Setting up a Command Line for the Oxford NCRM Summer School

Introduction

This is a quick guide to help set up efficient working environments to facilitate research in SocioGenomics (and other related fields). Importantly, it is to help enable access to both the productive Command Line Interface (CLI) and commonly used genomic software packages which would be otherwise be inaccessible to Windows users. Mac OS-X and Linux users are welcome to follow this guide in order to set-up a virtual environment which is identical to the one which will be used in Computer Class 1 of the NCRM Summer School: CentOS – a popular Linux distribution.[1] By using a popular, extremely accessible virtual machine such as VirtualBox, we are able to emulate another computer system within our regular environment without having to delete our regular operating system or re-partition it. Further motivation for using the CLI and *nix-like sytems (i.e. – Linux and Mac OS X) can be found in the accompanying lecture notes and will be briefly discussed in the class. Before we begin, let's define some important terms:

- **Host**: The host operating system is the physical computer on which the VirtualBox is installed. There are versions of VirtualBox for Windows, Mac OS X, Linux and Solaris hosts.
- **Guest**: This is the operating system that is running inside the virtual machine. Theoretically, VirtualBox can run any x86 operating system.
- **Virtual Machine (VM)**: This is the special environment that VirtualBox creates for your guest operating system while it is running.

Setting Up VirtualBox

The first thing to do is to head over to the VirtualBox download page:

https://www.virtualbox.org/wiki/Downloads

and download the appropriate binary for your host. We are going to proceed with the Windows host, as this is likely the most commonly required host in our SocioGenomic context. Click to install and move through the appropriate file-path settings and permission requests and start VirtualBox. Click on 'New' to create a new 'Guest' system, and type in CentOS. The 'Version' of Linux will default to Red Hat, as CentOS attempts to provide a free, enterprise-class, community-supported computing platform functionally compatible with its upstream source, Red Hat Enterprise Linux. You can click through the remainder of the settings accepting default parameters. The first option will ask you about the amount of memory which you wish to dedicate to the VM: a minimum of 1024mb is recommended, but the more that you can spare, the better. The next option will ask you to set up the hard disk. We suggest that you follow the default option of 'Create a virtual hard disk now' (and the default for type of storage and allocation). This will allow you to use this same hard disk on the virtual machine for the duration of the NCRM summer school. We suggest that you allow up to 8.00 GB of storage space in order be sure that you can have all the files required for the summer school stored on one VM at once. Finally, clicking 'Create' should create the VM, and give you a new VM in at the top left of the main VirtualBox window. The next step is to download a copy of the operating system to use as a Guest, as all we have done so far is set up the parameters for it. To do this, we need to download an '.iso' installation file. For CentOS, there are a couple of options found at:

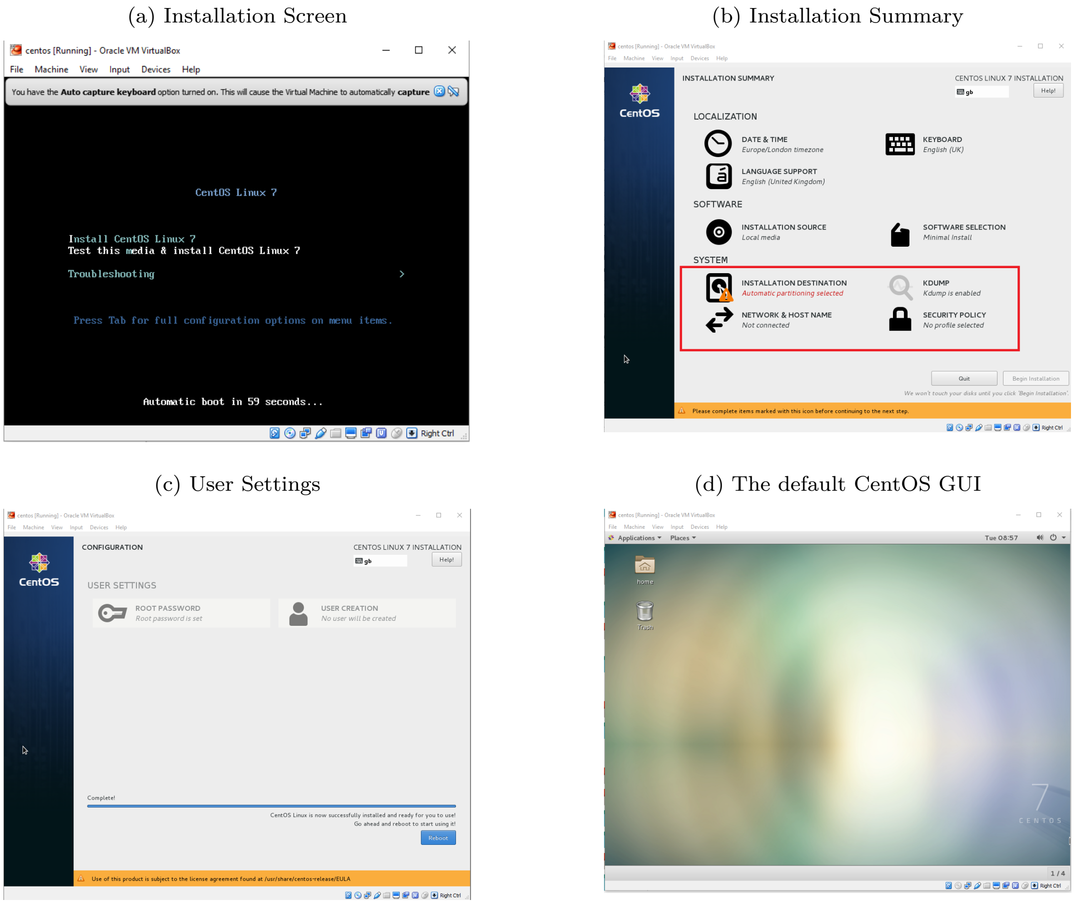
https://www.centos.org/download/

We strongly recommend the 'Everything ISO': which will give you a full Graphical User Interface.[2] However, this ISO is 7.7 GB, and will require a reasonable internet connection and some spare disk space. The alternative option is the 'Minimal ISO', which will only give you a command line prompt and limited functionality. To begin the installation, right click on the new VM which you recently configured, and then click on 'Start'. Your VM should open, asking you to locate the '.iso' file which you just downloaded. A handy hint: hitting the right CTRL button at any time will return you to your main Host operating system (and give you back control of your mouse and keyboard). You should click on the 'Installation Destination', and accept the default partitioning (8192 MB as per above). **Importantly, you need to choose the 'Software Selection'**

[1] CentOS is also one of the most commonly used operating systems on High-Performance Computers and currently runs on the majority of the Advanced Research Computing cluster at the University of Oxford and is also the operating system of choice for the Command Line Tools for Genomic Data Science course taught at John Hopkins University

[2] Note: using CentOS is by no means a binding requirement. Feel free to choose other more generalized, potentially accessible distributions such as Ubuntu or Mint.

Figure 1: Setting up CentOS



(a) Installation Screen

centos [Running] - Oracle VM VirtualBox

You have the **Auto capture keyboard** option turned on. This will cause the Virtual Machine to automatically **capture**

```
CentOS Linux 7

Install CentOS Linux 7
Test this media & install CentOS Linux 7

Troubleshooting                                  >

Press Tab for full configuration options on menu items.

Automatic boot in 59 seconds...
```

(b) Installation Summary

INSTALLATION SUMMARY

CENTOS LINUX 7 INSTALLATION

LOCALIZATION

DATE & TIME: *Europe/London timezone*

KEYBOARD: *English (UK)*

LANGUAGE SUPPORT: *English (United Kingdom)*

SOFTWARE

INSTALLATION SOURCE: *Local media*

SOFTWARE SELECTION: *Minimal Install*

SYSTEM

INSTALLATION DESTINATION: *Automatic partitioning selected*

KDUMP: *Kdump is enabled*

NETWORK & HOST NAME: *Not connected*

SECURITY POLICY: *No profile selected*

We won't touch your disks until you click 'Begin Installation'.

Please complete items marked with this icon before continuing to the next step.

(c) User Settings

CONFIGURATION

CENTOS LINUX 7 INSTALLATION

USER SETTINGS

ROOT PASSWORD: *Root password is set*

USER CREATION: *No user will be created*

Complete!

CentOS Linux is now successfully installed and ready for you to use!
Go ahead and reboot to start using it!

Use of this product is subject to the license agreement found at /usr/share/centos-release/EULA

(d) The default CentOS GUI

setting: and then GNOME Desktop. Similarily, You need to set up the Network Option to utilize the Wi-fi or ethernet of your host machine.[3] If you do not see a graphical interface or have nework connectivity after installing, a failure to set these options will be why. After setting up a root password, CentOS will then install and be ready to run in your VM.

Installing Python, R and PLINK

To install Python, we can use the `wget` command to download our favourite Python distribution (Anaconda):

```
wget https://repo.continuum.io/archive/Anaconda2-4.4.0-Linux-x86_64.sh
```

and then install using `bash Anaconda2-4.4.0-Linux-x86_64.sh`.[4] To check that this has worked correctly, we can check what version of Python we have installed with Python -V at the command line. Similarly, to install R, we need a couple more commands. We first need to install epel (Extra Packages for Enterprise Linux (EPEL)): `sudo yum install epel-release`, update all our packages with `sudo yum update`, and then reboot the machine: `sudo shutdown -r now`. Finally, install R with: `sudo yum install R -y`. To check information about this install, type `R --version` into the command line. Finally, to install plink, we can again use the `wget` command to download the necessary `.zip` file:

```
zzz.bwh.harvard.edu/plink/dist/plink-1.07-x86_65.zip
```

which we can then unzip with a simple `unzip plink-1.07-x86_65.zip`, `cd` into the folder, and then execute with `./plink`.

[3] The reason why this section is in in boldface is because **if you have any issues in setting up your VM, this is likely why!**

[4] See the corresponding lecture notes to better understand these and other commands.