16S Intermediate Bioinformatics Training - Software setup and testing

[1. Install Singularity on Ubuntu](#)

[2. Install Nextflow](#)

[3. Download the Rstudio Singularity image](#)

[4. Running RStudio on a cluster](#)

[5. Running Rstudio on a server](#)

[6. Running the DADA2 Nextflow pipeline on test data](#)

1. Install Singularity on Ubuntu

Singularity can be installed system wide. The most up to date instructions are found here.

Lets install Singularity v3.1.1

```
$ sudo apt-get update && sudo apt-get install -y \
    build-essential \
    libssl-dev \
    uuid-dev \
    libgpgme11-dev \
    squashfs-tools \
    libseccomp-dev \
    pkg-config

$ export VERSION=1.12.5 OS=linux ARCH=amd64 && \
    wget https://dl.google.com/go/go$VERSION.$OS-$ARCH.tar.gz && \
    sudo tar -C /usr/local -xzvf go$VERSION.$OS-$ARCH.tar.gz && \
    rm go$VERSION.$OS-$ARCH.tar.gz

$ echo 'export GOPATH=${HOME}/go' >> ~/.bashrc && \
    echo 'export PATH=/usr/local/go/bin:${PATH}:${GOPATH}/bin' >>
~/.bashrc && \
    source ~/.bashrc

$ go get -u github.com/golang/dep/cmd/dep

$ go get -d github.com/sylabs/singularity

$ export VERSION=v3.1.1 # or another tag or branch if you like &&
\
```

```
    cd $GOPATH/src/github.com/sylabs/singularity && \
    git fetch && \
    git checkout $VERSION

$ ./mconfig && \
    make -C ./builddir && \
    sudo make -C ./builddir install

$ singularity version
3.1.1
```

Looks OK.

2. Install Nextflow

Nextflow needs to be installed in each user's home directory (permissions assigned to the user) and be available on the users path.

Requirements: Java 1.8 or later is required. Also see Nextflow setup instructions here.

```
$ mkdir /home/user/nextflow
$ cd  /home/user/nextflow
$ curl -s https://get.nextflow.io | bash
$ echo "export PATH=$PATH:/home/user/nextflow/" >>
/home/user/.bashrc
$ sudo su user
$ nextflow -v
nextflow version 19.04.1.5072
```

Nextflow version 19.04 is fine.

3. Download the Rstudio Singularity image

Download the Rstudio Singularity image here.

4. Running RStudio on a cluster

This setup is focus on running a RStudio Singularity container on a SLURM cluster. For PBS/Torque or SGE clusters the only difference would be in the way that you would submit your interactive jobs.

Firstly one should configure ssh in such a way that it is simple to connect to a worker node once a job is running. The easiest way it to add the following to your local `~/.ssh/config` file:

```
Host *.ilifu.ac.za
```

```
    User USERNAME
    ForwardAgent yes

Host slwrk-*
    Hostname %h
    User USERNAME
    StrictHostKeyChecking no
    ProxyCommand ssh headnode nc %h 22
```

One should substitute in your headnode, workernode and USERNAME settings in the above script.

Next is the process of starting an interactive job and launching RStudio. To begin start an interactive job – below is an example of launching a single node / 1 core job with 8Gb of ram:

```
USERNAME@slurm-login:~$ srun --nodes=1 --ntasks 1 --mem=8g --pty
bash
USERNAME@slwrk-103:~$
```

Once the interactive session has begun on a specific node (in this case slwrk-103), RStudio can be launched as follows:

```
USERNAME@slwrk-103:~$ RSTUDIO_PASSWORD='Make your own secure
password here' /cbio/images/bionic-R3.6.1-RStudio1.2.1335-bio.simg

Running rserver on port 37543
```

This will launch an RStudio server listening on a random free port (in this case `37543`). Now one needs to port-forward from your local machine to the host machine. One connects to the appropriate node by running:

```
$ ssh slwrk-103 -L8082:localhost:37543
```

On your local machine. Specifically what this does is forward traffic on your local machine's port 8082 to the worker node's port 37543 (and it knows how to connect to `slwrk-103` by using the `.ssh/config` settings above). One may use any free local port – ssh will complain if you choose something that is not free with an error message approximating:

```
bind [127.0.0.1]:8000: Address already in use
channel_setup_fwd_listener_tcpip: cannot listen to port: 8000
```

Finally in your browser you can connect to `http://localhost:8082` and you can login with your `USERNAME` and the `RSTUDIO_PASSWORD` which you set.

5. Running Rstudio on a server

ssh into server

```
$ ssh USERNAME@SERVERNAME
```

Start the server: (assume image is stored under /share/images)

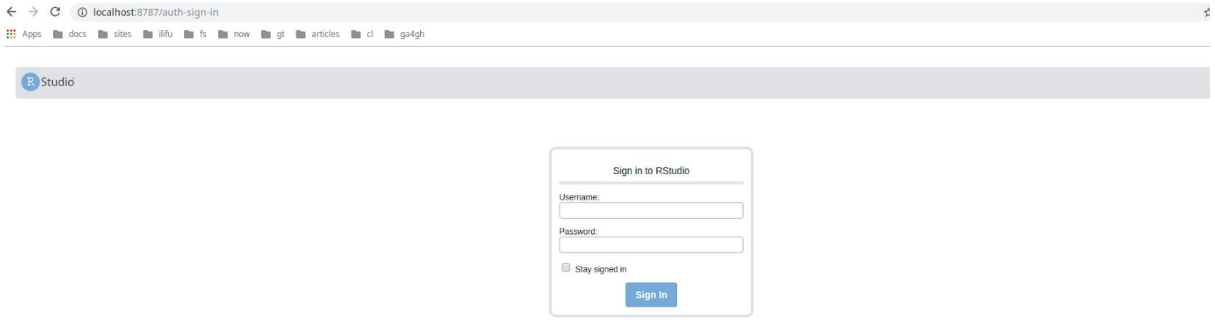
```
$ RSTUDIO_PASSWORD='Make your own secure password here'
/share/images/bionic-R3.6.1-RStudio1.2.1335-bio.simg

Running rserver on port 37543
```

Now use that port number in all the port settings below

From local machine open up a new terminal and start a new session:
`ssh USERNAME@SERVERNAME -L8082:localhost:37543`

3) Go to `http://localhost:8082` and you'll be prompted for a username/password. The username is your usual `server` system `USERNAME` and the password is the `RSTUDIO_PASSWORD`.



DADA2 should be installed already

```
> library("dada2")
Loading required package: Rcpp
Registered S3 methods overwritten by 'ggplot2':
method from
[.quosures rlang
c.quosures rlang
```
print.quosures rlang
> packageVersion("dada2")
[1] ‘1.12.1’

Looks OK.

6. Running the DADA2 Nextflow pipeline on test data

```
$ mkdir $HOME/test-data
$ cd $HOME/test-data
$ wget http://web.cbio.uct.ac.za/~gerrit/downloads/dog_stool.tgz
$ tar -xzvf dog_stool.tgz

$ mkdir $HOME/ref-data
$ cd $HOME/ref-data
$ wget
https://zenodo.org/record/1172783/files/silva_nr_v132_train_set.fa
.gz
$ wget
https://zenodo.org/record/1172783/files/silva_species_assignment_v
132.fa.gz

$ cd $HOME
$ git clone https://github.com/grbot/16S-rDNA-dada2-pipeline
$ cd $HOME/16S-rDNA-dada2-pipeline

$ nextflow run main.nf -profile standard
--reads="$HOME/test-data/*_R{1,2}.fastq.gz" --trimFor 24 --trimRev
25 --reference="$HOME/ref-data/silva_nr_v132_train_set.fa.gz"
--species="$HOME/ref-data/silva_species_assignment_v132.fa.gz"
--outdir="$HOME/out"

N E X T F L O W  ~  version 19.04.1
Launching `main.nf` [soggy_gilbert] - revision: 1696132777
===================================
uct-cbio/16S-rDNA-dada2-pipeline  ~  version 0.4
===================================
Run Name        : soggy_gilbert
Reads           : /home/gerrit/test-data/*_R{1,2}.fastq.gz
trimFor         : 24
trimRev         : 25
truncFor        : 248
truncRev        : 212
truncQ          : 2
maxEEFor        : 2
maxEERev        : 2
maxN            : 0
maxLen          : Inf
```

```
minLen          : 50
rmPhiX          : T
minOverlap      : 20
maxMismatch     : 0
trimOverhang    : F
species         :
/home/gerrit/ref-data/silva_species_assignment_v132.fa.gz
pool            : pseudo
Reference       :
/home/gerrit/ref-data/silva_nr_v132_train_set.fa.gz
Max Memory      : 384 GB
Max CPUs        : 40
Max Time        : 3d
Output dir      : /home/gerrit/out
Working dir     : /home/gerrit/16S-rDNA-dada2-pipeline/work
Container       : docker://quay.io/cbio/16s-rdna-dada2-pipeline
Current home    : /home/gerrit
Current user    : gerrit
Current path    : /home/gerrit/16S-rDNA-dada2-pipeline
Script dir      : /home/gerrit/16S-rDNA-dada2-pipeline
Config Profile : standard
=======================================
[warm up] executor > local
WARN: Singularity cache directory has not been defined -- Remote
image will be stored in the path:
/home/gerrit/16S-rDNA-dada2-pipeline/work/singularity
Pulling Singularity image
docker://quay.io/h3abionet_org/h3a16s-fastqc [cache
/home/gerrit/16S-rDNA-dada2-pipeline/work/singularity/quay.io-h3ab
ionet_org-h3a16s-fastqc.img]
Pulling Singularity image
docker://quay.io/cbio/16s-rdna-dada2-pipeline [cache
/home/gerrit/16S-rDNA-dada2-pipeline/work/singularity/quay.io-cbio
-16s-rdna-dada2-pipeline.img]
executor >  local (27)
[6c/5aa1b6] process > runFastQC                    [100%] 4 of 4 ✔
[2b/6c88b3] process > runMultiQC                   [100%] 1 of 1 ✔
[35/2bbc83] process > filterAndTrim                [100%] 4 of 4 ✔
[7d/1a6c45] process > runFastQC_postfilterandtrim  [100%] 4 of 4 ✔
[64/9eafb4] process > LearnErrorsFor               [100%] 1 of 1 ✔
[7b/a4322c] process > mergeTrimmedTable            [100%] 1 of 1 ✔
[ca/cdc08f] process > LearnErrorsRev               [100%] 1 of 1 ✔
[29/012c81] process > runMultiQC_postfilterandtrim [100%] 1 of 1 ✔
[aa/13e057] process > SampleInferDerepAndMerge     [100%] 4 of 4 ✔
[57/22aca5] process > SequenceTable                [100%] 1 of 1 ✔
[f1/517201] process > mergeDadaRDS                 [100%] 1 of 1 ✔
```

```
[01/46b8b3] process > ChimeraTaxonomySpecies       [100%] 1 of 1 ✔
[8f/d1c503] process > ReadTracking                 [100%] 1 of 1 ✔
[e7/9874c0] process > AlignAndGenerateTree         [100%] 1 of 1 ✔
[d1/3e5f94] process > BiomFile                     [100%] 1 of 1 ✔
Completed at: 14-Aug-2019 13:18:53
Duration    : 27m 48s
CPU hours   : 0.8
Succeeded   : 27
```

Looks OK.