**Installation instructions**

**Conda is not installed (long instructions):**

If you don’t have conda installed on your directories, complete these steps first. To see if you have conda installed you can open up a terminal, log in and look at the bottom of your terminal screen where you type your commands (Figure 1). If conda is installed you should have a (base) or a (XYZ) with the name of your conda environment. Conda is a general purpose environment and package manager allowing us to install nearly any program locally and is crucial for not just this pipeline but can be a powerful tool for any analysis.

A screen shot of a computer

Description automatically generated  
**Figure 1.**

If you are missing conda follow the following steps

1. Open up a terminal and connect to the compute cluster you will be using (e.g., petrichor HPC)
   1. Hint, you can use many different programs to do this like putty, or WSL, or PowerShell
2. Navigate to the directory you would like conda installed. Remember to change the user-ID to your username
   1. `cd /scratch3/user-ID/`
3. Make a new directory for conda
   1. `mkdir miniconda
4. Start an interactive session in your terminal which allows for downloading on the HPC (CSIRO specific step)
   1. `sinteractive -A [Accountcode] -c 1 -m 4G -t 1:00:00 -p io`
   2. Note: your account code is your OD-code. It can be found by typing `get\_project\_codes` into the terminal
5. Download the latest miniconda set up file (in this code, change the path to where you make the miniconda directory
   1. `wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86\_64.sh -O /scratch3/user-ID/miniconda/miniconda.sh`
6. Run the conda build script (remember to change the filepaths to where you saved the conda.sh
   1. `bash /scratch3/user-ID/miniconda/miniconda.sh -b -u -p /scratch3/user-ID/miniconda/`
7. Run the following commands in the terminal (remember to update your user path)
   1. `/scratch3/user-ID/miniconda/bin/conda init bash`
   2. `/scratch3/user-ID/miniconda/bin/conda init zsh`
8. Close your terminal and open a new one. After you open a new terminal conda will be installed and you will have a `(base)` in the corner of your screen

**Conda is installed (short instructions):**

If conda is installed, we can begin to install MetaDIVE.

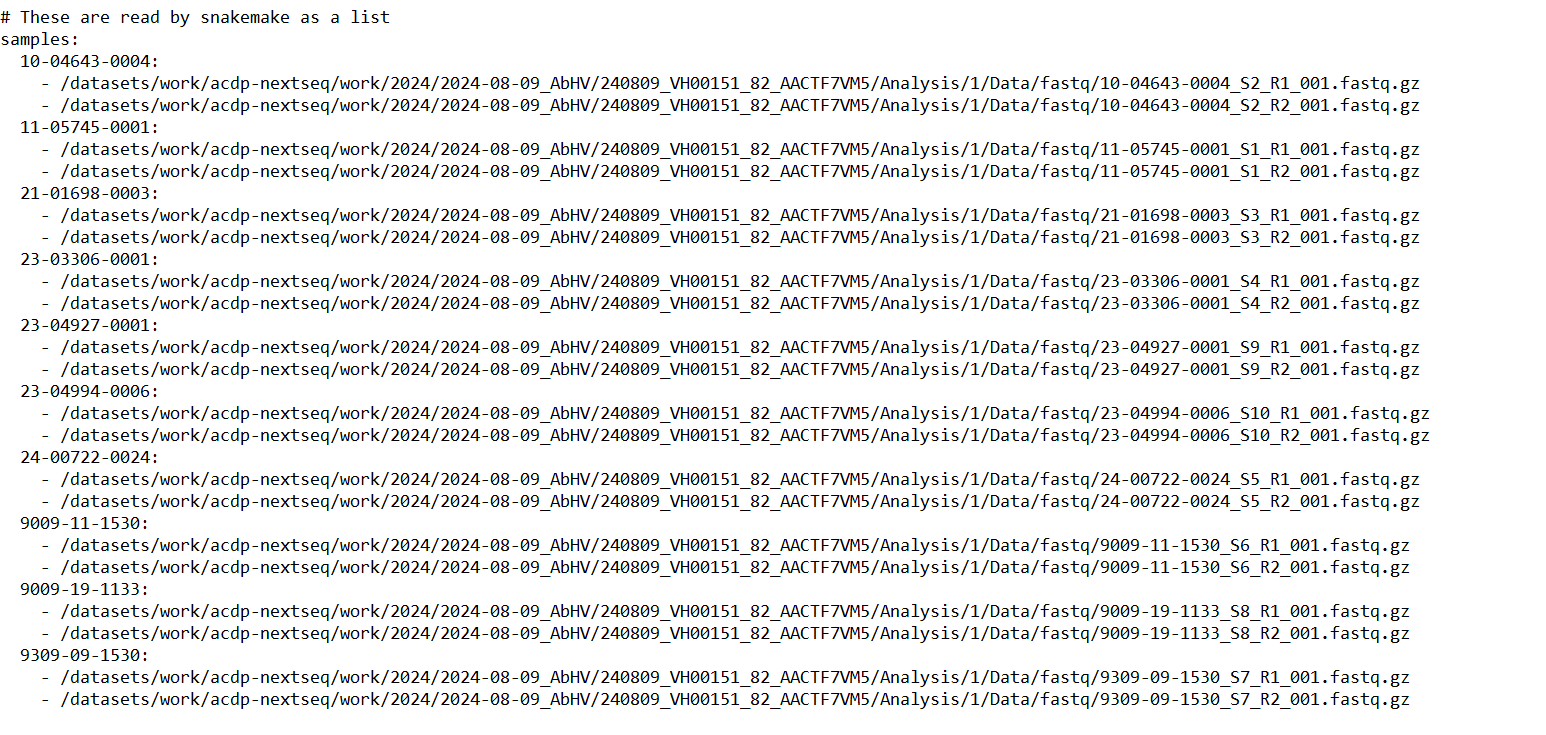
1. Start an interactive session in your terminal which allows for downloading on the HPC (CSIRO specific step)
   1. `sinteractive -A [Accountcode] -c 1 -m 4G -t 6:00:00 -p io`
   2. Note: your account code is your OD-code. It can be found by typing `get\_project\_codes` into the terminal
2. Install the program github large file storage into your conda environment. In your terminal you can be in any directory and run the following.
   1. `conda install conda-forge::git-lfs`
   2. This will install the program into your conda base environment after downloading the program. Once the program is finished downloading files a prompt will appear asking if you want to install the files y/n
   3. Type `y` into the terminal
3. Initialise the program git-lfs (activate it for use)
   1. `git lfs install`
4. In the terminal create and then move to the folder you want MetaDIVE to save to (this will be where the pipeline is built and where analysis will happen so having a folder name related to your project is useful to remember what is being done.
   1. `mkdir /scratch3/user-ID/analysis\_samplesX`
   2. `cd /scratch3/user-ID/analysis\_samplesX`
5. Download MetaDIVE off github (this uses the github programs installed earier)
   1. `git clone https://github.com/James-ODwyer/MetaDIVE.git`
6. Now you have downloaded the MetaDIVE pipeline into the directory. MetaDIVE is structured with an outer folder saying MetaDIVE and in this folder there are 3 main sub folders called pipeline, databases, and envs as shown in the below images (Figures 2, 3)

A white background with black dots

Description automatically generatedA white background with blue text

Description automatically generated  
Figures 2 and 3.

1. Navigate to the newly made envs directory
   1. `cd /scratch3/user-ID/analysis\_samplesX/MetaDIVE/envs`
2. **\*Some of the next steps will be easier using a file transfer program as well as a terminal\* for example on Windows you can use a program like WinSCP to connect to the HPC and move files around/open and edit them\***   
   Open up the file titled `Create\_conda\_environments.sh` and change line 2 of the file to your OD-code (CSIRO specific step). Save and close the file.
   1. E.g., change `#SBATCH --account=OD-229285` to your account number
   2. Hint: all files labelled as `.sh` here are text files. They can be opened up in programs like notepad, notepad++, or any internal editor on a file transfer program like WinSCP. Alternatively, you can edit a file directly in a terminal using the command `nano Create\_conda\_environments.sh` and saving the file after changed
3. Run the file `Create\_conda\_environments.sh`
   1. E.g., `sbatch Create\_conda\_environments.sh`
   2. Hint. This will take about 15-20 minutes to run and will install all programs required to run MetaDIVE directly into different conda environments.
4. Navigate to the databases directory
   1. `cd /scratch3/user-ID/analysis\_samplesX/MetaDIVE/databases`
5. Open up the files titled `download\_and\_build\_databases.sh`and `download\_kraken\_DBs.sh`and change line 2 of the file to your OD-code (CSIRO specific step). Save and close the file.
   1. E.g., change `#SBATCH --account=OD-229285` to your account number
6. Run the file `download\_and\_build\_databases.sh`
   1. E.g., `sbatch download\_and\_build\_databases.sh`
   2. Hint: this will take about 90-120 minutes to run and will install and index all databases for maximum speed during analysis. This only has to be done once and all future analyses of MetaDIVE can use these databases.
7. Navigate to the scripts folder in the pipelines folder (MetaDIVE/pipeline/scripts)
   1. `cd /scratch3/user-ID/analysis\_samplesX/MetaDIVE/pipeline/scripts`
8. Open the following script files up (they are text files, so can be opened up in any text editor) and change line 2 of the file to your OD-code (CSIRO specific step). Save and close each file.
   1. `downloadgenome.sh`
   2. `download\_viral\_genomes.sh`
   3. `download\_variety\_genomes\_high\_completion\_viruses.sh`
   4. Hint: the folder these files are in is at /scratch3/user-ID/analysis\_samplesX/MetaDIVE/pipeline/scripts . More instructions on editing text files are given in step 9.
9. Navigate to the pipeline folder
   1. ` cd /scratch3/user-ID/analysis\_samplesX/MetaDIVE/pipeline`
10. Open up the file titled run\_snakemake.sh and change lines 2,3,6,7,8 as described
    1. Line 2: update the OD-code to your account number e.g., change `#SBATCH --account=MyODcode `
    2. Line 3: Update the name of the project to something informative for your work e.g., change to `#SBATCH --job-name MY\_project`
    3. Line 6: Change the number of CPUs to use (as a general rule 8-16 for Miseq data, 16-32 for Nextseq data, and up to 48 for Novaseq or larger data) the larger the number the faster the analysis but the more compute resources required. E.g., #SBATCH --cpus-per-task 32
    4. Line 7: Change the memory to use (Memory required will depend on analyses you want to do, how many reads you have e.g., miseq vs Nextseq, the assembler you want to use, and how complicated your datset is e.g., less for an isolate, more for a pooled mosquito metagenome, even more for a complex soil or water sample ) e.g., #SBATCH --mem 120G
       1. Hint, there are more details in the config.yaml file describing memory. If unsure and you have >100m reads to analyse, set to around 90-120G
    5. Line 8: Change the time allocated to the analysis. Depending on the size of the data, CPUs used and complexity of the dataset, between 16 and 140 hours will be required. If unsure, set to 72-96 hours (the time is read as hh/mm/ss) as if the time is exceeded the run will halt and can be restarted again from where it was up to. E.g., #SBATCH --time 96:00:00
11. If you want to run trinity as the assembler create a directory where you want trinity to store temporary files.
    1. E.g., `mkdir /scratch3/user-ID/analysis\_samplesX/MetaDIVE/trinity\_temp
12. In the pipelines directory, open the file called config.yaml. This is the parameter file for the metagenomics run and the next steps will involve editing this. This file is formatted like a **Python** script and so spacing is important. All the spacing of all variables is already done correctly so care just needs to be taken to not delete the spacing already present
    1. Update the names and paths of raw data you want to analyse (Figure 4). (~ line 5)  
       The sample names format can be best described as   
       samples:   
       [space][space]samplename:  
       [space][space][space][space]-[space]Full \_path\_to\_samp1\_R1\_raw.fastq.gz  
       [space][space][space][space]-[space]Full \_path\_to\_samp1\_R2\_raw.fastq.gz  
       [space][space]samplename2:  
       [space][space][space][space]-[space]Full \_path\_to\_samp2\_R1\_raw.fastq.gz  
       [space][space][space][space]-[space]Full \_path\_to\_samp2\_R2\_raw.fastq.gz  
       etc etc (note the ‘-‘ present after the 4 spaces and before the path )
    2. Update the settings you want to use for the analysis run (~ lines 44 -144). All settings have a detailed description above them and what to type to change each setting (Figure 5).
    3. Update the API key to connect to NCBI (~ line 59). This is a key that can be generated by anyone with an NCBI or NIH account that allows for faster access to downloading from genbank. (For sequencing team, feel free to leave it as is if you don’t have a key right now).
    4. There are two databases used by this program that are not included in the databases folder. These are the Diamond Blast database and the Blastn nucleotide database. These two databases are very large and very common in bioinformatics and so it is best to install them once somewhere everyone on an HPC can access them. These are currently pointing to the CSIRO general databases and can remain the same when used on CSIRO HPCs but for other HPCs the paths will need to be updated updated (~Lines 197-200) (diamond\_database: blast\_nucleotide\_database: )  
       If they are not on your HPC these will need to be built following the installation instructions from each program
13. After all edits to the parameters are complete you can start analysing your samples. In the pipeline directory (you should still be in this directory from the last edits done) run
    1. `sbatch run\_snakemake.sh`



**Figure 4.** example of sample names and paths.

A white background with black text

Description automatically generated

**Figure 5.** Some settings which can be changed including sensitivity, various filtering settings, and assembly choice.