MG_HW11: Building Anvi'o profiles

Bonnie Hurwitz and Murat Meren

Abstract

Citation: Bonnie Hurwitz and Murat Meren MG_HW11: Building Anvi'o profiles. **protocols.io**
dx.doi.org/10.17504/protocols.io.gepbtdn
Published: 15 Nov 2016

Protocol

Login to the HPC

Step 1.

login to the HPC

cmd COMMAND
```
ssh hpc
ice
```

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.

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

Download the bam files

Step 5.

Download the bam files form the HPC to your computer

cmd COMMAND

```
scp 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.

Create a list of samples for downstream processing

Step 6.

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
```

My SAMPLE_IDs contains: SRR1647045 SRR1647046 SRR1647047 SRR1647048 SRR1647049 SRR1647141 SRR1647142 SRR1647143

Index the bam files

Step 7.

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
```

ANNOTATIONS

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

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

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]

Creating an anvi'o profile database

Step 8.

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:

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

- 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 10X 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 Pa --sample-name Pa
```
Look up your sample name in the original sample to metadata table from your report. We will need to run this step on each of the samples individually, and use the appropriate names. Run this step 8x, once for each sample.

Merge all anvi'o profiles

Step 9.

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.

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

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
```


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