

## FluSeq Guide

Version 1.0 for Illumina Miseq system.

For Research Use only. Not for use in diagnostic procedures.

### General Description

FluSeq is a bioinformatics program developed and tested for an influenza A genome-wide amplicon-based high-throughput sequencing (HTS) method, using Illumina Miseq system [1]. It includes contig assembly, blastn, filters for contamination reads, and consensus sequence calling. Users are advised to refer the end-to-end laboratory protocol described by Lee, et al. 2016 [1]. The users are encouraged to modify the FluSeq program written in simple python scripts for other clinical virus amplicon-based HTS method.

### 1.0 INSTALLATION

Installation instructions are available for LINUX/UNIX or MacOSX. The entire analytic workflow was implemented and tested by the author on CentOS-6.6/RedHat Linux and MacOSX, but not on other operating systems. The operating system should contain an updated java.

#### Pre-requisite

1. Installation of Python 3.4.3 (<https://www.python.org/downloads/>)
  - Upon installation, log in as root from Terminal (Linux and MacOSX):

```
# pip install pandas
# pip install numpy
# pip install ZODB
# pip install lxml
# pip install xlrd
# pip install xlwt
# pip install beautifulsoup4
# pip install scipy
# pip install transaction
```
2. Installation of VICUNA-v1.3  
(<http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/vicuna>)
  - The vicuna\_config.txt used by this analytic workflow is stored in FluSeq Folder. Compilation of vicunAnalysis is not required during the installation.
  - path to the compiled vicuna executive file needs to be changed accordingly in the FluSeq-v1.0.py later, i.e. Line 100: First argument of the subprocess.Popen, to "YourPathInstalled/VICUNA\_v1.3/bin/vicuna-omp-v1.0"

- Optional: Line 104 - logging.info: Change program path to "YourPathInstalled/VICUNA\_v1.3/bin/vicuna-omp-v1.0".
  
- 3. Installation of BLAST 2.2.31+ software  
([https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=Download](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download))
  
- 4. Installation of BWA-0.6.2 software  
(<http://sourceforge.net/projects/bio-bwa/files/>)
  
- Users are advised to install BWA-0.6.2 software but not the latest version, as bwa-aln/sampe in this version was found to be more stable in reads alignment for this study.
  
- 5. Installation of samtools-1.2  
(<http://sourceforge.net/projects/samtools/files/samtools/1.2/>)
  
- 6. Installation of GATK-3.4-46 software  
(<https://www.broadinstitute.org/gatk/download/>)
  
- Path to the GenomeAnalysisTK.jar executive file needs to be changed accordingly in the FluSeq-v1.0.py later, i.e. Lines 233, 259, and 269: Third or Fourth argument of the subprocess.Popen, to "YourPathInstalled/GenomeAnalysisTK.jar".
  
- Optional: Lines 243, 265, and 278: logging.info: Change program path to "YourPathInstalled/GenomeAnalysisTK.jar".
  
- 7. Installation of picard-tools-1.138  
(<https://github.com/broadinstitute/picard/releases/tag/1.138>)
  
- Path to the GenomeAnalysisTK.jar executive file needs to be changed accordingly in the FluSeq-v1.0.py later, i.e. Line 146: Third argument of the subprocess.Popen, to "YourPathInstalled/picard-tools-1.138/picard.jar".
  
- Optional: Line 152: logging.info: Change program path to "YourPathInstalled/picard-tools-1.138/picard.jar".

## Procedure

1. Download the FluSeq.tar, decompress, and mv the FluSeq Folder to ~.  

```
$ tar -xvf FluSeq.tar
$ mv FluSeq ~
```
  
2. Change working directory to FluSeq  

```
$ cd ~/FluSeq
```

3. Execute FluSeq  
\$ ./FluSeq-v1.0.py InputFolder  
or  
\$ python3 FluSeq-v1.0.py InputFolder  
or  
\$ python3.4 FluSeq-v1.0.py InputFolder

## 2.0 FluSeq INPUT FILES and FOLDER

This section describes the input FluSeq-v1.0 requires.

### Folder and File Naming

The top-level input folder can be named in any combination alphanumeric characters. It is advisable that a folder can be capitalised at first letter for a better folders and files organization, eg. Folder20151231 (**Figure 1** – Folder in red).

The input folder should contain paired-end reads in fastq.gz format generated by Miseq Reporter or fastq2bcl2, eg. SampleID\_L001\_R1\_001.fastq.gz and SampleID\_L001\_R2\_001.fastq.gz (**Figure 1**). Also, it should contain the ResequencingRunStatistics.xml that can be copied from the top-level run folder named according to Miseq <ExperimentName>, eg. YYMMDD\_machinename\_experimentnumber\_flowcellnumber.

### 3.0 Contamination Learning

The Input folder for in-run contamination learning should contain paired-end reads in fastq.gz format generated by Miseq Reporter or fastq2bcl2, eg. SampleID\_L001\_R1\_001.fastq.gz and SampleID\_L001\_R2\_001.fastq.gz (**Figure 1** – Folder in orange). Also, it should contain the ResequencingRunStatistics.xml that can be copied from the top-level run folder named according to Miseq <ExperimentName>, eg. YYMMDD\_machinename\_experimentnumber\_flowcellnumber.

### Procedure

1. Change working directory to FluSeq  
\$ cd ~/FluSeq
2. Execute FluSeq  
\$ ./ms2ContLearning.py InputFolder 151206  
or  
\$ python3 ms2ContLearning.py InputFolder 151206  
or

```
$ python3.4 ms2ContLearning.py InputFolder 151206
```

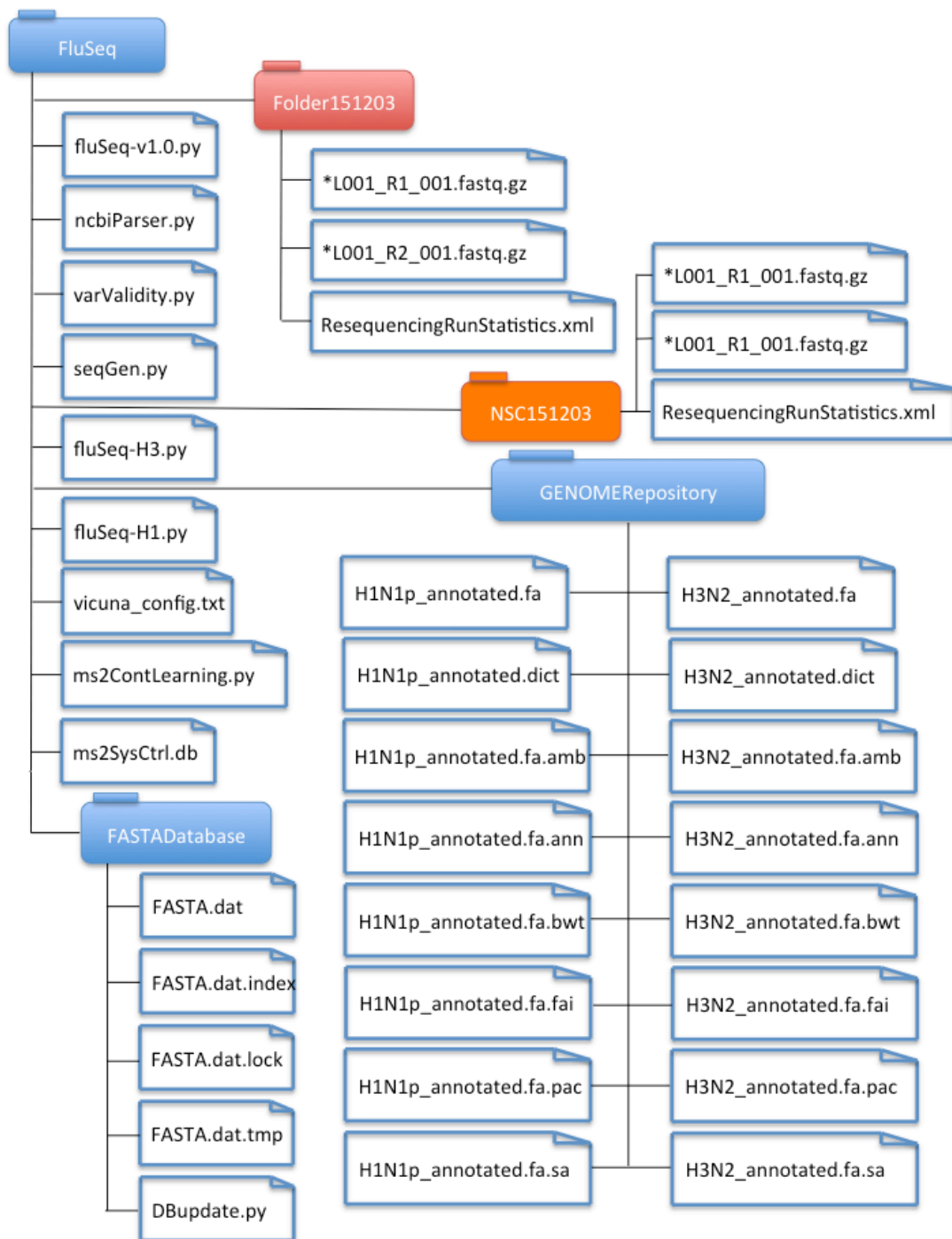
#### **4.0 Influenza Sequence Database**

This section describes the sequence database used for blastn. It contains all influenza sequences available in GenBank, according to the last update.

##### **Database Update**

1. Download all influenza sequences available from NCBI Influenza Virus Resource with a customized FASTA define as  
“>{accession}|{strain}|{segment}|{serotype}” and save the fasta file as FASTA.fa in FASTADatabase directory
2. Change working directory to FASTADatabase  
\$ cd ~/FluSeq/FASTADatabase
3. Execute DBupdate  
\$ ./ DBupdate.py FASTA.fa  
or  
\$ python3 DBupdate.py FASTA.fa  
or  
\$ python3.4 DBupdate.py FASTA.fa

**Figure 1.** Directory of FluSeq and input folder. Input folders for run analysis and contamination learning are colored in red and orange, respectively.



## Reference:

[1] Hong Kai Lee, Chun Kiat Lee, Julian Wei-Tze Tang, Tze Ping Loh, and Evelyn Siew-Chuan Koay. Pairwise comparison of contamination-controlled high-throughput and Sanger sequencing for influenza A genome sequencing. [Article in preparation]