D I F F R N A : A RNA-Seq analysis pipeline

Izem Mouhoubi Théo Roncalli Gustavo Magaña López

2021-11-30

Contents

Chapter 1

Prerequisites

1.1 Hardware and Operating System

The pipeline was developed and tested on Ubuntu 20.04.3 LTS on top of the `(GNU/Linux 5.4.0-88-generic x86_64)` kernel. The output of the commands `uname` and `neofetch` are provided to further detail our configuration.

```
$ uname -a
Linux machine329a7396-059f-41aa-94c7-4c41b4ec8290 5.4.0-88-generic #99-Ubuntu SMP Thu Sep 23 17:2

$ neofetch
            .-/+oossssoo+/-.               ubuntu@machine3f9ae3dc-6a4d-46b7-8131-04f00a1be146
        `:+ssssssssssssssssss+:`           -------------------------------------------------
      -+ssssssssssssssssssyyssss+-         OS: Ubuntu 20.04.3 LTS x86_64
    .ossssssssssssssssssdMMMNysssso.       Host: OpenStack Compute 18.2.1-1.el7
   /ssssssssssshdmmNNmmyNMMMMhssssss/      Kernel: 5.4.0-88-generic
  +ssssssssshmydMMMMMMMNddddyssssssss+     Uptime: 13 hours, 14 mins
 /sssssssshNMMMyhhyyyyhmNMMMNhssssssss/    Packages: 719 (dpkg), 4 (snap)
.ssssssssdMMMNhsssssssssshNMMMdssssssss.   Shell: bash 5.0.17
+sssshhhyNMMNyssssssssssssyNMMMysssssss+   Theme: Adwaita [GTK3]
ossyNMMMNyMMhsssssssssssssshmmmhssssssso   Icons: Adwaita [GTK3]
ossyNMMMNyMMhsssssssssssssshmmmhssssssso   Terminal: /dev/pts/0
+sssshhhyNMMNyssssssssssssyNMMMysssssss+   CPU: Intel (Haswell, no TSX, IBRS) (16) @ 2.294GHz
.ssssssssdMMMNhsssssssssshNMMMdssssssss.   GPU: 00:02.0 Cirrus Logic GD 5446
 /sssssssshNMMMyhhyyyyhdNMMMNhssssssss/    Memory: 635MiB / 64323MiB
  +sssssssssdmydMMMMMMMNddddyssssssss+
   /ssssssssssshdmNNNNmyNMMMMhssssss/
    .ossssssssssssssssssdMMMNysssso.
      -+sssssssssssssssssyyyssss+-
        `:+ssssssssssssssssss+:`
```

```
            .-/+oossssoo+/-.
```

This configuration was actually a virtual machine hosted on Biosphere's RAIN-Bio a cloud service maintained by the French Institue of Bioinformatics (*Institut Français de Bioinformatique*).

1.2 BioPipes: a Biosphere-commons app

The instance of the virtual machine we used is called *BioPipes*. It provides the most notable bioinformatics pipeline tools:

- nextflow
- snakemake
- cwltool

1.3 Main Tools

Their versions are specified to maximise reproducibility:

```
$ conda --version
conda 4.11.0
$ nextflow -v
nextflow version 21.10.0.5640
$ docker --version
Docker version 20.10.11, build dea9396
```

Detailed information about our development docker installation :

```
$ docker info
Server:
 Containers: 0
  Running: 0
  Paused: 0
  Stopped: 0
 Images: 4
 Server Version: 20.10.11
 Storage Driver: overlay2
  Backing Filesystem: xfs
  Supports d_type: true
  Native Overlay Diff: true
  userxattr: false
 Logging Driver: json-file
 Cgroup Driver: cgroupfs
 Cgroup Version: 1
 Plugins:
  Volume: local
```

```
 Network: bridge host ipvlan macvlan null overlay
 Log: awslogs fluentd gcplogs gelf journald json-file local logentries splunk syslog
Swarm: inactive
Runtimes: io.containerd.runc.v2 io.containerd.runtime.v1.linux runc
Default Runtime: runc
Init Binary: docker-init
containerd version: 7b11cfaabd73bb80907dd23182b9347b4245eb5d
runc version: v1.0.2-0-g52b36a2
init version: de40ad0
Security Options:
 apparmor
 seccomp
  Profile: default
Kernel Version: 5.4.0-88-generic
Operating System: Ubuntu 20.04.3 LTS
OSType: linux
Architecture: x86_64
CPUs: 16
Total Memory: 62.82GiB
Name: machine3f9ae3dc-6a4d-46b7-8131-04f00a1be146
ID: XT4Y:2HUL:HXEA:CDXV:ERC7:Z7JZ:YYRU:WZBT:ERCU:6GGA:OBZ6:QLXE
Docker Root Dir: /mnt/docker-data
Debug Mode: false
Registry: https://index.docker.io/v1/
Labels:
Experimental: false
Insecure Registries:
 127.0.0.0/8
Live Restore Enabled: false
```

Chapter 2

Dependencies

2.1 Software

The pipeline runs on nextflow a domain-specific language created to automate data-analysis pipelines whilst maximising reproducibility. Nextflow enables scientists to focus on their analyses, isolating different parts of the pipeline into processes whose dependencies can be dealt with using containers and virtual environments with technologies such as Docker, Singularity, and Anaconda.

The recommended way to install `nextflow` is via `conda`, using the environment file.

```
conda env create -f nextflow_conda_env.yml # will create an env called "nextflow"
conda activate nextflow
# You can edit the file at your choice, specially if the environment name conflicts
# with a preexisting conda env on your system
```

Docker should be installed as well:

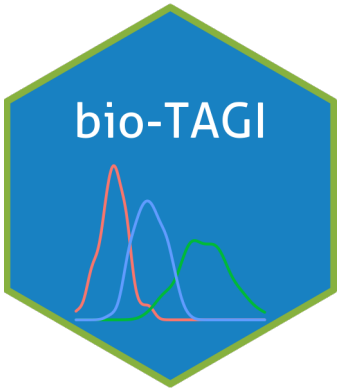
```
sudo apt install docker
```

Once nexflow is installed, it will automatically retrieve the docker images used within the pipeline.

Chapter 3

Workflow

Nextflow workflows should form a *DAG* (i.e. directed acyclic graph), which represents the flow of data through the different steps required to produce the final result.



This pipeline will generate a set of figures, representing differential gene expression analysis of RNA-Seq data.

Chapter 4

Methods

We describe our methods in this chapter.

Math can be added in body using usual syntax like this

4.1 math example

p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt{(\frac{p(1-p)}{n})} \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$

You can also use math in footnotes like this[1].

We will approximate standard error to 0.027[2]

[1] where we mention $p = \frac{a}{b}$

[2] p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt{(\frac{p(1-p)}{n})} \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.