ShigaTyper User Tutorial

Contents

System Set up

ShigaTyper was developed on Ubuntu 16.04.3 LTS installed on VMware Player ver 14.1.1 in a Windows 7 and tested on Ubuntu 18.04.1 LTS on an Oracle VirtualBox ver 5.2.22 in a Windows 10 operating system.

1. If you have a Mac or Linux computer, you do not need to install VM ware or Ubuntu. Please skip directly to Installation of Anaconda.
2. If you have Windows 10, you can install an Ubuntu directly without VMware or VirtualBox, although ShigaTyper was not tested there. See a tutorial here for installation of Ubuntu on Windows 10: https://tutorials.ubuntu.com/tutorial/tutorial-ubuntu-on-windows#0
3. Ubuntu is only one of the free Linux distributions. Another popular Linux distribution, CentOS, is also frequently used for NGS analysis. ShigaTyper has not been tested on CentOS though.

Fomats

Input: ShigaTyper accepts Illumina paired-end reads in fastq.gz or fastq format. Please contact us if you would like to run single-end reads. ShigaTyper also accepts assembled genome in fasta format but this has not been validated. Moreover, it takes much longer to assemble a genome than running ShigaTyper directly with raw reads.

Output: Each sample will have a report in html format. A summary table is given in the jupyter notebook “RunNotebook_010819.ipynb”, which can be manually exported to an html report.

1.1. Installation of VMware Workstation Player (Administrator privilege required)

1.1.1. Go to the following website to download the latest version of VMware Workstation Player: https://my.vmware.com/en/web/vmware/free#desktop_end_user_computing/vmware_workstation_player/14_0

1.1.2. On the second panel in black, you can choose the version from the dropdown menu (the latest version is 15.0.0). Click on “VMware Workstation [version number] Player for Windows 64-bit Operating Systems” to download the software to your hard drive.

1.1.3. Double click the executable to install.

1.2. Installation of Oracle VirtualBox (Administrator privilege required, if you have installed VMware Workstation Player, skip to 2. Installation of Ubuntu).

1.2.1. Go to the following website to download Oracle VirtualBox for Windows hosts: https://www.virtualbox.org/wiki/Downloads#VirtualBox%20Downloads

1.2.2. Click the downloaded executable to launch the Setup Wizard. Here is a good tutorial to follow: https://itsfoss.com/install-linux-in-virtualbox/

1.2.3. Set the disk size at a minimal of **50 GB**.

2. Installation of Ubuntu Guest Addition

Installing Ubuntu on VMware Workstation Player (Administrator privilege required)

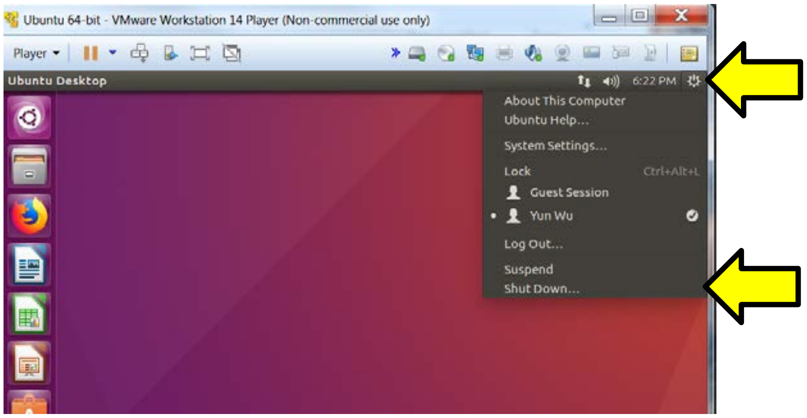
2.1. Download Ubuntu from the following hyperlink: https://www.ubuntu.com/download/desktop.

2.2. Follow procedures in the detailed instruction with images on how to install Ubuntu on VMware Workstation Player at the following link. Set the max disk size at a minimal of **50 GB**. https://websiteforstudents.com/how-to-install-ubuntu-16-04-17-10-18-04-on-vmware-workstation-guest-machines/

2.3. Follow the instruction all the way through and install VMware tool.

2.4. Allocate CPU and memory to Ubuntu guest addition, if the resources aren't enough.

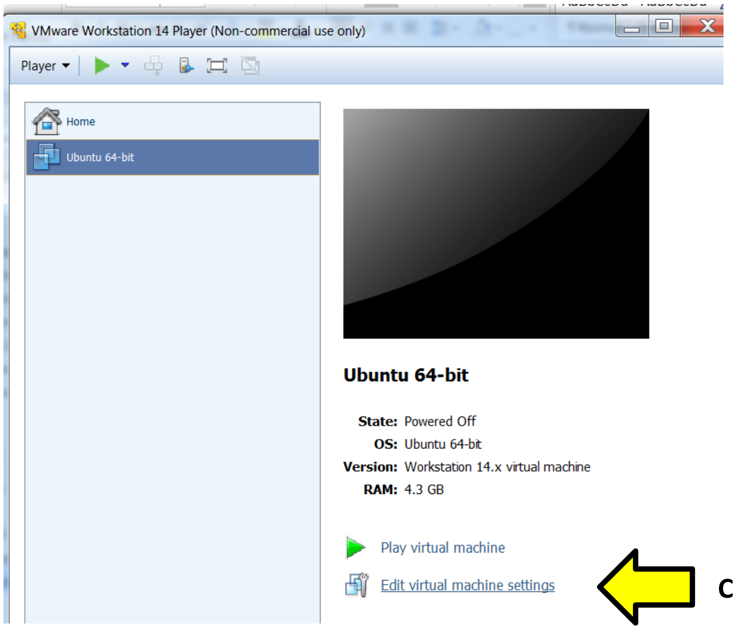
2.4.1. Shut down the Ubuntu VM by clicking the power button on the upper right corner of the Ubuntu Desktop then "Shut Down". Select "Shut Down".



First click here to show the drop down menu

Select here to shut down

2.4.2. In VMware Workstation Player, select the installed Ubuntu VM and click "Edit virtual machine settings".


VMware Workstation 14 Player (Non-commercial use only)
Player
Home
Ubuntu 64-bit
Play virtual machine
Edit virtual machine settings

Ubuntu 64-bit

State: Powered Off
OS: Ubuntu 64-bit
Version: Workstation 14.x virtual machine
RAM: 4.3 GB

Click here to edit VM settings

2.4.3. In the "Hardware" tab, click on "Memory" and move the slider bar on the right to allocate at least **4 GB**, if not more, memory to the VM (8 GB recommended).

2.4.4. In the "Hardware" tab, click on "Processor" and set from the drop-down window on the right. Allocate at least **4 CPU cores**, if not more, to the VM (8 cores recommended).

2.5. Set up a shared folder between Windows and Ubuntu.

2.5.1. Create a folder in your Windows file system to accommodate all the WGS files, if there isn't already one.

2.5.2. If your Ubuntu is on, shut down the Ubuntu VM by clicking the power button on the upper right corner of the Ubuntu Desktop. Select "Shut Down".

2.5.3. In VMware Workstation Player, select the Ubuntu VM and click "Edit virtual machine settings" (same as 2.4.2).

2.5.4. In the "Options" tab, click "Shared Folders" on the left pane.

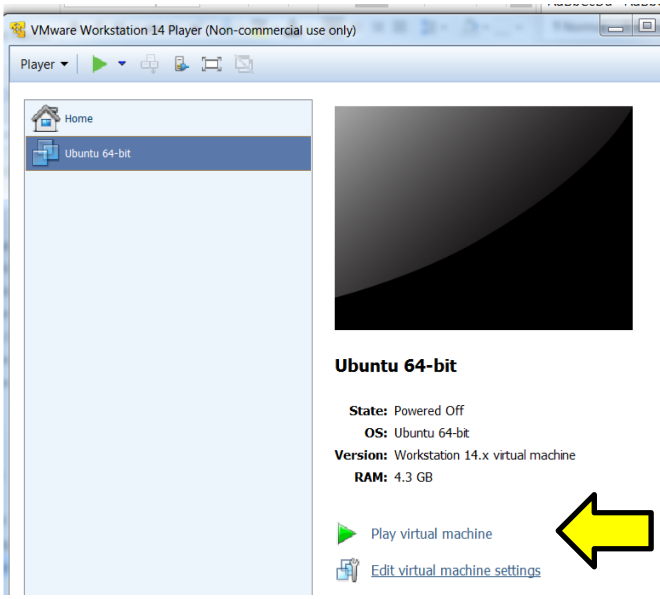
2.5.5. Click "Always enabled" on the right pane and click "Add" in "Folders" below.

2.5.6. The "Add Shared Folder Wizard" will pop up. Click "Next".

2.5.7. Select the folder to be shared (the folder created in 2.5.1) and give it a name.

2.5.8. Click "OK" to close the settings.

2.5.9. Select Ubuntu 64-bit VM and click on "Play virtual machine". Log onto Ubuntu.

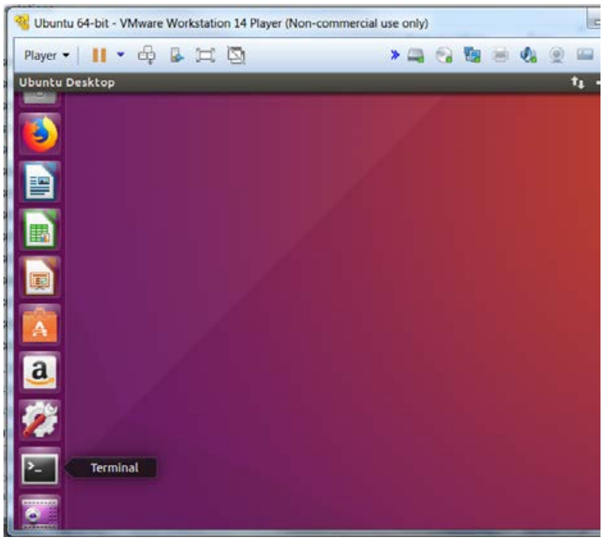

VMware Workstation 14 Player (Non-commercial use only)
Player
Home
Ubuntu 64-bit
Play virtual machine
Edit virtual machine settings

Ubuntu 64-bit

State: Powered Off
OS: Ubuntu 64-bit
Version: Workstation 14.x virtual machine
RAM: 4.3 GB

Click here to turn on Ubuntu

2.5.10. Locate "Terminal" from the panel on the left of Ubuntu Desktop. Click to open the command line window.


Ubuntu 64-bit - VMware Workstation 14 Player (Non-commercial use only)
Player
Ubuntu Desktop
Terminal

2.5.11. Enter the following command to see if the VM is aware of the existence of the shared folder (note: case-sensitive):

```
vmware-hgfsclient
```

This should return the name of the shared drive into the terminal window.

2.5.12. Run the VMware config tools by entering the following command in the terminal:

```
sudo vmware-config-tools.pl
```

2.5.13. Enter password following the prompt. Accept the default values.

2.5.14. [Optional] set up a short cut on your Ubuntu desktop by entering the following command:

```
ln -s /mnt/hgfs/[name of your shared drive] ~/Desktop/[name of your shared drive]
```

3. Installation of Anaconda with Python 3

3.1. Locate the latest Anaconda for Python 3 in the following link:
https://www.anaconda.com/download/#linux

3.1.1. Right click on the “Download” button and Select “Copy link address”

3.2. Logon to your Ubuntu VM and open “Terminal”.

3.3. Follow instruction exactly as described in the following link for Anaconda Installation:
https://www.digitalocean.com/community/tutorials/how-to-install-anaconda-on-ubuntu-18-04-quickstart

3.4. In step 2 of the tutorial, using “curl” to download anaconda, paste saved link address for anaconda that you copied in step 1, click on “Edit” then “Paste”. (If you don’t see “Edit”, hover your mouse over “Terminal” on the top of Ubuntu Desktop.)

3.4.1. Note that the tutorial described an earlier version of Anaconda so the file name is different from the one you will download. Use the filename that you just downloaded from step 4. Do not simply copy and paste from the webpage.

3.5. Skip step 9 unless you do want to create an environment.

4. Installation of ShigaTyper dependencies

First set up bioconda channels. In Terminal, type the following commands:

```
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
```

4.1 Minimap2

In Terminal, type the following:

```
conda install -c bioconda minimap2
```

4.2 fastp

In Terminal, type the following:

```
conda install -c bioconda fastp
```

4.3 htslib

In Terminal, type the following:

```
conda install -c bioconda htslib
```

4.4 samtools

In Terminal window, type the following:

```
conda install -c bioconda samtools
```

4.5 bcftools

In Terminal, type the following:

```
conda install -c bioconda bcftools
```

4.6 papermill

In Terminal window, type the following:

```
conda install -c conda-forge papermill
```

or (conda installation is recommended):

```
pip install papermill
```

4.7 version control

ShigaTyper was developed using fastp 0.12.2, minimap2 2.9, htslib/samtools/bcftools 1.7, papermill 0.12.5 and tested on fastp 0.19.5, minimap2 2.14, htslib/samtools/bcftools 1.9, and papermill 0.14.2. Identical interpretations were obtained from WGS data of 62 *Shigella* isolates.

5. Download Jupyter Notebooks and reference sequence database

5.1. There are 3 files to be downloaded from Github, 2 jupyter notebooks and 1 fasta file that is the reference sequence database. Name of the current versions are listed below:

5.1.1.batch_010819.ipynb

5.1.2.RunNotebooks_010819.ipynb

5.1.3.ShigellaRef6.fasta

5.2. Do not right-click the name of the file to save the ipynb files. The file saved this way will not run in Jupyter Notebook. Follow the steps below to properly download:

5.2.1.Click on the file name to reveal its formatted contents.

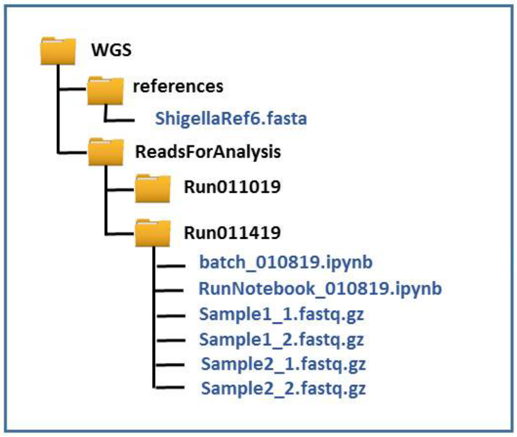
5.2.2.Click on the “Raw” button on the upper right corner, this will display the code (in json format).

5.2.3.Right click anywhere and select “Save as” or press “Ctrl+s” to save it as a jpynb file. (Sometimes Github will attempt to save it as a .txt file.) Make sure the names are exactly as shown above.

5.3. Create a folder under your shared drive called “references”. Place the reference sequence database “ShigellaRef6.fasta” in this folder.

5.4. Download the .jpynb files to wherever you can remember them. Copy and paste them to the folder of analysis when it’s time for analysis.

5.5. Recommended file organization structure as follows:



5.6. Folder name “references” and file names “ShigellaRef6.fasta” and “batch_010819.ipynb” have to be exactly as is and are case-sensitive. Names of folders in which the WGS reads are stored don’t matter as long as the WGS reads are in 2 layers of folders in the shared drive where the “references” folder is.

5.7. <u>Change file path/name of the reference sequence file</u>: the reference (ShigellaRef) is set at "../../references/ShigellaRef6.fasta" with respect of the working directory in in “batch_010819.ipynb” in the 4th code chunk in RunNotebook_010819.ipynb. Follow instruction there if a different directory or file is used.

6. Run ShigaTyper Analysis

6.1. Please see previous page for a diagram of file organization scheme.

6.1.1. If this is the first time you run ShigaTyper Analysis, create a folder under your shared drive. (Name of this folder doesn't matter. In the example figure, it is named "ReadsForAnalysis".)

6.1.2. In this folder, create another folder in which you place the paired end WGS files in fastq.gz format. (Names of this folder doesn't matter. In the example figure, it is named after the date of analysis "Run011419".)

6.1.3. Make sure WGS reads for each of the samples have a unique name before the symbol "_" (e.g., Sample1, Sample2).

6.1.4. Copy and paste the two jupyter notebooks into the same folder where all the fastq.gz files are.

6.2. Open terminal in your shared drive.

6.3. In VMware Workstation Player, if you have set up a short cut on the Ubuntu Desktop, right click the folder and Choose "Open in Terminal".

6.4. In VMware Workstation Player, if you don't have a short cut on the Ubuntu Desktop, Open "Terminal" and type:

```
cd /mnt/hgfs/[name of your shared drive]
```

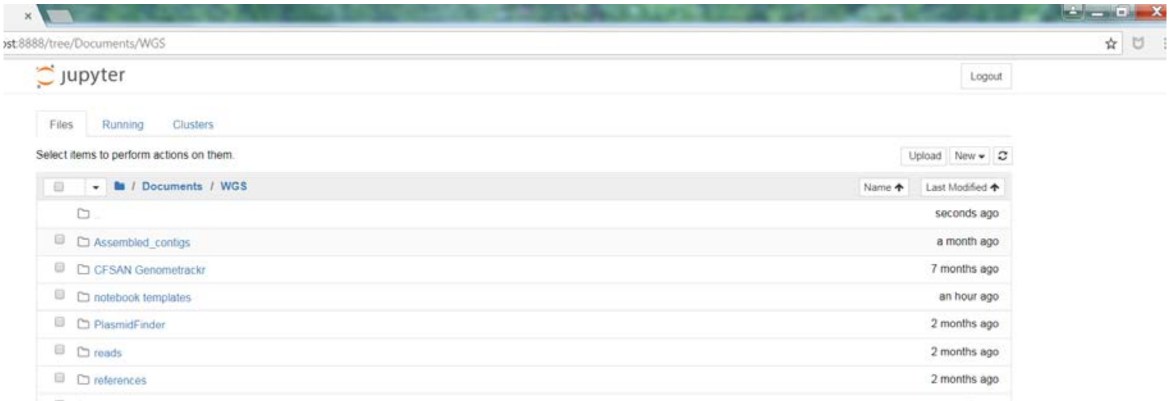
6.5. In Oracle VirtualBox, open Terminal and type:

```
sudo mount -t vboxsf [name of your shared drive] ~/Share/
```

6.6. Turn on Jupyter Notebook by typing in Terminal:

```
jupyter notebook
```

6.7. Jupyter Notebook will turn on in the default internet browser (Firefox in Ubuntu) like the figure shown below:



Select items to perform actions on them.

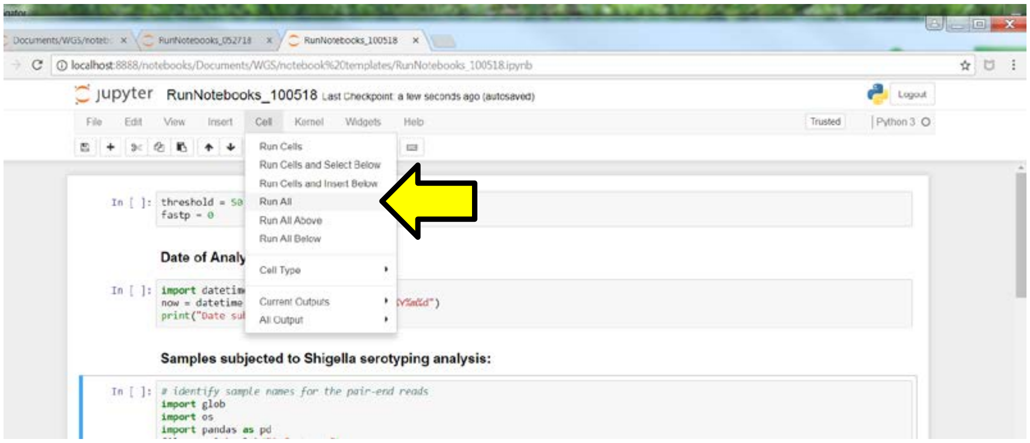
Name	Last Modified
..	seconds ago
Assembled_contigs	a month ago
CFSAN Genometrackr	7 months ago
notebook templates	an hour ago
PlasmidFinder	2 months ago
reads	2 months ago
references	2 months ago

6.8. Navigate to the folder that contains the jupyter notebooks and fastq.gz files to be analyzed.

6.9. Click on RunNotebooks_010819.ipynb to open the notebook in a new tab.

6.9.1. [Optional] If you would like to skip quality inspection, in the second line of the very first cell, change the code "fastp = 0" to "fastp = 1".

6.10. Click on "Cell" on the top then "Run All" to start the analysis as shown in the following figure (the file in the figure is an earlier version but you get the idea):


Documents/WGS/noteb
RunNotebooks_052718
RunNotebooks_100518
localhost:8888/notebooks/Documents/WGS/notebook%20templates/RunNotebooks_100518.ipynb
jupyter
RunNotebooks_100518 Last Checkpoint: a few seconds ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run Cells
Run Cells and Select Below
Run Cells and Insert Below
Run All
Run All Above
Run All Below
Cell Type
Current Outputs
All Output

```
In [ ]: threshold = 50
        fastp = 0
```

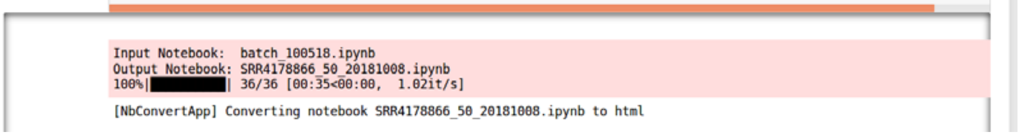
Date of Analy

```
In [ ]: import datetim
        now = datetime                    Y%m%d")
        print("Date sul
```

Samples subjected to Shigella serotyping analysis:

```
In [ ]: # identify sample names for the pair-end reads
        import glob
        import os
        import pandas as pd
```

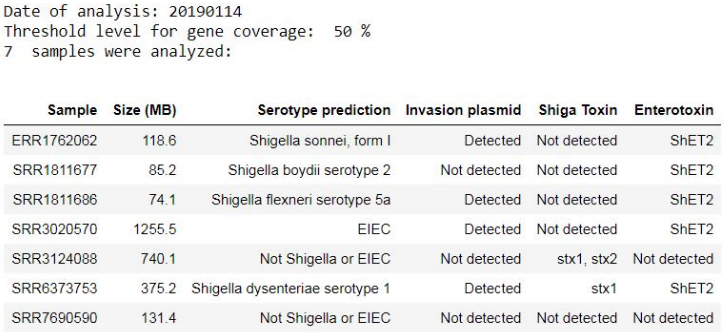
6.11. The notebook will automatically run to finish. If you have used this notebook to analyze samples before, results from the previous analysis will be erased and replaced by results from the new samples. You can see the progress for each of the sample:



```
Input Notebook:  batch_100518.ipynb
Output Notebook: SRR4178866_50_20181008.ipynb
100%|██████████| 36/36 [00:35<00:00,  1.02it/s]
[NbConvertApp] Converting notebook SRR4178866_50_20181008.ipynb to html
```

6.12. A html report will be automatically generated for each of the sample in the same directory.

6.13. When the analysis of all samples is complete, a summary table is generated at the end of RunNotebook_010819.ipynb as below:



```
Date of analysis: 20190114
Threshold level for gene coverage:  50 %
7  samples were analyzed:
```

Sample	Size (MB)	Serotype prediction	Invasion plasmid	Shiga Toxin	Enterotoxin
ERR1762062	118.6	Shigella sonnei, form I	Detected	Not detected	ShET2
SRR1811677	85.2	Shigella boydii serotype 2	Not detected	Not detected	ShET2
SRR1811686	74.1	Shigella flexneri serotype 5a	Detected	Not detected	ShET2
SRR3020570	1255.5	EIEC	Detected	Not detected	ShET2
SRR3124088	740.1	Not Shigella or EIEC	Not detected	stx1, stx2	Not detected
SRR6373753	375.2	Shigella dysenteriae serotype 1	Detected	stx1	ShET2
SRR7690590	131.4	Not Shigella or EIEC	Not detected	Not detected	Not detected

6.14. Click on "File" then "Save and Checkpoint" to save the results. You can change the filename of the notebook to more properly reflect the samples (for example, "ShigellaSamples_011419") by clicking on the title, "RunNotebooks_010819", on the top next to "jupyter".

6.15. To download a html version of the summary, click on "File", "Download as", then "HTML (.html)".

6.16. Shut down Jupyter Notebook.

6.16.1. Click on "File" then "Close and Halt" to close the notebook.

6.16.2. Click "Logout" on the upper right corner in the main page.

6.16.3. Quit browser.

6.16.4. Enter "Ctrl + c" in Terminal then type "y" upon prompt to quit Jupyter Notebook.

6.17. Click the power button on the upper righthand corner of the Ubuntu VM then select "Shut Down". Click "Shut Down" to shut down the VM.