README for Virtual Machine

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 experiments using the yeast Saccharomyces Cerevisiae. 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)

Also, two reference sequences are stored:

- S288C
- W303

If the VM is setup according to the installation guide, there should appear a shared folder on the Desktop. 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

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 ~/ (the home directory). 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 about all posible settings see the SATAY_Users_Notes). The workflow checks if the data file is present in the shared folder or at the location *~/Documents/data_processing/[datafolder]* (where *datafolder* can be set by the user in the user settings block in *processing_workflow.sh*). If the data file is present in the shared folder, it moves the data file to the above location. 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 including a copy of the adapters.fa file where the adapter sequences are stored. 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 you do not delete the original quality report in meantime). 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 (see the adapters.fa file for an example).

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.

If the names of the folders are changed, the paths in the workflow might not work properly anymore. If you change the paths (e.g. to the software or the reference genomes), make sure you change all the relevant paths in the workflow accordingly. The default workflow is setup to respect the folder structure as is defined in the figure below.

The default folder structure during processing is given in the following figure. Note that the subfolder in the data_processing directory might not be present before a workflow has started, but will automatically be generated when the workflow is running. The orange boxes are folders, the blue boxes indicate the most important files for the processing.