Building Anvi'o profiles + Interactive Viewer

James Thornton Jr,Bonnie Hurwitz

Abstract

Protocol details how to create and merge Anvi'o profiles to view them using anvi-interactive.

Citation: James Thornton Jr,Bonnie Hurwitz Building Anvi'o profiles + Interactive Viewer. **protocols.io** dx.doi.org/10.17504/protocols.io.kwpcxdn
Published: 25 Nov 2017

Protocol

Login to the HPC

Step 1.

Login to the HPC

cmd COMMAND
```
ssh hpc
ocelote
```

Move to your user directory

Step 2.

Move to your user directory on bh_class

cmd COMMAND
```
cd /rsgrps/bh_class/username
```

Go to your BAM file directory

Step 3.

If you are here, you must be done with your contigs database, and have your BAM files ready. Good! It is time to initialize your BAM files, and create an *anvi'o profile* for your samples.

Check and make sure you have "bam" files here. If not you need to run using a previous protocol.

cmd COMMAND
```
cd /rsgrps/bh_class/username/read_recruit/bam
ls *.bam
```

Go to the Anvi'o terminal on your computer

Step 4.

Now we need to move all of the bam files from the HPC to our personal computers to input into Anvi'o. Go to the Anvi'o terminal on your computer, where you created the contigs database in the previous

protocol.

NOTES
James Thornton Jr 21 Nov 2017

Remember, we use Anvi'o locally. (Docker for Windows users)

Download the bam files
Step 5.

Download the bam files form the HPC to your computer

cmd COMMAND
```
scp username@sftp.hpc.arizona.edu:/rsgrps/bh_class/username/read_recruit/bam/*bam .
```
Notice the "." at the end of the line, that means download to the current directory.

Download the bam files
Step 6.

ADDITIONALLY

Download your partners bam files.

cmd COMMAND
```
$ scp username@sftp.hpc.arizona.edu:/rsgrps/bh_class/partners/read_recruit/*bam .
```

Create a list of samples for downstream processing
Step 7.

To help with downstream processing. Create a list of the samples you have. We will call this file:

SAMPLE_IDs

cmd COMMAND
```
ls *bam | sed 's/\.bam//' > SAMPLE_IDs
```

Index the bam files
Step 8.

Anvi'o requires BAM files to be sorted and indexed. In most cases the BAM files you get back from your mapping software will not be sorted and indexed. You need to initialize your BAM files:

cmd COMMAND
```
for sample in `cat SAMPLE_IDs`; do anvi-init-bam $sample.bam -o $sample.i.bam; done
```

NOTES
Amy Hudson 21 Nov 2016

Is (`) specific to the bash command line? I don't think I've seen it yet in perl6.

Bonnie Hurwitz 21 Nov 2016

backticks are in perl5 and essentially mean execute this "command" to get some sort of input. In perl6, there is the concept of processes that are "run" (that is the actual command). You can see examples in all of Ken's test suite programs.

Emma Skidmore 22 Nov 2016

When I run this step it gives me an error of too few arguments.

usage: anvi-init-bam [-h] [-o FILE_PATH] BAM_FILE

James Thornton Jr 29 Nov 2016

PC users

When you scp your files using Cygwin, move those files to a new folder in Documents. Then in docker quickstart terminal navigate to that folder and do pwd to get the full path. Then to launch Anvio:

docker run --rm -v /path/to/files:/my_data -p 8080:8080 -it meren/anvio:latest

Additional troubleshooting- if having issues do docker ps and see if there are existing sessions. If so do docker kill [session id]

James Thornton Jr 21 Nov 2017

Reminder for PC users using the Docker image:

docker run --rm -v /path/to/files:/my_data -p 8080:8080 -it meren/anvio:latest

Creating an anvi'o profile database

Step 9.

In contrast to the contigs database, an anvi'o profile database stores sample-specific information about contigs. Profiling a BAM file with anvi'o using `anvi-profile` creates a single profile that reports properties for each contig in a single sample based on mapping results. Each profile database links to a contigs database, and anvi'o can merge single profiles that link to the same contigs database into merged profiles (which will be covered later).

In other words, the profiling step makes sense of each BAM file separately by utilizing the information stored in the contigs database. It is one of the most critical (and also most complex and

computationally demanding) steps of the metagenomic workflow.

The simplest form of the command that starts the profiling looks like this:

```
anvi-profile -i SAMPLE-01.bam -c contigs.db
```

When you run `anvi-profile` it will:

- Process each contig that is longer than `2,500 nts` by default. You can change this value by using `--min-contig-length` flag. But you should remember that the minimum contig length should be long enough for tetra-nucleotide frequencies to have enough meaningful signal. There is no way to define a golden number for minimum length that would be applicable to genomes found in all environments. We empirically chose the default to be 2,500, and have been happy with it. You are welcome to experiment, but we advise you to never go below 1,000. You also should remember that the lower you go, the more time it will take to analyze all contigs. You can use –list-contigs parameter to have an idea how many contigs would be discarded for a given `--min-contig-length` parameter. If you have an arbitrary list of contigs you want to profile, you can use the flag `--contigs-of-interest` to ignore the rest.

- Make up some output directory, and sample names for you. We encourage you to use `--output-dir` parameter to tell anvi'o where to store your output files, and `--sample-name` parameter to give a meaningful, preferably not-so-long sample name to be stored in the profile database. This name will appear almost everywhere, and changing it later will be a pain.

Processing of contigs will include:

- The recovery of mean coverage, standard deviation of coverage, and the average coverage for the inner quartiles (Q1 and Q3) for a given contig. Profiling will also create an HD5 file where the coverage value for *each nucleotide position* will be kept for each contig for later use. While the profiling recovers all the coverage information, it can discard some contigs with very low coverage declared by `--min-mean-coverage` parameter (the default is 0, so everything is kept).

- The characterization of single-nucleotide variants (SNVs) for every nucleotide position, unless you use `--skip-SNV-profiling` flag to skip it altogether (you will definitely gain a lot of time if you do that, but then, you know, maybe you shouldn't). By default, the profiler will not pay attention to any nucleotide position with less than `10`X coverage. You can change this behavior via `--min-coverage-for-variability` flag. Anvi'o uses a conservative heuristic to not report every position with variation: i.e., if you have 200X coverage in a position, and only one of the bases disagree with the reference or consensus nucleotide, it is very likely that this is due to a mapping or sequencing error, and anvi'o tries to avoid those positions. If you want anvi'o to report everything, you can use `--report-variability-full` flag. We encourage you to experiment with it, maybe with a small set of contigs, but in general you should refrain reporting everything (it will make your databases grow larger and larger, and everything will take longer for -99% of the time- no good reason).

- Finally, because single profiles are rarely used for genome binning or visualization, and since clustering step increases the profiling runtime for no good reason, the default behavior of profiling is to *not cluster* contigs automatically. However, if you are planning to work with single profiles, and if you would like to visualize them using the interactive interface without any merging, you can use `--cluster-contigs` flag to initiate clustering of contigs. In this case anvi'o would use default clustering configurations for single profiles, and store resulting trees in the profile database. You *do not* need to use this flag if you are planning to merge multiple profiles (i.e., if you have more than one BAM files to work with, which will be the case for most people).

cmd COMMAND

```
anvi-profile -i SRR1647045.i.bam -c contigs.db --output-dir Ocean1 --sample-name Ocean1
```

Use appropriate output names (Ocean, Acid, Canine, Human, Gut, Oral, etc...)

NOTES

James Thornton Jr 21 Nov 2017

IMPORTANT: You must use the bam file that has been indexed (SRR34443.i.bam)

James Thornton Jr 21 Nov 2017

Repeat this step for both yours and your partners bam files.

ANNOTATIONS

Eldridge Wisely 01 Dec 2017

I keep getting this error message:

Config Error: At least one contig name in your BAM file does not match contig names stored in

the contigs database. For instance, this is one contig name found in your BAM

file: '42114', and this is another one found in your contigs database: '7293'.

You may be using an contigs database for profiling that has nothing to do with

the BAM file you are trying to profile, or you may have failed to fix your

contig names in your FASTA file prior to mapping, which is described here:

http://goo.gl/Q9ChpS

How should I fix my contig names? It seems like my partner's naming system is different from mine, which is again different from the contigs...

Merge all anvi'o profiles

Step 10.

You have all your BAM files profiled! Did it take forever? Well, sorry about that. But now you are golden.

The next step in the workflow is to to merge all anvi'o profiles.

When you run `anvi-merge`,

- It will merge everything and create a merged profile (yes, thanks, captain obvious),
- It will attempt to create multiple clusterings of your splits using the default *clustering configurations*. Please take a quick look at the default clustering configurations for merged profiles –they are pretty easy to understand. By default, anvi'o will use euclidean distance and ward linkage algorithm to organize contigs, however, you can change those default values with `--distance`and `--linkage` parameters (available options for distance metrics and linkage algorithms are listed in this release note). Hierarchical clustering results are necessary for comprehensive visualization, and human guided binning, therefore, by default, anvi'o attempts to cluster your contigs using default configurations. You can skip this step by using `--skip-hierarchical-clustering` flag. But even if you don't skip it, anvi'o will skip it for you if you have more than 20,000 splits, since the computational complexity of this process will get less and less feasible with increasing number of splits. That's OK, though. There are many ways to recover from this. On the other hand, if you want to teach everyone who is the boss, you can force anvi'o try to cluster your splits regardless of how many of them are there by using `--enforce-hierarchical-clustering` flag. You have the power.
- It will attempt to run CONCOCT to bin your splits automatically. CONCOCT can deal with hundreds of thousands of splits. Which means, regardless of the number of splits you have, and even if you skip the hierarchical clustering step, there will be a collection in the merged profile database (which will be called CONCOCT) with genome bins identified by CONCOCT in an automatic manner. From which you can generate a summary, or run the interactive interface with `--collection-name CONCOCT` parameter (more later on these). But if you would like to skip default CONCOCT clustering, you can use `--skip-concoct-binning` flag.

cmd COMMAND

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

Launch Interactive Viewer

Step 11.

Launch the Anvi'o interactive viewer.

cmd COMMAND

```
$ anvi-interactive -p SAMPLES-MERGED/PROFILE.db -c contigs.db
```

Launch Interactive Viewer

Step 12.

For a detailed overview of the anvi'o interactive viewer and all of the things you can do with it. See the weblink below. We will go over major components of the viewer in class.

LINK: http://merenlab.org/2016/02/27/the-anvio-interactive-interface/

Launch Interactive Viewer

Step 13.

Click on the draw button in the layers tab in anvi'o to draw a circle phylogram. Try zooming in and out of the viewer to see more or less detail. Try switching the colors for each of the body sites, and redraw the figure.

Through the layers tab you can,

- **Change general settings for the tree** (i.e., switching between circle or rectengular displays, changing tree radius or width), **and layers** (i.e., editing layer margins, or activating custom layer margins).
- **Load or save states** to store all visual settings, or load a previously saved state.
- **Customize individual layers** by switching between different **display modes**depending on the layer type (i.e., 'text' or 'color' mode for categorical layers, or 'bar' or 'intensity' mode for numerical layers), **set normalization** (i.e., 'square-root', or 'log' normalization), **minimum, and maximum** cutoff values for numerical layers, or set **layer height**, and **layer margin** (i.e., its distance from the previous layer).
- Use the **multi-selector** at the bottom to change settings for multiple layers at once.

Launch Interactive Viewer

Step 14.

In the bins tab, click on the 'Load bin collection' button. Select 'CONCOCT' from the pop up window.

Questions:

How many bins did the CONCOCT algorithm find? These are meant to represent different 'genomes' in your metagenomic sample. Do you see bins that are comprised of reads from only certain skin sites?

Anvi'o allows you to create selections of items shown in the display (whether they are metagenomic contigs, 16S rRNA tags, or any other type of information). Bins tab allow you to maintain these selections. Any selection on the tree will be added to active bin in this tab (the state radio button next to a bin defines its activity). Through this tab you can,

- **Create or delete bins**, **set bin names**, **change the color of a given bin**, or sort bins based

on their name, the number of units they carry, or completion and contamination estimates (completion / contamination estimates are only computed for genomic or metagenomic analyses).
- View **the number of selected units** in a given bin, and see the **list of names in the selection** by clicking the button that shows the number of units described in the bin.
- **Store a collection of bins**, or **load a previously stored collection**.

Can you create better bins?

Launch Interactive Viewer

Step 15.

In the mouse tab, mouse over different regions of the graph. You can mouse over different regions of the graph to see data about individual contigs. You can also mouse over different levels to see more or less of the data in each of the circles, going from center (with the most info) to the ends (with the least).

Launch Interactive Viewer

Step 16.

You can also right click on a contig and select "inspect" to see the detailed informaton about that contig, including genes and annotations (kegg and pfam) you loaded in the contigs db. Or, you can do a blast on the fly.

Launch Interactive Viewer

Step 17.

Samples tab is for the additional data you provide the interface through a samples database (see samples order and samples information sections above). Through this layer you can,

- **Change the order of layers** using automatically-generated or user-provided orders of layers using the Sample order combo box,
- **Customize individual samples information entries**.

Changes in this tab can be reflected to the current display without re-drawing the entire tree unless the sample order is changed.

Launch Interactive Viewer

Step 18.

The interactive viewer is fun to play with, but you might also want to create a summary of the data. You can do this from the anvi'o command line:

cmd COMMAND

```
anvi-summarize -p SAMPLES-MERGED/PROFILE.db -c contigs.db -o SAMPLES-SUMMARY -C CONCOCT
```

Launch Interactive Viewer

Step 19.

Was Anvi'o able to determine the taxonomy for each of the bins? Click on one of the buttons that says "NONE". What taxa are present in that bin? Do you see a mixture?

Launch Interactive Viewer

Step 20.

How complete are the bins from CONCOCT? Look at the "Compl." tab to get an estimate fomr Anvi'o on how complete your bins are based on the composition of core bacterial genes.

Launch Interactive Viewer

Step 21.

When we try to estimate the completeness of a genome bin, we identify single-copy genes that appear more than once as "contamination".

Jed Fuhrman suggested that the use of "redundancy" would be more appropriate, since these are not always contaminations (sometimes there are hits due to not-very-specific HMM profiles, etc), *and* the word "contamination" is a bit scary.

Compare with partners data

Step 22.

Do you have similar results across the two time points?

Compare with partners data

Step 23.

Can you use some of anvi'os interactive features to create better and more complete bins?