2.

Boot your computer as a Mac and use the Terminal to ssh in to Liverpool.

Preparing your contigs database for anvi'o

Step 3.

Make a new directory called “anvio” inside your project folder, then change into that directory.

Copy the following into your new directory:

1) assembled contigs with the cleaned up names (i.e. >c_000000000001)

2) all of your sorted .bam and .bai files that you already mapped to those contigs last week

cmd COMMAND

```
mkdir project_directory/anvio
cd project_directory/anvio
```

Preparing your contigs database for anvi'o

Step 4.

The first thing you have to do is make contigs database, which contains the sequences of your contigs, plus lots of information about those contigs-- we'll be combining together your Prodigal information and your Interproscan information into this contigs database, and we'll also be characterizing the taxonomy of your contigs. First, you make the database.

1. anvi-gen-contigs-database is the anvi’o script that makes the contigs database.
2. –f is the fasta file with your contigs that you have already assembled and fixed.
3. –o provides the name of your new contigs database.

cmd COMMAND

```
anvi-gen-contigs-database -f [project file assembled contigs] -o contigs.db
```

Preparing your contigs database for anvi'o

Step 5.

Now we will search our contigs for archaeal and bacterial single-copy core genes. This will be useful later on because when we try to disentangle genomes from this metagenome, these single-copy core genes can be good markers for how complete your disentangled genome is.

cmd COMMAND

```
anvi-run-hmms -c contigs.db
```

Preparing your contigs database for anvi'o

Step 6.

Now we are going to figure out the taxonomy of our contigs using a program called centrifuge. Centrifuge is a program that compares your contigs to a sequence database in order to assign taxonomy to different sequences within your metagenome. We're going to use it first to classify your contigs.

If you would like to know more, go here: http://merenlab.org/2016/06/22/anvio-tutorial-v2/ and here:

http://www.ccb.jhu.edu/software/centrifuge/

First, export your genes from anvi'o.

cmd COMMAND
```
anvi-get-dna-sequences-for-gene-calls -c contigs.db -o gene-calls.fa
```

Preparing your contigs database for anvi'o

Step 7.

Now run centrifuge.

cmd COMMAND
```
centrifuge -f -x /usr/local/CENTRIFUGE/p_compressed gene-calls.fa -S centrifuge_hits.tsv
```

Preparing your contigs database for anvi'o

Step 8.

Now import those centrifuge results for your contigs back in to anvi'o. It has a parser written into the software that can automatically read and import centrifuge output.

cmd COMMAND
```
anvi-import-taxonomy -c contigs.db -i centrifuge_report.tsv centrifuge_hits.tsv -
p centrifuge
```

Preparing your contigs database for anvi'o

Step 9.

Now anvi'o needs to combine all of this information—your mapping, your contigs, your open reading frames, your taxonomy—together. To do this, use the anvi-profile script. **Do this for every single sorted bam file you have (you should have one from each of the samples in your project region).**

1. anvi-profile is the name of the program that combines the info together
2. The –i flag provides the name of your processed bam file that you created in step 10 above.
3. The –c flag provides the name of your contigs database that you created in step 3 above.
4. The -M flag sets a minimum contig length. We're going to use 200 so you can use as many contigs as possible. In a project for publication, you'd want to use at least 1000, because the clustering of contigs is dependent on calculating their tetranucleotide frequencies (searching for patterns of kmers). You need to have a long enough contig to calculate these frequences accurately. But for our purposes, let's use 200.

cmd COMMAND
```
anvi-profile -i [your sorted bam file] -c contigs.db -M 200
```

ANNOTATIONS

Vianne Gao 14 Feb 2017

use sorted bam file mapped to your own contigs from week 6

Preparing your contigs database for anvi'o

Step 10.

Now merge all of these profiles together using a program called anvi-merge.

cmd COMMAND
```
anvi-merge */RUNINFO.cp -o SAMPLES-MERGED -c contigs.db
```

Preparing your contigs database for anvi'o

Step 11.

The next step requires visualization, and we're going to do that inside liverpool because that's where anvi'o is installed. Open up X2GO (see previous weeks' labs) and log in to liverpool. Open up a terminal window inside X2GO. Change directory into your anvi'o directory.

cmd COMMAND
```
cd project_directory/anvio
```
This is in your new Terminal window inside X2GO!

ANNOTATIONS
Vianne Gao 14 Feb 2017

To open a terminal in X2GO, right click and click on 'open terminal here'.

Preparing your contigs database for anvi'o

Step 12.

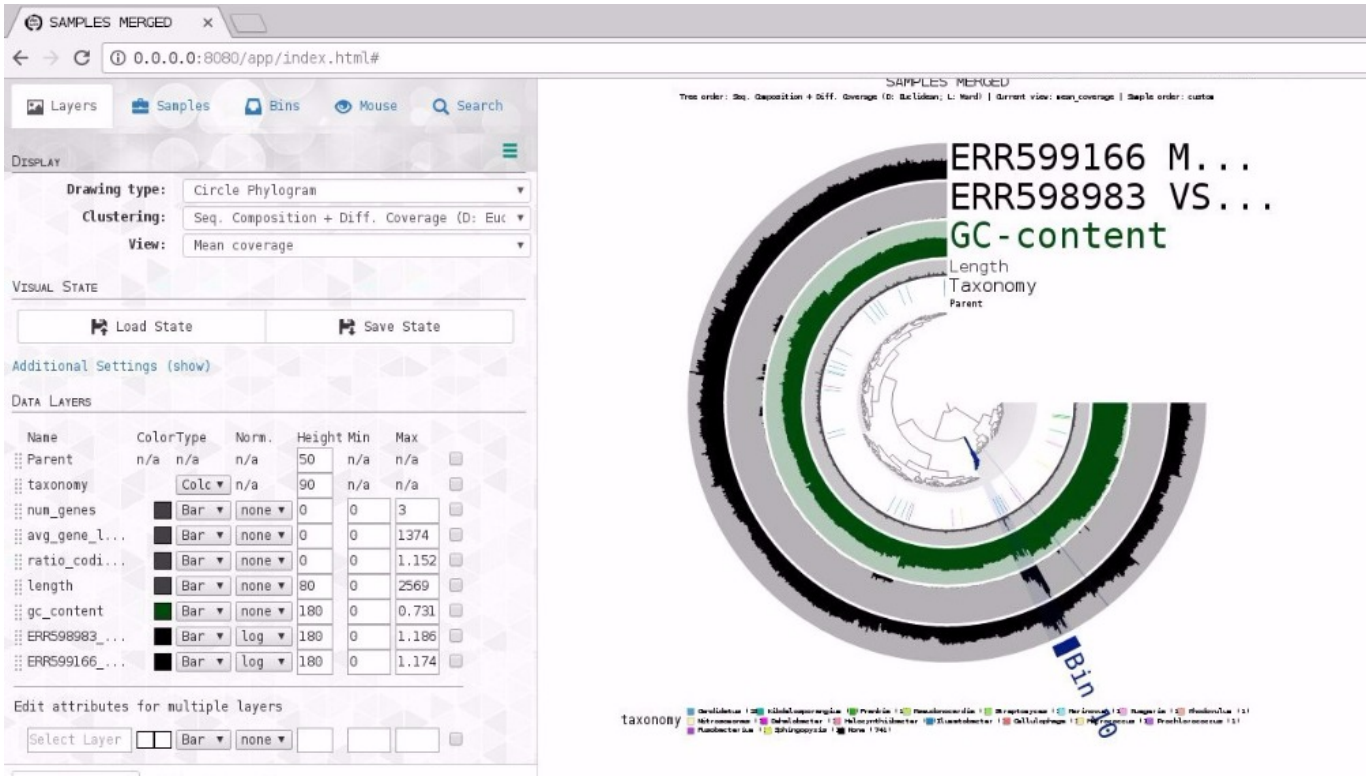
Now type this to open up the visualization of your contigs:

cmd COMMAND
```
anvi-interactive -p SAMPLES-MERGED/PROFILE.db -c contigs.db
```
This is also in your Terminal window inside X2GO!

Preparing your contigs database for anvi'o

Step 13.

What you should see is something like this:



What you are looking at:

-the tree on the inside shows the clustering of your contigs. Contigs that were more similar according to k-mer frequency and coverage clustered together.

-the rings on the outside show your samples. This is a visualization of the mapping that you did last week, but now we can see the mapping across the whole sample. There is one black line per contig. The taller the black line, the more mapping for that contig.

-the "taxonomy" ring shows the centrifuge designation for the taxonomy of that particular contig.

-the "GC content" ring shows the average percent of bases that were G or C as opposed to A or T for that contig.

We will go over the process for making bins together in class.

Because your datasets are fairly small, your bins are also going to be very small. Your percent completeness will be very low. Try to identify at least one bin according to patterns in the mapping of the datasets as well as the GC content.

When you are done making your bins, be sure to click on 'Save collection', give it a name ('my_bins' works), and then click on 'Generate a static summary page.' Click on the link it gives you. It will provide lots of information about your bins. In the boxes under the heading 'taxonomy,' you can click on the box to get a percentage rundown of how many contigs in your bin matched specific taxa according to centrifuge, if any matched.

Once you have completed your binning process, take a screenshot of your anvi'o visualization and save it as 'Figure 1.' Write a figure caption explaining what your project dataset is, and which datasets you mapped to your sample.

Preparing your contigs database for anvi'o

Step 14.

You will find your new bin FASTA files in the directory called '/project_directory/anvio/SAMPLES-MERGED/SUMMARY_my_bins'. 'general_bins_summary.txt' provides just that, with information about the taxonomy, total length, number of contigs, N50, GC content, percent complete, and percent redundancy of each of your bins. This is reflected in the summary html page you generated earlier when you clicked 'Generate a static summary page.' If you go to the directory 'bin_by_bin', you will find a series of directories, one for each bin you made. Inside each directory is a wealth of information about each bin. This includes (among other things):

- a FASTA file containing all of the contigs that comprise your bin (i.e. 'Bin_1-contigs.fa')
- mean coverage of each bin across all of your samples (i.e. 'Bin_1-mean-coverage.txt')
- files containing copies of single-copy, universal genes found in your contigs. These could be aligned and made into a tree to show all of your bins in a single tree. (i.e. 'Bin_1-Rinke_et_al_hmm-sequences.txt' and 'Bin_1-Campbell_et_all-hmm-sequences.txt.'
- information about single nucleotide variability in your bins-- the number of SNVs per kilobasepair. (i.e. 'Bin_1-variability.txt')

Preparing your contigs database for anvi'o

Step 15.

Now that you have a nice summary of the bins you've made, you're going to profile the single nucelotide variants in one bin.

We've been interchangably calling these "SNPs" and "SNVs," so let's clarify. A single nucleotide polymorphism (SNP) is a single nucleotide change in an isolate genome. For example, Carl Zimmer was identfying SNPs in his genome relative to a human reference genome. Here, we're looking at variation within a microbial community, so the variation at a single position in the genome (or "site") may have, for example, some As and some Ts. This means that some members of the microbial community had As there and others had Ts. This is a little different from SNPs, so we're calling them single nucleotide variants (SNVs).

To profile the SNVs in your bins, you invoke the following command for a specific bin you're interested

in. For example, if you're interested in Bin_1, type this:

cmd COMMAND

```
anvi-gen-variability-profile -c contigs.db -p SAMPLES-MERGED/PROFILE.db -C my_bins -
b Bin_1 -o variability_Bin1.txt
```

You can type this either in the terminal in X2Go or in your local terminal that is connected to liverpool remotely through ssh. This is assuming you want to visualize the variability across samples for Bin 1. If you want to focus on a different bin, specify which one.

Preparing your contigs database for anvi'o

Step 16.

The command spits out a giant text file with many columns. This is a very useful, large set of data that shows information for every single SNV in your bin. If you wished to pursue this farther for your projects, you could import this data file into Excel (or manipulate it with Python if you're into that) and parse to your heart's content. You could explore the data, make graphs, look for patterns, compare SNVs between bins or between samples. Here is what the columns mean:

1. **entry_id** refers to the unique id for the line in the output file. It is the only column that contains a unique id for each line.

2. **unique_pos_identifier** refers to the unique identifier for a given nucleotide position. Since each sample in the profile database can report variability for every nucleotide position, a **unique_pos_identifier** can appear in the file as many times as the number of the samples in the analysis. This column can be used to pull frequencies of nucleotides for a given nucleotide position from all samples.

3. **sample_id** corresponds to the sample name a given particular line is reported from. This column allows the linking of SNVs and the sample(s) they were identified from.

4. **pos** refers to the nucleotide position in the split.

5. **pos_in_contig** refers to the nucleotide position in the contig (why is this called *pos_in_contig*, and the one before is not called *pos_in_split*? Well, we have been wondering about that for a long time, too).

6. **corresponding_gene_call** refers to a unique gene caller id (`-1`, if the position falls out of a gene call).

7. **in_partial_gene_call** indicates whether the gene call is incomplete (i.e., starts with a start codon, stops with a stop codon, etc). `1` if incomplete, `0` if both start and stop positions are detected, or if the position is not in a gene.

8. **in_complete_gene_call** indicates the gene completion status. `1` for complete, `0` if incomplete, or if the position is not in a gene.

9. **base_pos_in_codon** refers to the position of the nucleotide in a codon. `1`, `2` or `3` for codon positions, `-1` if the position is not in a detected gene.

10. **codon_order_in_gene** refers to the order of the codon in the gene call, starting from the start position. `-1` if the position is not in a called gene.

11. **coverage** refers to the coverage, the number of recruited reads mapping to this position.

12. **cov_outlier_in_split** indicates whether the coverage of this position is marked as an outlier compared to all other positions in the split. `1` if outlier, and `0` if not.

13. **cov_outlier_in_contig** has the same purpose with the one above, except it is at the contig-level.

14. **departure_from_reference** refers to the ratio of nucleotides in a given position that diverge from the reference nucleotide. If a position with a coverage of `100X` has the nucleotide `A` in the reference, the departure from reference would be `0.2` if the frequency of mapping nucleotides are as follows: `A: 80X`, `T: 0X`, `C: 12X`, and `G: 8X`. Note that the departure from reference can dramatically change across samples, revealing subtle differences at the single nucleotide level.

15. **competing_nts** refers to the two most represented nucleotides. Note that if competing nucleotides are stored as `CT`, it does not necessarily mean that `C` occurs more than `T`, since they are ordered alphabetically. For positions positions that does not have any variation, competing nucleotides will contain two identical nucleotides. I hope here you are asking yourself here "*why would a position without any variation would appear in this table?*". The answer is `--quince-mode`. See below.

16. **reference** refers to the reference nucleotide in the mapping context.

17. **A** refers to the number of mapped reads covering this position with a `A`.

18. **T** refers to the number of mapped reads covering this position with a `T`.

19. **C** refers to the number of mapped reads covering this position with a `C`.

20. **G** refers to the number of mapped reads covering this position with a `G`.

21. **N** refers to the number of mapped reads covering this position with an ambiguous base.

22. **consensus** refers to the most frequent nucleotide mapping to this position. This column is used to define the **departure_from_consensus** ratio.

23. **departure_from_consensus** refers to the ratio of nucleotides in a given position that diverge from the most frequent nucleotide. The departure from reference would be `0.5` if the frequency of mapping nucleotides for this position is `A: 10X`, `T: 20X`, `C: 30X`, and `G: 70X`. Compared to the **departure from reference**, this column relies less on the reference sequence and thus might better reflect variation in environmental samples.

24. **n2n1ratio** refers to the ratio of the second most frequent nucleotide to the consensus nucleotide. This value would be `0.42` if the frequency of mapping nucleotides for this position is `A: 10X`, `T: 20X`, `C: 30X`, and `G: 70X`.

25. **contig_name** refers to the contig name as it appears in the contigs database.

26. **split_name** refers to the split name.

See the link below for more information about this.

LINK:
http://merenlab.org/2015/07/20/analyzing-variability/

Preparing your contigs database for anvi'o

Step 17.

Our last step is to visualize the single nucleotide variability of that particular bin. Type this:

cmd COMMAND
```
anvi-script-snvs-to-interactive variability_Bin1.txt -o SNVs_Bin1
```
This is assuming you want to visualize the variability across samples for Bin 1. If you want to focus on a different bin, specify which one. You could type this in your regular Terminal window (connected to liverpool via ssh) or in the X2Go Terminal. It doesn't matter which one.

Preparing your contigs database for anvi'o

Step 18.

Now type this in the X2Go terminal:

cmd COMMAND
```
anvi-interactive -d SNVs_Bin1/view_data.txt -s SNVs_Bin1/samples.db -t SNVs_Bin1/tree.txt -
p SNVs_Bin1/profile.db -A SNVs_Bin1/additional_view_data.txt --
title "SNV profile for Bin1" --manual
```
This is assuming you want to visualize the variability across samples for Bin 1. If you want to focus on a different bin, specify which one.

Preparing your contigs database for anvi'o

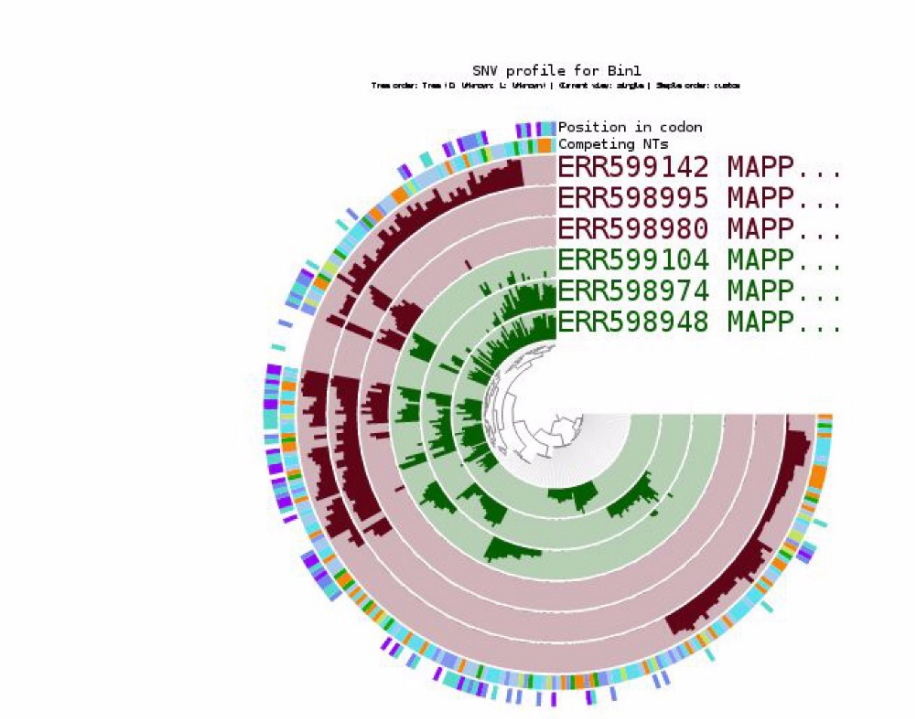
Step 19.

You will see something that looks like this. Each leaf is a nucleotide position, and bars indicate the divergence from the reference-- the taller the bar, the more the SNVs diverge from the reference. The shorter the bar, the more similar they are to the reference. If the bar is somewhere in the middle, it means that about half of the SNVs are one base pair and the other half are another: for example, at that position, half of the microbes had a T there and half had a G.

You can probably see that some samples had many more SNVs than others. You can probably also see that some SNVs were unique to some samples, and others were variable in all of the samples.

This is simply a way to help you visualize the total density and distribution of single nucleotide variants in your bin.

Take a screenshot of this visualization and save it as 'Figure 2.' Write a figure caption, and be sure to indicate which bin you selected.



Analyzing your anvi'o results

Step 20.

You now have a wealth of information at your fingertips about your bins. This week you will simply characterize a single bin.

For the document you will submit to Moodle this week:

Submit the screenshots of your anvi'o visualizations (Figures 1 and 2).

Select a single bin (preferably the one you show in Figure 2) and characterize the following:

1) How many contigs were in your bin?

2) What was the taxonomy of the contigs in your bin, according to anvi'o? (See the html page generated by 'Generate a static summary page' and click on the box for your bin under the column 'Taxonomy.')

3) What was the N50 of your bin?

4) What was the average coverage of your bin with mapped reads from your own sample? How about the average coverage of mapped reads from each of the other samples you mapped? What does this tell you about the abundance of this particular microbial lineage in each of your samples?

5) In which sample did this particular genome bin have the highest density of single nucleotide variants per 1000 base pairs (SNVs/kbp), and in which sample did this particular genome bin have the lowest SNVs/kbp? (hint: see SUMMARY_my_bins/bin_by_bin/Bin_1/Bin_1_variability.txt if you want to find this information for Bin_1.) Explain which sample is which-- for example, rather than simply saying that sample ERR598974 had the most SNVs/kbp, state that sample ERR598974 is the Arabian Sea deep chlorophyll maximum.

6) Imagine that in sample A, the microbial lineage represented by your bin was under very strong natural selection. That is, a very specific strain was favored in that environment. In sample B, there was less selection pressure on that particular strain, and as a result mutations were not eradicated from the population as quickly.

a) In which sample would you expect the genome bin to have the highest SNV variability? Why?

b) Given your answer in a), and based on the data you showed in question #5, in which of the sample sites do you expect your own bin to be under the strongest selection

(or strongest evolutionary pressure)? Explain why.

Submit this document on the Moodle by lab time next week.