|  |
| --- |
| ***Insert Laboratory Specific Name Here*** |
| **SanitizeME: Host DNA Removal Standard Operating Procedure** |

1. **Purpose**

This procedure outlines the steps for using SanitizeMe (MINIMAP2 and SAMTOOLS) to remove host contaminant DNA from an input file (fq,fq.gz,fastq,fastq.gz).

1. **Scope**

This document applies to all staff that perform filtering, cleaning and preprocessing of DNA sequence files prior to other downstream bioinformatic analyses of these files. This process removes reads from a fastq file which map to another reference file (in this case, for the removal of human host contaminants from DNA sequencing).

1. **Related Documents**

|  |  |
| --- | --- |
| **Title** | **Document Control Number** |
| Data Retention and Storage Guidance |  |
| *“Lab-developed Risk Assessment/Mitigation document”* |  |

1. **Responsibilities**

| **Position** | **Responsibility** |
| --- | --- |
| Bioinformatician or Analytical Personnel | * Install the required dependencies mentioned in this document and SanitizeMe * Utilize the procedure to perform host contaminant filtering with your sample fastq files of interest |

1. **Equipment/Materials**
   1. The GUI version of SanitizeMe will require direct access to a Linux PC or x11 forwarding (e.g., VcXsrv). Additionally, SanitizeMe will require environmental dependencies (Minimap2, SAMtools, Python 3, and Gooey/colored modules) that can be installed using Miniconda/Anaconda (see procedure for installation of dependencies).
2. **Installation Procedure**
   1. If Git is not installed, install Git for Debian systems using the following commands:

**sudo apt update**

**sudo apt install git**

* 1. Clone the code from the repository (repo) to your home directory (or desired location):

**git clone https://github.com/jiangweiyao/SanitizeMe**

* 1. If Conda is not installed on the system, install Miniconda by running the prepackaged script install\_miniconda.sh to install into your home directory (if Git was used to clone to a different directory please reference the correct path in your command) by using the following command:

**. ~/SanitizeMe/install\_miniconda.sh**

* 1. Install Miniconda, Download human\_g1k\_v37, and build the conda environment using the single command below. (If you used git to clone to a different directory, please reference the correct path in your command.):

**. ~/SanitizeMe/install\_all.sh**

1. **Sample Usage Procedure**
   1. Activating the environment makes the software you installed in that environment available for use. You will see “(SanitizeMe)” in front of bash after activation. Activate with the following command:

**conda activate SanitizeMe**

* 1. Run the GUI version with the following command. Select the path to the folder containing your desired sequence files to be filtered in the InputFolder field. Specify the human reference file you would like to use for the mapping step of the process, note that the human\_g1k\_v37 reference file is downloaded by the install\_all.sh file and available for use if a custom human reference is not needed for your process. Next, input the path to the directory you would like the process output to be stored. Increase or decrease the number of threads to use as well as select your long read technology. Then, hit Start to run the program.

**SanitizeMe\_GUI.py**

* 1. GUI Usage Shown Below:

Screenshot:

Sanitize Me Paired GUI
Input Folder
Reference
Output Folder
Threads


* 1. Run the command-line interface (CLI) version with the test files using the following command, assuming the repo was cloned into your home directory and the reference file is in the repo folder. Alternatively, when running this tool with your particular data set, specify a folder containing your fastq files that require filtering after the -i parameter. If you would prefer to use a different reference sequence, download the file if needed and specify the path to the alternate reference file after the -r parameter instead. Change the command below as appropriate. The -i parameter is the directory path containing input fastq files, the -r parameter is the file path for the preferred reference sequence, and the -o parameter is the directory path where you would like SanitizeMe to place output files.

**SanitizeMe\_CLI.py -i ~/SanitizeMe/test/ -r ~/SanitizeMe/human\_g1k\_v37.fasta.gz -o ~/dehost\_output/test\_CLI**

* 1. Get help for the CLI version with the following command:

**SanitizeMe\_CLI.py -h**

* 1. When you are finished running the workflow, exit out of your environment by running conda deactivate. Deactivating your environment exits out of your current environment and protects it from being modified by other programs. You can build as many environments as you want and enter and exit out of them. Each environment is separate from each other to prevent version or dependency clashes. The author recommends using conda/Bioconda to manage your dependencies.

1. **References**
2. Yao, J., 2020. *Cdcgov/Sanitizeme*. [online] GitHub. Available at: <https://github.com/CDCgov/SanitizeMe> [Accessed 11 August 2020].
3. **Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rev #** | **DCR #** | **Changes Made to Document** | **Date** |
|  |  |  |  |

1. **Approval**

Approved By: Date:

Author

Print Name and Title

Approved By: Date:

Technical Reviewer

Print Name and Title

Approved By: Date:

Quality Manager / Designee

Print Name and Title