Angiosperms 353 metagenomic pipeline

fuck you

2023-02-08

This document is gonna make your eyes burn. And not in a *Heaven's Gonna Burn Your Eyes* kinda way. Their will be heinous crimes committed against multiple shell script languages, python, and R - heck I'll probably throw in a couple against tex.

Essentially *woe is me*, my masters was a tragicomedy of stupidity, and here I try to converge the most import details as scattered across HackMD scripts. The things that confuse you in here confuse me too bud, grab an old style a few bottles of Malort or Gin, and let's please hope you still smoke and live in a progressive enough area you can do so indoors - and grab a fresh pouch of Stokkebye.

I am going to assume that you use Linux. If you are windows, well it's time to switch it over. If you are on Mac, hopefully you bought it hot.

If you are like me, and you didn't want to go head's in on a linux distro here is some advice to get a dual boot going on Windows 10.

1: Installing Ubuntu on a dual boot advice

> 'Free your mind and your software will follow' - George Clinton mostly

I was able to largely follow two guides to accomplish this.

My instructions below are brief.

For the dualboot installation download the most recent stable version of Ubuntu as an ISO image. I have previously used Balena Etcher to flash the image onto a USB thumbdrive, but have switched to RUFUS for the last couple boots. I recommend RUFUS now.

Following this use an administrator account to access the 'Command Line' and 'Run as Administrator'.

Now enter 'diskmgmt.msc', in the gui click on the 'C:' drive and 'Shrink Volume' of it. Shink to between 20,000-80,000mib. I did 75,000 - should be way overkill.

To dual boot I had to go with the old approach of the F12 at startup. If I remember this was the same route I had to use for Mint on a Lenovo ThinkPad T-530. So it could be a Lenovo thing. 1a: Troubleshooting Windows Shutdown

As a consequence of dual booting it may be hard to shutdown your hardware from the WindowsOS. This is a workaround, which will make it take longer to log into Windows, but will let your hardware shutdown.

This Unix & Linux StackExchange Thread had a solution which worked for me. https://unix.stackexchange.com/questions/247184/unable-to-shutdown-windows-after-installing-grub

```
(1) Go to control panel.
(2) Find power options.
(3) From the left menu click on "Choose what the power button does".
(4) Click on "Change settings that are currently unavailable."
(5) Go to the bottom of the page, uncheck "Turn on fast startup" and save changes.
```

After doing the typical updates and processes associated with installing Linux you are good to go! 1b: Troubleshooting Time Zones

You'll find on your Windows OS the time is probably wrong. Run this on the linux side of the computer to fix this. PS. to access the command line, **CONTROL + ALT + T**.

```
timedatectl set-local-rtc 1
```

Now you need to update the time manually in windows and you should be good.

2: Check to see if the data transfer from the genomics core you was clean

Strange things happen here and we do not know why.

Here we have some quick lines to inspect the files you transferred over in order to figure out if you are missing something.

```
$ mkdir /media/sagesteppe/Genomics/data_summaries

# A) find distinct Sample ID's.
$ cd /media/sagesteppe/Genomics/Original_zip
$ find . -regex ".*\_S[0-9][0-9]?[0-9]?_.*" > ../data_summaries/sample_ids.txt
$ sed -i -E "s/.*(S[0-9][0-9]?[0-9]?).*/\1/" ../data_summaries/sample_ids.txt
$ sort -u ../data_summaries/sample_ids.txt > ../data_summaries/unique_seqs.txt

# B) count the distinct samples to determine how many are present.
$ wc -l ../data_summaries/unique_seqs.txt

# C) determine which samples have each of 4 reads associated with them.
$ sort ../data_summaries/sample_ids.txt | uniq -c > ../data_summaries/reads_per_sample.txt

# D) Identify the missing & extra reads.
$ awk ' $1 < 7 { print }' ../data_summaries/reads_per_sample.txt > ../data_summaries/missing_reads.txt
$ awk ' $1 > 8 { print }' ../data_summaries/reads_per_sample.txt > ../data_summaries/extra_reads.txt

# E) Report the missing & extra reads.
$ cat ../data_summaries/missing_reads.txt
$ cat ../data_summaries/extra_reads.txt

# F) Determine which reads are empty.
" find ~/lists -empty " # not solved.

# G) Determine if any samples are missing.
$ sed -i -E 's/S//g' ../data_summaries/unique_seqs.txt # remove 'S'
$ sort -n ../data_summaries/unique_seqs.txt -o ../data_summaries/unique_seqs.txt # arrange numerically

$ min=$(sort -n ../data_summaries/unique_seqs.txt | sed -n '1p') #lowest sample number
$ max=$(sort -n ../data_summaries/unique_seqs.txt | sed -n '$p') # highest sample number
$ seq "$min" 1 "$max" > ../data_summaries/full_seq.txt # create range

$ difference=$(( $(wc -l ../data_summaries/full_seq.txt |  sed 's/[^0-9]*//g') - $(wc -l ../data_summar
$ echo "$difference"  # compared to range of all sample numbers

$ comm -3 ../data_summaries/unique_seqs.txt ../data_summaries/full_seq.txt
# you may get a message in your output like "comm: file 2 is not in sorted order" - but for me it is so
# H) Determine file size of read. (In MiB)
$ du -a --block=1M > ../data_summaries/fsize_raw_reads.txt

$ cd
```

3: Install somestuff

Download Anaconda

You can go to the Anaconda downloads page and find the version relevant to you and copy the link. It is easier to install from command line then clicking and dragging IMO.

```
sudo apt-get update -y
sudo apt-get upgrade -y

cd /tmp
wget https://repo.anaconda.com/archive/Anaconda3-2021.11-Linux-x86_64.sh
bash Anaconda3-2021.11-Linux-x86_64.sh
```

Now you will need to respond to some prompts in the terminal. I let my conda install directly within my user profile. You should now have conda installed.

so your terminal looked like this: `sh reed@reed-steppe:~` BUT NOW it looks like this: `sh (base) reed@reed-steppe:~`, maybe you didn' notice a change here if you are new to this, but if you got the based ya got what ya want.

Check the version of Conda, and check for updates and install them.

```
conda --version
conda update conda
```

CONDA ON

From here on out we are going to work in the Conda environment.

We will now create an environment to store all of our metagenomics materials in. These environments serve to isolate the contents of projects.From now on pretty much everything we are going to do should be occuring in this 'environment', to illustrate the point

```
conda create --name metagenomics
conda activate metagenomics # we always need to activate this !!!!
conda info --envs
```

Now we will turn to installing biopython - make sure you are in the right project ('metagenomics' !)

We will now install biopython, which is prone to fits of rage for most people, but generally plays nicely with conda `sh conda install biopython`.

`sh $PATH` bash: /home/reed/anaconda3/envs/metagenomics/bin:/home/reed/anaconda3/condabin:/usr/local/sbin:/usr/local/ No such file or directory

Notice that each of the above paths are *within* the anaconda3/envs.

Github

All linux are meant to ship with GH, I think even if you do the super spartan install. Although maybe they just ship with Git since gh changed hands. Anyways.

```
$ git --version
# if you do not have it...
$ sudo apt install git-core
```

Homebrew

this is supposedly way easier on Mac's, however we should not condone sweat shop labor.

```
# ensure you are still in metagenomics folder if not conda activate metagenomics !!!!
sudo apt install curl
git clone https://github.com/mossmatters/HybPiper.git

/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install.sh)"
echo 'eval "$(/home/linuxbrew/.linuxbrew/bin/brew shellenv)"' >> /home/reed/.profile
eval "$(/home/linuxbrew/.linuxbrew/bin/brew shellenv)"
sudo apt update -y
brew install blast bwa gcc parallel spades samtools
parallel --bibtex

sudo apt-get update -y
sudo apt-get upgrade -y

sudo apt-get install -y exonerate # now let's grab exonerate!!!
```

Ensure the file is in a suitable location such as…

```
cd /home/reed/HybPiper
python3 reads_first.py --check-depend
```

this should hopefully say all packages can be found!

```
cd /test_dataset
bash run_tests.sh
```

if you run this and all kinds of stuff happens that is good.

Create a big bad bioinformatics folder

```
mkdir sequenceData
cd sequenceData
```

copy the original files over to your computer

```
cp -R /media/reed/Genomics/Original_zip .
```

FastQC

this worked for me

```
sudo apt update -y
sudo apt upgrade -y
sudo apt install fastqc # install

$ mkdir /media/sagesteppe/Genomics/Pre_trim_FastQC # to hold output
$ cd /media/sagesteppe/Genomics/Original_zip # fastqs to evaluate
$ fastqc *fastq.gz --outdir=/media/sagesteppe/Genomics/Pre_trim_FastQC -t 3 # torun process
$ threads=$(nproc --all)
$ threads=$(($threads - 1))
$ echo $threads # to print results.

# hopefully dynamic using all threads - 1
$ fastqc *fastq.gz --outdir=/media/sagesteppe/Genomics/Pre_trim_FastQC -t echo$threads
```

Trimmomatic

Most of my trimmomatic advice came from a blog hosted by the University of Texas at Austin, and John Zhang.

This is what worked for me, I did a few things they did not. Realistically all of this advice should be dumb and you should be able to install it layers beneath this.

```
cd /usr/bin
wget http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip

unzip Trimmomatic-0.39.zip
rm Trimmomatic-0.39.zip
export PATH=$PATH:/home/reed/Trimmomatic-0.39/

java -jar Trimmomatic-0.39/trimmomatic-0.39.jar
```

OK so trimmomatic is installed, but it is a long path to call it, so we will create a bash script to do that. One way to do that is this.

```
nano Trimmomatic-0.39/trimmomatic #You will now have a strange new terminal before you. Don't Panic. Pl
```

–NANO INTERFACE START–

```
!/bin/bash
java -jar $HOME/Bioinformatics/Trimmomatic-0.39/trimmomatic-0.39.jar $@`
```

Note that the #!/bin/bash argument has to be on it's own line. It is defining that the .txt file you are writing is meant to be processed by Bash.

Now to safely exit nano, on your keyboard: > **ctrl + X**

and to to close the terminal click **'Y'** to say 'YES'

Now this is an important part, it may ask if you want to update the name of the script you just wrote - do not! My script name is never displayed, but you have already supplied the name and the directory for it to be saved too!

On your keyboard hit **ENTER**

–NANO INTERFACE END–

Now we will need to make it so that this script is executable. We can do this as so: `{sh} $ chmod +x ~/Bioinformatics/Trimmomatic-0.39/trimmomatic`

And now you should be able to run Trimmomatic by: `sh trimmomatic`

Note if you closed your terminal you will need to re-add the folder to the $PATH. You can add this to the $PATH forever,

Install Mega353 for custom target files

```
python3.9 -m pip install pandas # dependency
cd ~/.local/bin
~/.local/bin$ git clone https://github.com/chrisjackson-pellicle/NewTargets.git
$ cd ~/.local/bin/NewTargets
~/.local/bin/NewTargets$ unzip mega353.fasta.zip
# cd is not the best way to do the  last step but how i did it.
```

Install R and Rstudio

these steps worked to install R

```
sudo apt update -qq
sudo apt install --no-install-recommends software-properties-common dirmngr
sudo wget -qO- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc | sudo tee -a /etc/apt/
sudo add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu $(lsb_release -cs)-cran40/"
apt install --no-install-recommends r-base
sudo usermod -a -G staff sagesteppe
sudo apt-get update -y
sudo apt-get install -y libssl-dev
sudo apt-get install libcurl4-openssl-dev
sudo add-apt-repository ppa:c2d4u.team/c2d4u4.0+
sudo apt install --no-install-recommends r-cran-rstan
sudo apt-get install libtiff-dev libjpeg-dev libpng-dev
sudo apt-get install libblas-dev liblapack-dev
```

```
$ sudo add-apt-repository universe
$ sudo apt-get install gdebi-core
$ sudo apt install dpkg-sig
$ cd /home/sagesteppe/Downloads

~/Downloads$ gpg --keyserver keyserver.ubuntu.com --recv-keys 3F32EE77E331692F

#$ wget --no-check-certificate https://www.rstudio.com#/products/rstudio/download/#download
#$ curl -k -O -L https://www.rstudio.com/products/rstudio/download/#download" "rstudio-2021.09.1-372-am
 # currently download by hand :-(   #)

~/Downloads$ dpkg-sig --verify rstudio-2021.09.1-372-amd64.deb
~/Downloads$ ls *.deb
~/Downloads$ sudo gdebi ./rstudio-2021.09.1-372-amd64.deb
~/Downloads$ rstudio # ensure installation works.
> quit() # (in gui)
~/Downloads$ rm rstudio-2021.09.1-372-amd64.deb
~/Downloads$ cd
```

Process read data

```
conda activate metagenomics #(if not already)

cd sequenceData
mkdir raw_reads
find Original_zip/ -type f -print0 | xargs -0 mv -t raw_reads/ # copy all of our reads out of the origi
mkdir Trimmed_reads
```

```
cp ~/Trimmomatic-0.39/adapters/TruSeq3-PE.fa raw_reads

trimmomatic PE -phred33  1a_S76_L002_R1_001_paired.fastq.gz 1a_S76_L002_R2_001_paired.fastq.gz -baseout
```