Set up and run PneumoCaT

written by Georgia Kapatai – 30th March 2018

The following process assumes access to a terminal.

Step 1: Install Miniconda

The Miniconda is a smaller version of Anaconda, a python distribution, with installation and package management tools. It provides large selection of packages and commercial support. It is an environment manager, which provides the facility to create different python environments, each with their own settings. Miniconda gives you access to conda package manager and Python. Once Miniconda is installed, you can use the conda command to install any other packages and create environments. To install Miniconda open a terminal window and **Copy & Paste the following** commands into the terminal:

```
wget -c http://repo.continuum.io/miniconda/Miniconda-latest-Linux-x86_64.sh
bash Miniconda-latest-Linux-x86_64.sh
```

The `wget` command will download the miniconda installation file from its website and then the second line initiates installation. Once the installation was initiated you will first be asked to review the license. **Hit ENTER** and continue pressing enter until you reach the end of the license agreement. Then type yes and press ENTER. You will then be asked about the location of the installation; default is in your home directory so if you are ok with that **hit ENTER** or else specify a different location. I went for default. Once it finishes the installation you are asked whether to add the PATH in your .bashrc file. By adding the PATH in your .bashrc file you are telling your computer to look for software in this PATH so saying yes would make your life much easier. You have now installed miniconda. To start using conda commands **close this terminal and open a new one**.

Step 2: Install Git

Git is an open source version control system used by many developers to store or share their code in data structures called a repository. A Git repository can be access either through the web (www.github.com) or via the command line. For the latter option you need to install `git,` a set of commands that allows you to clone repositories from the web, upload/download new versions of the repository and many more functions. PneumoCaT has its own Git repository, therefore to run PneumoCaT we need to clone this repository, which you can do by using the `git clone` command from the command line. There is the option to just download the repository by going to the PneumoCaT website click on Clone or Download button and click Download ZIP. However, if you download the repo

this way then the connection to the web version of the repo will be severed. The advantage of using the `git clone` command is that it maintains the connection to the web server which means if an updated version of the code is release you can very simply update your local version by typing `git pull` on the command line. To test whether git is installed in your computer type `git` on the command line. If not then follow the instructions below:

Installing Git on Linux

Open a terminal window. **Copy & paste the following** into the terminal window and **hit ENTER**. You may be prompted to enter your password.

```
sudo apt-get update
sudo apt-get upgrade
sudo apt-get install git
```

You can use Git now.

Installing Git on Mac

Open a terminal window. Before installing Git you need to install Homebrew. Homebrew is a package management system that simplifies installation of software on Apple's macOS operating system using the command line.

Copy & paste the following into the terminal window and **hit ENTER**.

```
ruby -e "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/master/install)" brew
doctor
```

You will be offered to install the *Command Line Developer Tools* from *Apple*. **Confirm by clicking** *Install*. After the installation finished, continue installing *Homebrew* by **hitting ENTER** again.

Once HomeBrew is **copy & paste the following** into the terminal window and hit ENTER.

```
brew install git
```

You can use Git now.

Step 3: Clone the PneumoCaT repository

To clone the PneumoCaT Git repository in your local environment **copy & paste the following** into the terminal window.

```
git clone https://github.com/phe-bioinformatics/PneumoCaT.git
```

Clicking on the url in the above command will take you to the website for PneumoCaT. That is the PHE release of PneumoCaT that contains instructions on how to use and it the dependencies required to run this tool.

Alternatively, you can get this link by searching for PneumoCaT in Google and clicking the link of the first result. You then click the Clone or download button located on the right hand side of the website and copy the link in the box that appears. Finally type `git clone` in your terminal and paste this link in your line.

Step 4: Install PneumoCaT dependencies

Now that you have a local copy of the PneumoCaT repository we need to make sure all dependencies mentioned in https://github.com/phe-bioinformatics/PneumoCaT are installed. This is where conda can make things very easy. Instead of having to go to the website for each software, download the source file and follow the associated instructions to install including installing the dependencies for each of the software, we can instead use `conda install` to install all the required dependencies and their associated dependencies. First though we need to set up certain channels in our conda, which contain some of our dependencies, not found in the default channel. That is the bioconda channel as well as the other channels bioconda depends on. **It is important to add them in this order** so that the priority is set correctly (that is, bioconda is highest priority).

The *conda-forge* channel contains many general-purpose packages not already found in the *defaults* channel.

```
conda config --add channels defaults

conda config --add channels conda-forge

conda config --add channels bioconda
```

Now that the packages are install we can continue with installing the dependencies.

```
conda install pysam numpy lxml

conda install –c bioconda bowtie2

conda install –c bioconda biopython
```

And you now have everything you need to run PneumoCaT.

Step 5: Running PneumoCaT

Now that all the dependencies are all installed you can run PneumoCaT on one of the examples within the PneumoCaT directory to make sure everything is in order. To do that copy and paste the following command:

```
cd  PneumoCaT
```

```
python PneumoCaT.py –i Examples/PHESPV0253
```

Running this command will generate a result directory, pneumo_capsular_typing, within Examples/PHESPV0253.

```
georgia@georgia-VirtualBox:~/PneumoCaT$ ls -l Examples/PHESPV0253/
total 164328
-rw-rw-r-- 1 georgia georgia 82148246 Mar 30 21:01 PHESPV0253.R1.fastq.gz
-rw-rw-r-- 1 georgia georgia 86110945 Mar 30 21:01 PHESPV0253.R2.fastq.gz
-rw-rw-r-- 1 georgia georgia     1169 Mar 30 21:01 PHESPV0253.results.xml
drwxrwxr-x 4 georgia georgia     4096 Mar 30 22:28 pneumo_capsular_typing
```

As you can see the Examples/PHESPV0253 directory contains the two fastq files required to run PneumoCaT and the expected result file (PHESPV0253.results.xml).

```
georgia@georgia-VirtualBox:~/PneumoCaT$ ls -l Examples/PHESPV0253/pneumo_capsula
r_typing/
total 193236
-rw-rw-r-- 1 georgia georgia        0 Mar 30 22:28 ComponentComplete.txt
-rw-rw-r-- 1 georgia georgia     2944 Mar 30 22:25 coverage_summary.txt
drwxrwxr-x 2 georgia georgia     4096 Mar 30 22:16 logs
-rw-rw-r-- 1 georgia georgia      503 Mar 30 22:25 PHESPV0253.results.xml
-rw-rw-r-- 1 georgia georgia 197843014 Mar 30 22:24 PHESPV0253-sorted.bam
-rw-rw-r-- 1 georgia georgia     9424 Mar 30 22:25 PHESPV0253-sorted.bam.bai
drwxrwxr-x 2 georgia georgia     4096 Mar 30 22:28 SNP_based_serotyping
```

Within the pneumo_capsular_typing directory you can see another PHESPV0253.results.xml file that corresponds to the results of the first step of the PneumoCaT process; that is mapping the reads to all the serotype capsular loci sequences. The presence of a SNP_based_serotyping directory indicates that the process did not stop at the first step, therefore we are dealing with a genogroup. The final output of the PneumoCaT process will be in the PHESPV0253.results.xml file within this SNP_based_serotyping directory.

```
georgia@georgia-VirtualBox:~/PneumoCaT$ ls -l Examples/PHESPV0253/pneumo_capsula
r_typing/SNP_based_serotyping/
total 161644
-rw-rw-r-- 1 georgia georgia     1169 Mar 30 22:28 PHESPV0253.results.xml
-rw-rw-r-- 1 georgia georgia 165506764 Mar 30 22:27 PHESPV0253.sorted.bam
-rw-rw-r-- 1 georgia georgia      256 Mar 30 22:27 PHESPV0253.sorted.bam.bai
-rw-rw-r-- 1 georgia georgia      703 Mar 30 22:28 variant_summary.yml
```

You can view the result file by running the following command:

```
less Examples/PHESPV0253/pneumo_capsular_typing/SNP_based_serotyping/PHESPV0253.results.xml
```

The predicted serotype can be seen on the 5th line of this file:

```
<ngs_sample id="PHESPV0253">
  <workflow value="PneumoCaT" version="1.0"/>
  <results>
    <!--(START) Serotype Distinction Results (START)-->
    <result type="Serotype_Distinction" value="07B"/>
    <result type="Serotype_Distinction_Total_Hits" value="9/9"/>
    <result type="Serotype_Distinction_Serotypes_Tested" value="07B,07C,40"/>
```

This should match the predicted serotype in the expected result file mentioned earlier.

And you have now successfully run you first pneumococcal genomic sample using the PneumoCaT tool. For further details on interpretation of the output see the InterpretingResults.pdf document.