Introduction to the Virtual Machine for SATAY analysis

Contents

Introduction

Welcome to the Ubuntu Virtual Machine that is created for processing of SATAY data. This file explains the layout of the virtual machine (VM) and how to perform the processing of raw sequencing data, specifically for SATAY data. All the required software tools are already preinstalled and ready to use. The installed software packages are

- FastQC (v0.11.5)
- BBMap (v38.84)
- Trimmomatic (v0.39)
- cmpfastq
- BWA (v0.7.17)
- SAMTools and BCFTools (v1.10)
- Sambamba (v0.7.1)
- java (v11.0.7)
- python code (transposonmapping_satay.py)

Also, the S288C reference sequence is stored.

If the VM is setup according to the installation guide, there should appear a shared folder on the Desktop with the name: *sf_VMSharedFolder_Ubuntu64_1*. This shared folder can be used for easy sharing of files between the VM and the host system (i.e. Windows). In the host system the files are located in the folder that was selected during the setup of the VM.

Running the workflow

When the virtual machine is setup as discussed in the Installation Guide, there is a shared folder located on the desktop. The workflow by default expects your data file to be present in this shared folder. So after putting the datafile in the shared folder of the Windows machine, you do not have to move the data file within the virtual machine.

The main operations are performed in the terminal app (located in the left bar on the screen). When you open this, the default location is ~/. In this location there is a workflow called *processing_workflow.sh* which can be ran the terminal using the command **bash processing_workflow.sh**. Before running the script, some settings might need to be changed in the workflow, which can be done by opening the file using the command **xdg-open processing_workflow.sh**, by changing the variables in the user settings block (for detailed explanation see the SATAY_Users_Notes). Most importantly is that the filename is set. It is possible to do the analysis for paired end reads, for which the variable `paired` has to be set to 't' (True) (else, set it to 'f' (False)). In this case either input two filenames (called filename1 and filename2) for the forward and reverse reads or enter a single file (filename1) that includes all reads. After the workflow is started, it gets the datafile from the shared folder and moves it to *~/Documents/data_processing/[datafolder]* (where *datafolder* can be set by the user in the user settings block

in *processing_workflow.sh*). In this location, three folders are generated where the output for the quality report, trimming and alignment are stored (for an overview of the folder structure, see the figure below). Also, a log file will be generated (saved together with the data file) that includes the name of the data file, a time stamp and the settings that were set in the user settings block. After the whole processing is completed, all the results including the data file and the log file are moved back to the shared folder.

The workflow starts with creating a quality report for the raw sequencing data. After this is complete, you will be asked if you want to continue. If you want the possibility to change the settings according to the outcome of the quality report, press no 'n'. The workflow stops and you can change the settings in the user settings section in the workflow. The workflow can be ran again and this time it will skip the quality report of the raw sequencing data (as long as the original quality report is not deleted). Press 'y' when asked if you want to continue. It then creates a new .fastq file with the trimmed data, apply a quality check on the trimmed data, aligns the data to a reference sequence, converts the resulting .sam file to its binary equivalent (.bam) and sorts and indexes this .bam file. Finally, all the results are moved back to the shared folder.

An important step during the trimming is the removal of adapter and barcode sequences. The sequences that need to be removed or checked needs to be placed in the adapters.fa file, which can be opened using the command **xdg-open ~/Documents/Software/BBMap/bbmap/resources/adapters.fa**. Enter each sequence by starting with a > symbol followed by a name of the sequence (this can be anything you want). The next line contains the literal sequence. Start the following sequence in the same way on the next line without any blank lines.

If you want to run all the software packages manually, please check the SATAY_Users_Notes.

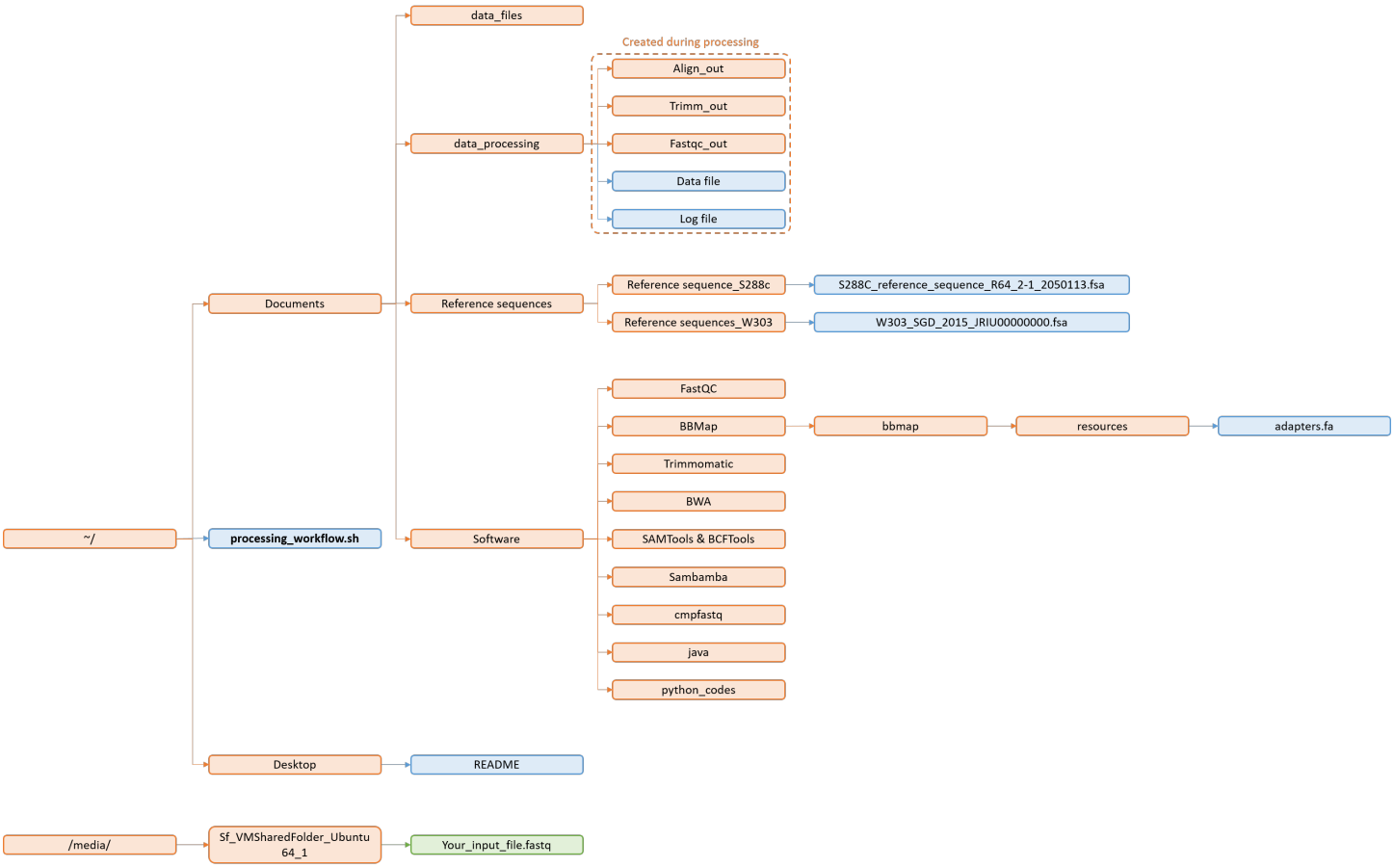
Virtual machine

The virtual machine runs on standard Ubuntu 64-bit. If you want to update the software, use the command line with the following commands:

- **sudo apt update**
- **sudo apt upgrade**

The password and user name are stored on the N-drive. When turning off, go to the top right corner and click on the sound and battery icon and power off the system.

The default folder structure is given in the following figure. Here, orange indicates directories and blue are some of the most important files. The green boxes indicate input files that you need to provide yourself.


data_files
Created during processing
Align_out
Trimm_out
data_processing
Fastqc_out
Data file
Log file
Documents
Reference sequences
Reference sequence_S288c
S288C_reference_sequence_R64_2-1_2050113.fsa
Reference sequences_W303
W303_SGD_2015_JRIU00000000.fsa
FastQC
BBMap
bbmap
resources
adapters.fa
Trimmomatic
BWA
~/
processing_workflow.sh
Software
SAMTools & BCFTools
Sambamba
cmpfastq
java
python_codes
Desktop
README
/media/
Sf_VMSharedFolder_Ubuntu 64_1
Your_input_file.fastq