Software installation guide for whole genome sequencing and transposon mapping

Contents

Introduction

This file discusses the installation procedures for the programs required for whole genome sequencing analysis and transposon mapping. The numbering corresponds with the order in which this software is used. Note that a part of the software is for windows and Linux, and a part is for Linux only. Therefore a virtual machine is included to run Ubuntu (or another unix based operating system to your liking). For a more detailed workflow and the commands of all the individual commands, see 'satay_analysis_notes.md'.

To run a program in the command line, you need to specify the entire path to the location where the program is installed. An easier way is to add the program to the windows path. This can be done with the command `setx PATH '/path/to/program'`. To check if the program is added to the path, use `echo $PATH$`.

Some programs requires java to be installed. To check if java is installed, run the command 'java -version' in the command line. If it is not installed, download and install it from <https://www.java.com/nl/download/>

0. VirtualBox-6.1.4

https://www.virtualbox.org/] https://ubuntu.com/download/desktop]

VirtualBox is used for running Linux based sofware on a Windows machine. For the operating system, recommended is to download Ubuntu 18.04 LTS 64-bit from https://ubuntu.com/#download, but any other Linux OS should be fine. Save the downloaded document at a convenient location on your computer. To install:

1. Run the VirtualBox-6.1.4-136177-Win application to install Virtualbox (VB) on your computer
2. Open VB -> click 'New' (blue icon)
3. Choose OS, storage location and name
4. Allocate RAM memory (adviced is to use the recommended amount to prevent problems)
5. Make new virtual hard drive (choose VHD)
6. Choose size of VHD (recommended is a minimal amount 50GB)
7. Go to settings of your new virtual machine (VM) (yellow gear symbol) and go to the storage tab.

8. Click 'empty' under 'Controller IDE'. Next to 'IDE Secondary master' click the blue CD icon and select 'choose/create a virtual optical disk'. Click 'Add' and choose your downloaded Linux OS.
9. Start the VM and run the Linux install process with the recommended settings. Note that this might freeze sometimes, so it might be necessary to reinstall it a few times)

The following steps are to add a shared folder between the Windows machine and the VM that allows for easier sharing of data (this is not obligatory). This requires to install 'guest additions' in the VM. If a folder is already shared and a new folder is wanted to be shared as well, perform only step 16, 17 and 18.

10. Open the terminal app
11. Enter `'sudo apt update'`
12. Enter `'sudo apt upgrade'`. Type 'Y' if the terminal asks if software needs to be installed. Restart the system
13. Enter `'sudo apt install build-essential dkms linux-headers-$(uname -r)'`
14. Click the 'devices' tab in the top of the window. Click 'Insert Guest Additions CD Image' and run the installer. Power off the VM when done.
15. In VB click 'settings' (yellow gear symbol) and go to the 'General' tab; 'Advanced'. Set both 'shared clipboard' and 'drag n drop' to bidirectional and click 'ok'
16. In VB click 'settings' again (yellow gear symbol) and go to the 'Shared Folders'. Add path to a folder that is going to be used as shared folder. Set 'Auto Mount' and click 'Ok'
17. Start the VM and check if the shared folder is present (either on the desktop, the files folder or in the media folder).
18. To be able to access the folder, special permission needs to be given. For this open the Terminal app.
19. Enter `'sudo adduser [username] vboxsf'` (where [username] should be replaced with your actual username of the VM).
20. To check, enter `'id [username]'`. This should give a list that needs to include 'vboxsf'
21. Restart the VM.

In order to run java based programs in Linux, Java needs to be installed. There are several ways of installing Java, but one of the simplest ways is using the command line with the following commands:

1. Enter `sudo -s` (enter password if requisted)
2. Enter `apt-get install openjdf-11-jre`. (If this version of java is not found, check which versions are available by entering `apt-get install openjdf` followed by a double tab. This should give a list of all available installations. It also gives you the opportunity to install the jdk (Java Developer Kit) instead of the jre (Java Runtime Environment), but if you don't plan to develop software in Java, the jre, what is used here, is good enough).
3. Restart the terminal. Check the installation by entering `java -version`

1. Fastqc (Windows)

[https://www.bioinformatics.babraham.ac.uk/projects/fastqc/]

Fastqc is used for quality checking. It is java based and therefore does not need to be installed. To run interactively, click the 'run-fastqc' batch file. To run non-interactively, enter in 'fastqc' in the command line.

2a. Trimmomatic-0.39 (Windows or Linux)

[http://www.usadellab.org/cms/?page=trimmomatic]

Trimmomatic is used for trimming fastq files. It is java based and therefore does not need to be installed. To run, enter 'trimmomatic-0.39' in the command line. The adapters folder contains some adapters that can be used during trimming if desired.

123Fastq-v1.1 (Windows) (optional)

[https://sourceforge.net/projects/project-123ngs/]

123Fastq is Fastqc and Trimmomatic combined in one interactive program. It is java based and therefore does not need to be installed. Click the 123fastq executable jar file to run the program.

2b. BBDuk-38.84 (Windows or Linux)

[https://jgi.doe.gov/data-and-tools/bbtools/]

BBDuk is an alternative for Trimmomatic for trimming of fastq files. It java based and therefore does not need to be installed. It is part of the bbtools packages (named the bbmap when downloaded). Once downloaded, unpack the .tar.gz package. Run the bbduk.sh executable in the bbmap directory. The adapter.fa file is included and located in the /resources directory.

3. BWA (Linux)

[http://bio-bwa.sourceforge.net/]

BWA is used for aligning the reads to a reference genome and to index the reference sequence. After downloading the software in the VM, install it by entering the following commands in the terminal:

```
bunzip2 bwa-0.7.17.tar.bz2
tar -xvf bwa-0.7.17.tar
cd bwa-0.7.17
sudo apt-get update
sudo apt-get install bwa
```

To run, enter `bwa` in the terminal.

4. SAMTools and bcftools (Linux)

[http://www.htslib.org/]

Samtools is used for processing after alignment, for example for converting SAM files to BAM files. After downloading the software in the VM, install it by entering the following commands in the terminal:

```
bunzip2 samtools-1.10.tar.bz2
tar -xvf samtools-1.10.tar
cd samtools-1.10
sudo apt-get update
sudo apt-get install samtools
```

To run, enter `samtools` in the terminal. Do the same protocol for bcftools.

5. Sambamba (Linux)

[https://lomereiter.github.io/sambamba/]

Sambamba is used for processing after alignment, for example for sorting and indexing the BAM files. After downloading the software in the VM, install it by entering the following commands in the terminal:

```
gunzip sambamba-0.7.1-linux-static.gz
chmod +x sambamba-0.7.1-linux-static
sudo ln -s /path/to/sambamba-0.7.1-linux-static /usr/local/bin
```

(where in the last line `/path/to/` needs to be replaced with the actual path.) To run, enter `sambamba` in the terminal.

6. IGV (Windows) (optional)

[https://software.broadinstitute.org/software/igv/]

IGV (Integrative Genomic Viewer) is used for visually check the results. Click the IGV_Win_2.8.0-installer and run the install process.

7. Matlab Transposon count (Windows)

[https://sites.google.com/site/satayusers/complete-protocol/bioinformatics-analysis]

This code relates the number of reads and transposon counts to the genes. This code is provided from the Kornmann-Lab.