**Metagenomic Protocol (From Quality Control to Mapping) for Metagenome-assembled Genomes (MAGs) using Anvi’o**

Anvi’o (advanced analysis and visualization platform for ‘omics data), an open-source software, provides a visualization platform for advanced analysis of ‘omics data. Anvi’o is able to analyze a range of multi-’omics data including metagenomics, genomics, pangenomics, phylogenomics, metapangenomics, metatranscriptomics, and microbial population genetics. The modular architecture of Anvi’o is different from any other existing software. This ensures interactivity, flexibility, extensibility, and most importantly, reproducibility. Using Anvi’o, the users can very easily avoid rigid workflows, and deal with ‘omics data. Snakemake is a tool for building computational workflows. The strategy here is to combine the Anvi’o with snakemake workflow that yielded better documentation and reproducibility. Combing the snakemake with Anvi’o allows the users to easily establish the workflows in any type of computer system. Further, the users can also set up parallelization of several independent analysis steps. The Anvi’o circumvents the requirements of complex computational demands and resources by directly converting raw input to data products and can easily be analyzed through Anvi’o interactive interface. This lesson will enable the users to perform every step demonstrated, reproduce the results, and improvise these steps for other projects. To guide the readers, we have introduced the purposes of the dir system:

1. input: the source of the input data is mentioned, along with explanation and guidelines of the input data are provided.
2. lib: the source code used within the workflow.
3. output: the output results of the workflow.
4. workflow: step by step of the pipeline.

Anvi’o Installation:

1. Conda setup: If the conda is not installed in the system, it is necessary to open a terminal such as iTerm.

*Command:*

*conda install*

To verify whether you already have conda installed, copy and paste the following command into your terminal.

*Command:*

*conda --version*

Always make sure that you work in an up-to-date conda environment by using the following command:

*Command:*

*conda update conda*

1. Anvi’o environment setup

You should create a new conda environment using the command:

*Command:*

*conda create -y --name anvio-7.1 python=3.6*

Then activate it using the command:

*Command:*

*conda activate anvio-7.1*

1. Installing Anvi’o

The first step is to download the python source package for the anvi’o release using the following command:

*Command:*

*curl -L https://github.com/merenlab/anvio/releases/download/v7.1/anvio-7.1.tar.gz \*

*--output anvio-7.1.tar.gz*

Then use the following command to install anvi’o:

*Command:*

*pip install anvio-7.1.tar.gz*

Input Data:

Source: <https://figshare.com/s/35ea294e2671d75f1d5c>

Anvi’o use the program *anvi-run-workflow* to run the workflows. For a particular workflow, the program will the users to prepare the config file.

*Command*

*anvi-run-workflow --list-workflows*

*Available workflows ....: contigs, metagenomics, pangenomics, phylogenomics, trnaseq*

After you have decided the process, you wish to use, the config file allows you to change the parameters and order of steps associated with that process. Even if you are satisfied with all the default parameters, the config file is required for all workflows. The users should ensure that the config files are proper and preparing a config file for a particular workflow could be challenging.

*Command*

*anvi-run-workflow -w WORKFLOW-NAME \*

*--get-default-config OUTPUT-FILE-NAME*

The --get-default-config will generate a default config file for a workflow that you can modify. Configurable flags and parameters will be contained in this file. You can either leave any parameters that you do not intend to change ‘as is,' or remove those that you do not want to change from your config file to make it shorter and cleaner.

There are three configurations in the config file:

1. **General workflow parameters:** You will need a name for your project and the workflow mode you want to employ.
2. **Parameters:** Parameters that are exclusively applicable to a single rule, such as the anvi'o profiling step's minimum contig length. Each program has unique parameters. The users should ensure to use parameters that appear in the config file that would be identical to the names used in the particular program. For instance, if there are multiple ways to use adjustable parameters or arguments – the users should use the longer one. Let’s take an example – *anvi-run-hmms* are able to accept with *-H* or *--hmm-profile-dir* parameters that specify the directory path of HMM profile. However, the users are only allowed to use *--hmm-profile-dir* in the config file.
3. **Names of the output directory:** This is about how anvi’o will deal with output directories and files.

Workflow:

Raw paired-end sequencing reads for shotgun metagenomes are the **default entry point** into the metagenomics workflow. The workflow's **default endpoint** is a merged profile database ready for bin refinement, as well as an annotated anvi'o contigs database. The steps in the workflow are as follows:

1. Using illumina-utils, quality-check metagenomic short reads and generate a thorough report for the outcomes of this step. Quality-check to be performed by removing the low quality reads. The combination of B-tail trimming and passed chastity filter to be used. Then reads to be removed that contained uncalled bases.
2. Selected programs for generating taxonomic profiles of short reads. These profiles are imported to individual databases of profiles that are available in the merged profile database of anvi’o.
3. Using megahit and/or idba\_ud and/or metaspades to assemble quality filtered metagenomic reads.
4. Using anvi-gen-contigs-database, create an anvi'o contigs database using assembled contigs. The contigs database should have annotations of the functions, taxonomy, and HMMs.
5. Using bowtie2, map short reads from metagenomes to contigs and then generate indexed and sorted BAM files.
6. Using anvi-profile to produce single anvi'o profiles from individual BAM files.
7. Using anvi-merge, merge to generate single anvi'o profiles.

You will only need a samples.txt file and a few FASTQ files. We will go through a mock example with three small metagenomes in this section. A limited number of reads were selected to create these metagenomes. The following samples.txt file can be found in your working directory:

*sample group r1 r2*

*P1 G01 M-1\_S21\_L001\_R1\_001.fastq.gz M-1\_S21\_L001\_R2\_001.fastq.gz*

*P2 G02 M-2\_S22\_L001\_R1\_001.fastq.gz M-2\_S22\_L001\_R2\_001.fastq.gz*

*P3 G03 M-3\_S23\_L001\_R1\_001.fastq.gz M-3\_S23\_L001\_R2\_001.fastq.gz*

This file details the raw paired-end reads locations for the samples and 'groups'. The default name is samples.txt for the samples\_txt file, but you can change it in the config file.

Let's have a look at the config file config-megahit.json in the working directory.

*{*

*"workflow\_name": "metagenomics",*

*"config\_version": “2”,*

*"samples\_txt": "samples.txt",*

*"megahit": {*

*"--min-contig-len": 1000,*

*"--memory": 0.4,*

*"threads": 7,*

*"run": true,*

*}*

*}*

Every customizable parameter will be given a default value. We normally start with a default config file and delete every line that we don't want to keep. We have everything now to start. Let’s generate a workflow graph at this stage.

*Command*

*anvi-run-workflow -w metagenomics \*

*-c config-megahit.json \*

*--save-workflow-graph*

We can now start the workflow:

*Command*

*anvi-run-workflow -w metagenomics \*

*-c config-megahit.json*

Expected Results:

After completing all the steps in this pipeline, the users will be able to utilize these generate profiles for downstream work like manual refining and functional analyses. This workflow explained the steps from QC to mapping for generation of MAGs. A general introduction was provided about the config and samples.txt files to connect raw sequencing reads with sample details. Anvi’o coupled with snakemake workflows to generate profiles that could be used for downstream analyses.