Sequence Download, Preparation, Assembly, and Analysis in anvi'o

Erica Holdridge

Adapted from: Assembly-Based Metagenomics Tutorial and Metagenomic Workflow

Updated: October 8, 2021

1. Download raw data from MG-RAST. There are a few different files available for download for each metagenome (read more about them here https://www.biostars.org/p/94875/ and here https://help.mg-rast.org/user_manual.html). I used the 299 file, which has been preprocessed to remove low quality sequences, dereplicated, and checked for host contamination.

   ```
   mkdir 01_QC/
   cd 01_QC/

   curl https://api-ui.mg-rast.org/download/${META_NUM}?file=299.1 >
   rawdata/${META_NUM}.fna
   ```

2. Create a co-assembly for all of your samples. Depending on how many samples you have and the depth of sequencing, this step may take several hours. Make sure you have Megahit installed.

   ```
   cd ~
   mkdir 02_ASSEMBLY/
   ```

   These two lines create environmental variables that correspond to all of your forward and reverse paried-end reads.

   ```
   R1s=`ls 01_QC/*_R1.fna | python -c 'import sys;
   print(",".join([x.strip() for x in sys.stdin.readlines()]))'`
   R2s=`ls 01_QC/*_R2.fna | python -c 'import sys;
   print(",".join([x.strip() for x in sys.stdin.readlines()]))'`

   megahit -1 $R1s -2 $R2s --min-contig-len 1000 -m 0.85 -o 02_ASSEMBLY/ -
   t 48
   ```

3. Now, you need to reformate the names of contigs into something anvi'o likes and remove sequences that are < 1000bp.

   ```
   mkdir 03_CONTIGS/

   anvi-script-reformat-fasta 02_ASSEMBLY/contigs.fa -o
   03_CONTIGS/contigs.fna -l 1000 --simplify-names
   ```

4. This step generates the BAM files you need to build anvi'o profiles **for each sample**. Anvi'o requires that your BAM files are sorted and indexed, so we will go ahead and do that during this step as well. Note: There are many other ways to do this, but this is one way that uses common tools. The tools you will need installed are Megahit (see Step 2),

Bowtie2, and Samtools. The first two lines of code below can help you check if they are already installed.

```
megahit -v
bowtie2 --version | head -n 1 | awk '{print $3}'
samtools -version

bowtie2-build scratch/rawdata/03_CONTIGS/contigs.fa
scratch/rawdata/04_MAPPING/contigs
```

Note: You must run the following **for each sample**.

```
bowtie2 -f --threads 40 -x 04_MAPPING/contigs -1
01_QC/${META_NUM}_R1.fna -2 01_QC/${META_NUM}_R2.fna -S
04_MAPPING/${META_NUM}.sam
samtools view -F 4 -bS 04_MAPPING/${META_NUM}.sam >
04_MAPPING/${META_NUM}-RAW.bam
anvi-init-bam 04_MAPPING/${META_NUM}-RAW.bam -o
04_MAPPING/${META_NUM}.bam
rm 04_MAPPING/${META_NUM}.sam 04_MAPPING/${META_NUM}-RAW.bam
```

5. It's *optional* but a good idea to annotate contigs with hits from NCBI's Cluster of Orthologous Groups (COGs). This line of code handles COGs setup (if you haven't done that already; if you have, skip this step).

```
anvi-setup-ncbi-cogs
```

6. Now you can create your anvi'o database using the contigs.fa file you just generated. The last line adds hits from hidden Markov models (HMMs) to your database, which is *optional* but a good idea.

```
mkdir anvio_output/

anvi-gen-contigs-database -f 03_CONTIGS/contigs.fa -o
anvio_output/contigs.db -n ProjectName
anvi-run-ncbi-cogs -c anvio_outputcontigs.db --num-threads 20
anvi-run-hmms -c anvio_output/contigs.db
```

7. You can take a look at what that database looks like with the following code. Note: This will open an interactive window in your web browser, so be sure you are set up to do that, especially if you are working on an HPC. See my other tutorial on how to do this, **Using anvi'o on an HPC with Slurm**.

```
anvi-display-contigs-stats anvio_output/contigs.db
```

8. Using the BAM files and anvi'o database you generated earlier, create an anvi'o profile **for each sample**.

```
mkdir anvio_output/profiles/

anvi-profile -i 04_MAPPING/${META_NUM}.bam -c anvio_output/contigs.db --output-dir anvio_output/profiles/${META_NUM} --sample-name ${META_NUM}_sample
```

9. The next step is to merge all of your anvi'o profiles.

```
anvi-merge anvio_output/profiles/*/PROFILE.db -o SAMPLES-MERGED -c contigs.db
```

10. Now you are all set up to use anvi'o interactive interface for further analyses. Check out their documentation for more information on everything anvi'o can do.