Wallaby Viral Assembly Pipeline

Version 0.3

Shawn.M.Rupp@asu.edu

Contents

Introduction

Wallaby is a series of scripts for automating the assembly and local alignment of viral DNA sequencing data.

Dependencies:
Python3
FastQC v0.11.3
Trimmomatic v0.36
ABySS 2.0.2

Kiwi v0.2 or above
MySQL
Cython
NCBI Blast+ 2.2.31+
usearch v10.0.240

Installation

FastQC
FastQC can be downloaded here. Download the zipped file into the /opt/ directory and unzip it.

Trimmomatic
Trimmomatic can be downloaded here. Download the zipped file into the /opt/ directory and unzip it.

ABySS
AbySS can be downloaded here. Follow their instructions to compile and install the program. Be sure to export the path to the compiled binaries.

Wallaby
Download the repository:

```
git clone https://github.com/icwells/Wallaby.git
```

Kiwi
If you wish to run BLAST or ublast, download Kiwi into the same parent directory as Wallaby (i.e. next to it) and follow the instructions in it’s ReadMe.

```
git clone https://github.com/icwells/Kiwi.git
```

Getting Started

Manifest File
Paired-end fastq reads are given to Wallaby in a simple tab-delimited text file. The first column contains the absolute path to each read (both forward and reverse). The second column contains the batch name: all reads with the same batch name will be assembled together and the final assembly will have that batch's name. The final column is the name of each individual sample. There should be exactly two files with a given sample name: the forward reads and the reverse reads.

```
path/to/forward/read.fastq.gz batch1    sample1
path/to/reverse/read.fastq.gz batch1    sample1
path/to/forward/read.fastq.gz batch2    sample2
path/to/reverse/read.fastq.gz batch2    sample2
```

In the event that each pair of reads is to be assembled together, then the batch and sample names can be the same. The fastq files may or may not be gzipped.

config.txt
Since Wallaby has many options, it uses a configuration file for some of the more basic options which are likely to stay the same between runs. These options include the number of threads, k (the length of k-mers for ABySS and default minimum length for Trimmomatic), trim settings for Trimmomatic (can use any valid Trimmomatic setting), the path to the database directory for running BLAST/ublast, and your MySQL username.

Any line starting with "#" will be ignored. When changing settings, be sure to leave a space between the equal sign and the setting (needless to say, don't change the setting name or the program will not recognize it).

When initially downloaded, it is named "config-REMOVE.txt" and it must be renamed to "config.txt". This is to prevent an edited configuration file from being overwritten if the "git pull" command is run. It must remain in the Wallaby directory and must be named "config.txt".

Running Wallaby

```
python wallaby.py –i path/to/manifest –o path/to/output/directory
```

Wallaby will run FastQC on all input files, read the results, and call Trimmomatic on any files that failed per base sequence quality. It will also print the name of any files which may still contain adapters. These steps can be skipped with the "--noqc" flag.

Next, Wallaby will call ABySS and make one assembly per input batch. It will then call `filterMetagenomicSequences.py` on the contig assemblies to sort contigs by descending length and identify putative circular sequences. It will write three output files: one for all contigs, one for circular contigs, and for all contigs with a minimum length greater than 250bp. The last file will contain circular contigs in both linear and circular formats in addition to the remaining linear contigs.

Lastly, Wallaby will perform local alignments using each of the minimum 250bp contig files against the appropriate database in the directory specified in the configuration file. The "--blast" flag will indicate that BLAST should be run; otherwise, ublast will be run since it is faster. A translated nucleotide search (i.e. blastx) will be run against the protein database. All hits with an e-value less than or equal to $1x10^{-5}$ will be subset and a nucleotide search (i.e. blastn) will be run against the nucleotide database. The output will then be concatenated with data from the MySQL database to provide more informative results. This step can be skipped with the "--noblast" flag and can be skipped to with the "--align" flag, allowing it to be run separately or resumed if it is interrupted.

Scripts

wallaby.py
Wraps all of the analyses programs. All functions for calling third party programs are contained in assemblyPipeline.py, but this script is only callable through wallaby.

```
python wallaby.py
  -h, --help  show this help message and exit
  -i I        Path to manifest file (formatted in three tab separated
               columns: paths to fastqs, batch name (all reads for a
               sample set), sample name (PE reads).
  -o O        Path to output directory. All output will be written
               here.
  --noqc      Skip FastQC and Trimmomatic.
  --noblast   Ublast/Blast will not be run on sorted contigs.
  --blast     Runs blast on sorted contigs (ublast is run by
               default).
  --align     Resume pipeline from blast/ublast.
```

filterMetagenomicSequences.py
This script will order contigs by length, update headers, and identify circular sequences.

```
python filterMetagenomicSequences.py
  -h, --help  show this help message and exit
  i           Path to input fasta file.
```