DUESSELPORE Webserver manual

This is the instruction of using Duesselpore webserver to process RNAseq data.
Video of instruction can be found at:

Data availability:

- Docker image on Dockerhub https://hub.docker.com/repository/docker/thachdt4/duesselpore
- Light weight testdata at https://iufduesseldorf-my.sharepoint.com/:u:/g/personal/thach_nguyen_iuf-duesseldorf_de/ES4BsdfJSKNHl-mDUR3BogcBEmdOawVTRy-eRXU3-XeG2A?e=Kq9O2e
- Full test data. https://iufduesseldorf-my.sharepoint.com/:u:/g/personal/thach_nguyen_iuf-duesseldorf_de/EWIk4CLauThHk61_5rItjEcBOP4CJstbyCN9yN3ty36A7g?e=zRUf1T

1. Install and configure webserver

1.1. System requirement

- CPU: 2.0 GHz (Intel architecture, acceleration support) 4 cores or higher (8 cores recommended)
- System memory (RAM): 8 GB or higher (16 GB recommended)
- Diskdrive: 200 GB free space (1TB recommended)
- Host operating system Window 10, Linux (Ubuntu >=18.04 or Fedora) or MacOS (Intel)

1.2. Installation

1.2.1 Install Docker and setup webserver Follow Docker install instruction from https://www.docker.com/get-started. After install Docker. Open the terminal (on Linux, MacOS) or WSL(on Window). You may have to run it as superuser. NGS_webserver/settings.py.

```
$docker run -it -p 8000:8000 thachdt4/duesselpore:running python3\
 /home/ag-rossi/projects/duesselpore/manage.py runserver 0.0.0.0:8000
```

Depend on your host Operating system and your IP address range, you may have to configure the Docker IP address (default is 172.17.0.2). IMPORTANT NOTE: If you change your Docker address, you have to add your IP Address in to the ALLOWED_HOST field in /home/ag-rossi/projects/duesselpore/

2.2. Using webserver

2.2.1. Access webserver Now you can use your webserver within your Local Area Network (LAN) with a regular web browser (e.g., Firefox or Google Chrome HTTP port: 8000) http://localhost:8000/duesselpore.

2.2.2. Data preparation Users can upload fastq files as ONE compressed zip file: each subfolder contains several replicas with one experimental condition. Your decompressor support most common compression program in Window and Linux such as zip, gunzip, 7z etc NOTE: files and folders' names must contain only alphabetic and numeric characters. Below is an example of data separated into two conditions, 'condition1' and 'condition2'. Please check the structure and the director name of your data carefully, all the name of analysis are generated by directory and file names.

```
fastq/(folder)
   condition1 (subfolder)
       condition1_replica1.fastq (single fastq file)
       condition1_replica2.fastq (single fastq file)
   condition2 (subfolder)
        condition2_replica1.fastq (single fastq file)
        condition2_replica2.fastq (single fastq file)
        condition2_replica3.fastq (single fastq file)
```

How to merge multiple fastq files into a single file: On Linux terminal:

```
$ cat /path/to/fastq/files/*.fastq > /your/new/location/output.fastq
```

On Window command prompt (path syntax is different):

```
$ type \path\to\fastq\files\*.fastq> \your\new\location\output.fastq
```

2.2.3. Setup running parameter: First, select one group among your groups as the reference group. Select the gene (transcriptome) counting method, then select the differential expression algorithm you want to analyze. To run the analysis, we have to set up other parameters of the analysis function. There are some optional parameters, e.g., ReadCountMinThreshold, Logfold change threshold, adjPValueThreshold. After submit we can wait for the result. Advanced users can customize the RNA.R code to develop a new workflow. The figure bellow exlain the web input form.

2.2.4. Collecting the results: The run time depends on your data size and the system speed. For our standard dataset, which contains 6 replicates, approximate totals 16 million reads (around 15 Gb), the run time is around 6 hours. For lightweight test data, running time is 1.5 hours. After the computation is completed, all the results are downloaded from the browser. The interactive HTML file is exported with different plots. Users can continue the offline analysis on the Linux virtual machine directory at /home/ag-

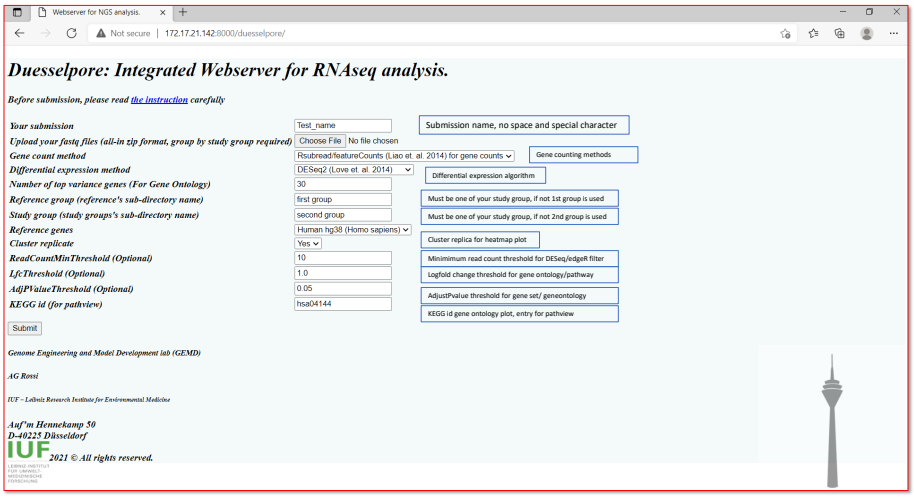


Figure 1: Input web form explaination

rossi/duesselpore/users_file/{your session id}. Experienced users will be able to further analyze their data by editing the R script. NGS data will occupy high volume space, therefore we recommend erasing the data on the virtual machine regularly. The Sample result is in the Support Information, or sample_result/report.html in sample_result.zip. Please note that while most of the analysis do not require an internet connection, gene ontology and disease pathway will require an internet connection.

Use Virtualbox version. (Not recommended)

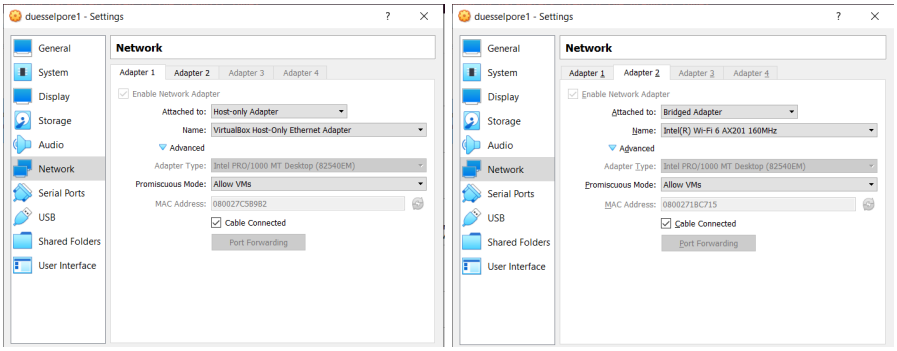
1.2.1 Download and install VMWare

- Virtualbox image at: https://iufduesseldorf-my.sharepoint.com/:u:/g/personal/thach_nguyen_iuf-duesseldorf_de/EZB0I5s_Gq5MnPbg1g69WvsBACVULQFQ3s2wJjc-pyN3PA?e=Jh7Hwv

Note: For inexperienced Linux user our software are tested with current version pipeline. We do not recommend upgrading the version on Linux Virtual machine. The webserver may crash when new software is updated. The webserver is in virtual environment which is isolated from your host.

- Download and install Virtualbox (VB) installation and VirtualBox 6.1.22 Oracle VM VirtualBox Extension Pack from https://www.virtualbox.org/wiki/Downloads. Already tested Virtualbox version 6.1.22 on 64 bit Ubuntu (18.04, 20.04), MacOs and Window 10.
- Download the webserver.ova image file from address above

After installing VB and its Extension Pack, open VirtualBox GUI and open File > Import Appliance to select webserver.ova downloaded file, then set up configuration based on your machine configuration. By default, our web server uses 4 cores CPU, 8 GB RAM. We recommend using 8 CPUs, 16 GB RAM, or more. A 30 GB partition for swap, which extends your virtual memory. This configuration keeps the Minimap2 program running in a low memory machine. However, hard disk read/write speed is much slower than RAM. So to speed up the program you should use higher memory. Hard disk data is dynamically allocated. Therefore when your data increases, the image file size also increases. We recommend deploying a VB image in the partition with at least 200 GB (depends on the number of users and data size, TB volume is highly recommended). Configure the network interface on your host site (your primary OS): Before we start the Virtual machine in the Virtual box configuration panel, we configure two network interfaces as in the figure below. The first network interface to the host machine via VirtualBox Host-Only Ethernet adapter and the second interface the internet via one of your host machine network interface. Network configuration is critical important for our webserver.



VM Network interface configuration

Figure 2: Network interface configuration

1.2.2. Login and configure webserver After booting up our guest OS, log in to your Virtual Machine (VM) with this default credential:

```
* user name: ag-rossi (default)
* password: 123456
```

Open the terminal, and we can get our web server IP address by this command on the guest terminal. The light configuration is for only the Human genome.

The program will download all reference genomes, genome annotation, reference transcriptome (cDNA) and other required packages. It also sets your IP address into the allowed IP list of the webserver then the IP address is printed out from the printout messages. The configuration step required internet connection, therefore you should configure webserver before field work.

```
$setup_webserver light
$runserver
```

If you want to use RNASeq for other organisms (Rat, Mouse, Zebrafish, C-elegans). These genome is much bigger than Human genome. Therefore, you have to extent your hard drive to at least 1 TB to use this command (beta version):

```
$setup_webserver full
$runserver
```

The webserver can access via three ways: the first way is your local network interface (normally start with 192.168.x.x), the second way your LAN IP address (depend on your LAN network for example 10.x.x.x), third way is directly on the Virtual machine (address: localhost or 127.0.0.1)