# **IPD (Infection Pathogen Detector)**

IPD is an in-silico GUI-based automated pathogen analysis pipeline for seamless analysis of data from heterogenous NGS platforms. IPD performs integrated variants analysis, along with systematic quantification of pathogen genomes. IPD additionally has an in-built SARS-CoV-2 analysis module, for assignment of viral clades of the samples analyzed and an automated report generation.

## **Prerequisites required for installation of IPD**

There are two automated interfaces for the tool either of them can be used by the user. It is developed using python3.

System Prerequisites:

* Pip3 (<https://pip.pypa.io/en/stable/installing/>)
* Python3 (<https://realpython.com/installing-python/>)
* Conda (<https://docs.conda.io/projects/conda/en/latest/user-guide/install/>)

With the following system prerequisites in place, users can use the following commands in the Linux / Unix environment to install the required python packages.

Python Packages:

* Pysam (<https://pypi.org/project/pysam/>): pip install pysam
* Tkinter (For only GUI: https://tkdocs.com/tutorial/install.html)

Packages Required for automated report generation:

* Numpy (<https://pypi.org/project/numpy/>): pip install numpy
* Matplotlib (<https://pypi.org/project/matplotlib/>): pip install matplotlib
* Pandas (<https://pypi.org/project/pandas/>): pip install pandas
* SciPy (<https://pypi.org/project/scipy/>): pip install scipy

**The installation of the IPD has been tested on Fedora/Red Hat OS**

IPD can be downloaded from <http://www.actrec.gov.in/pi-webpages/AmitDutt/IPD/IPD.html>. The downloaded tar file (ipd.tar.gz) should be untar using the following command.

$tar xvzf ipd.tar.gz

## Installation

Install script helps the user to sets up the tools, download and index the required databases (primary and secondary). On the terminal the user needs to go inside the IPD directory and then run the install command as shown below:

$ cd ipd

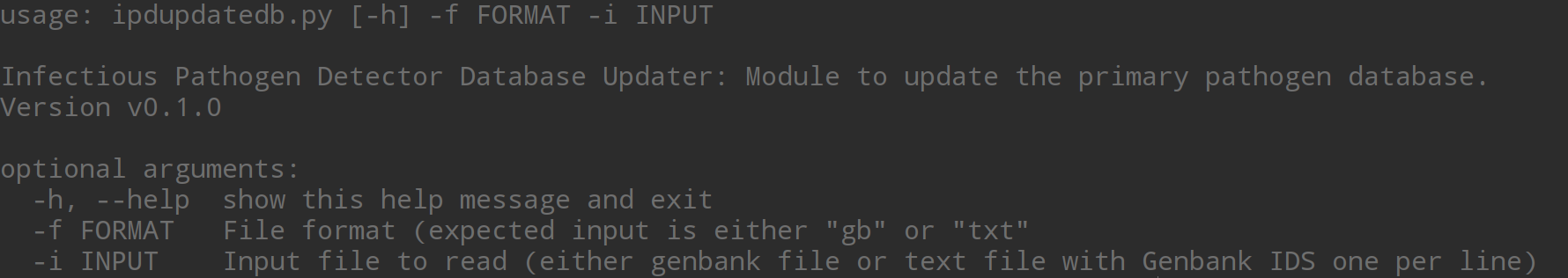
$ bash install.sh

Adding pathogens of choice to IPD database

We understand that users of IPD may want to include additional pathogen genomes, to perform the quantification and variant calling from the NGS data. Hence, we provide an additional script ("ipdupdatedb.py") along with the core IPD package. To update the current primary pathogen database, a user is required to either download and give the pathogen genome as a flat file in the GenBank format (single or multiple genomes are accepted) or provide a list of GenBank IDS in a text file (one ID per line), which would be downloaded automatically and the database is updated accordingly. The usage of the update script is described in the snapshot below.

Command:

$ python3 ipdupdatedb.py -f <format> -i <input file>



## Running IPD

The required scripts to run IPD are placed in the “src” directory of IPD. IPD can be run in two modes; command line (ipd\_cli.py) or the GUI mode (ipd\_gui.py). By going into the */src* directory of *ipd,* users can run the program as shown below:

For Command Line interface: (Usage of the command line interface is further described in Section 2.0 of the manual)

$ python3 ipd\_cli.py

For GUI: (Usage of IPD GUI is further described in Section 3.0 of the manual)

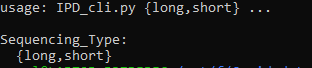
$ python3 ipd\_gui.py

1. Guide to use command-line interface of IPD

Scripts are present in the src folder. src should be used as the execution directory, output directory will be used as the working directory. The screenshot below explains the options available in the interface and its usage. There are two modes in the IPD command-line interface based on the sequencing type selected (long read sequencing and short read sequencing)

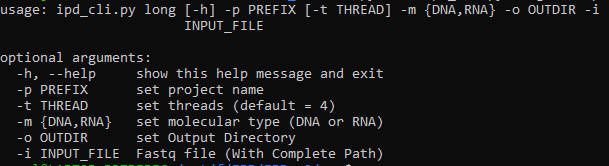
Command:

$ python3 IPD\_cli.py {long, short}



Single-end long read data is taken as an input. Prefix is set as the project name which is used as the prefix for all the output files.

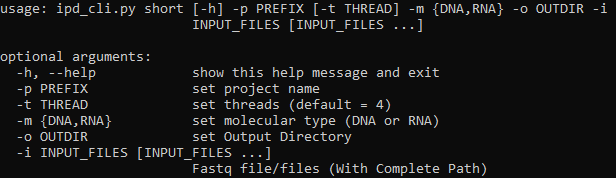
Command:

$ python3 ipd\_cli.py long -p <Project Name> -t <number of threads default 4> -m <DNA/RNA> -o <output directory> -i <input fastq/fastq.gz file with path>

Paired-end and single-end data can be given as input in case of short read sequenced input files. Files should be given with complete path. In case of paired end data, files should be given with a space in between.

Command:

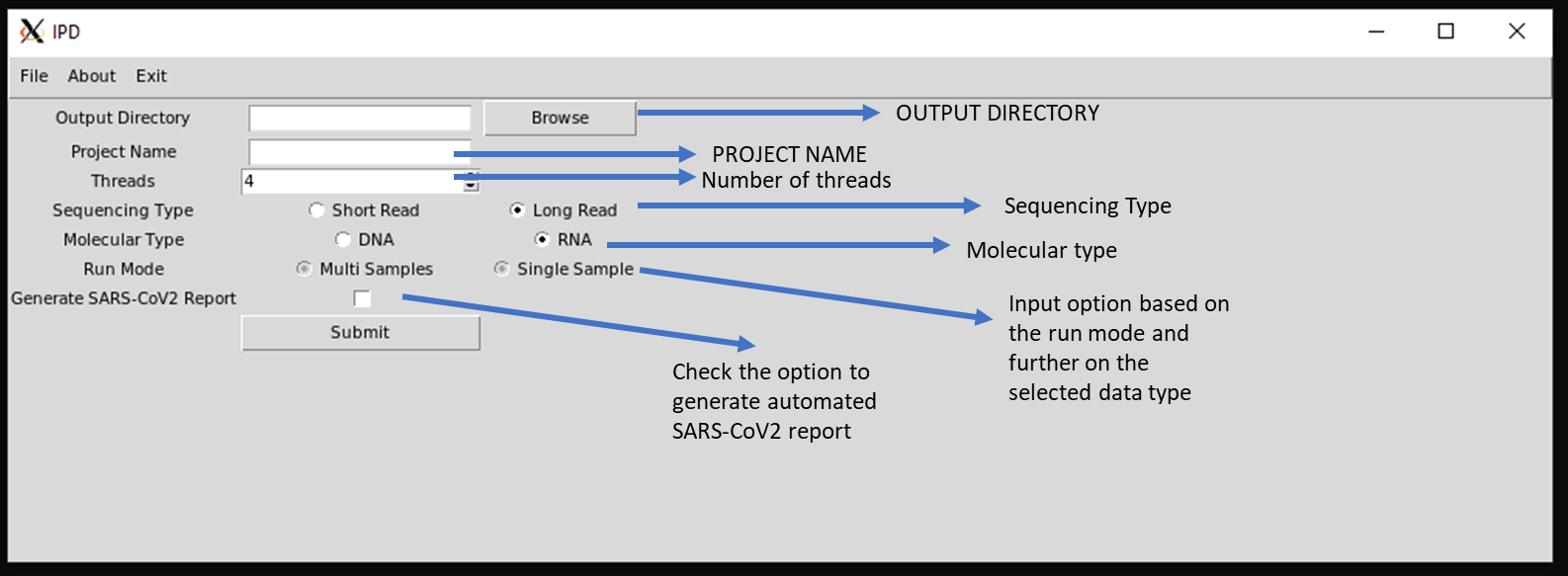
$ python3 ipd\_cli.py short -p <project name> -t <number of threads default 4> -o <output directory> -m <DNA/RNA> -i <input fastq or fastq.gz file/files>



1. Guide to use GUI of IPD

IPD graphical user interface is developed for the analysis of both long and short read to detect the abundance of pathogen and variants present in them. GUI code is kept in src, following command is used to access the same.

$python3 ipd\_gui.py



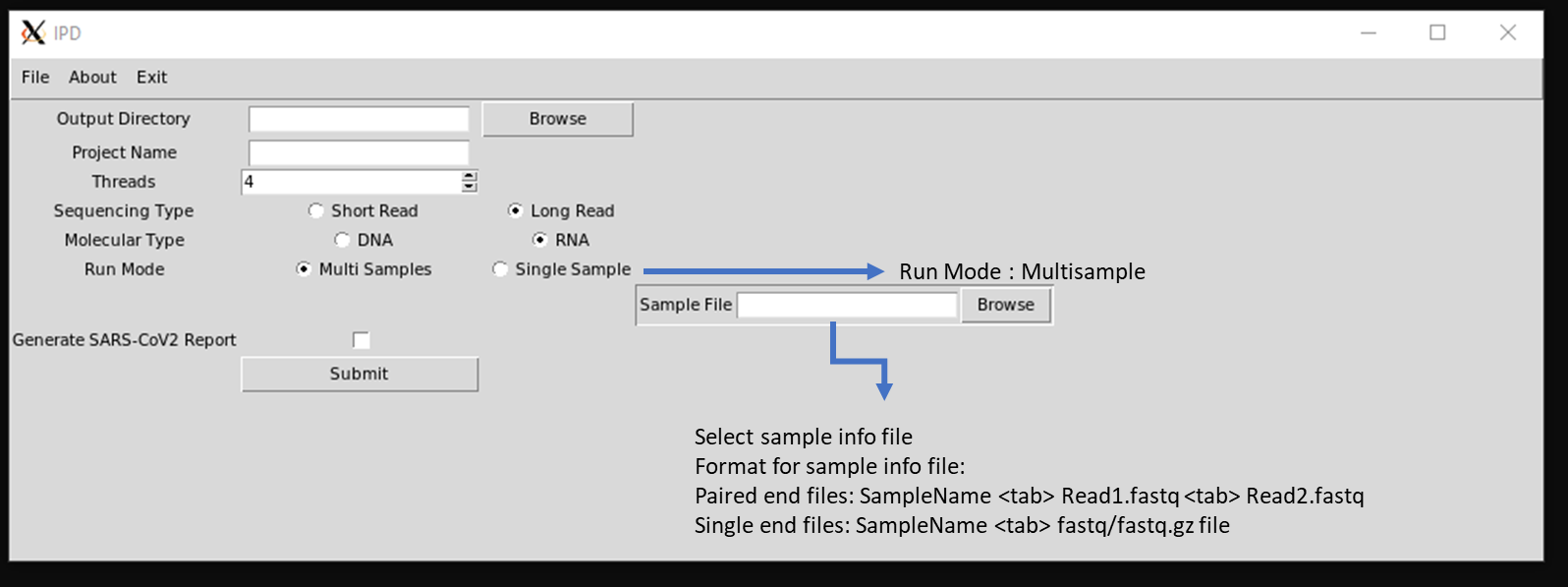
It has both multi-sample and single sample run mode. For Multi-sample run mode, user need to provide a sample info file. Sample name and Project name will be used as prefix for all the output files.

**Example Sample File format for paired-end data:**

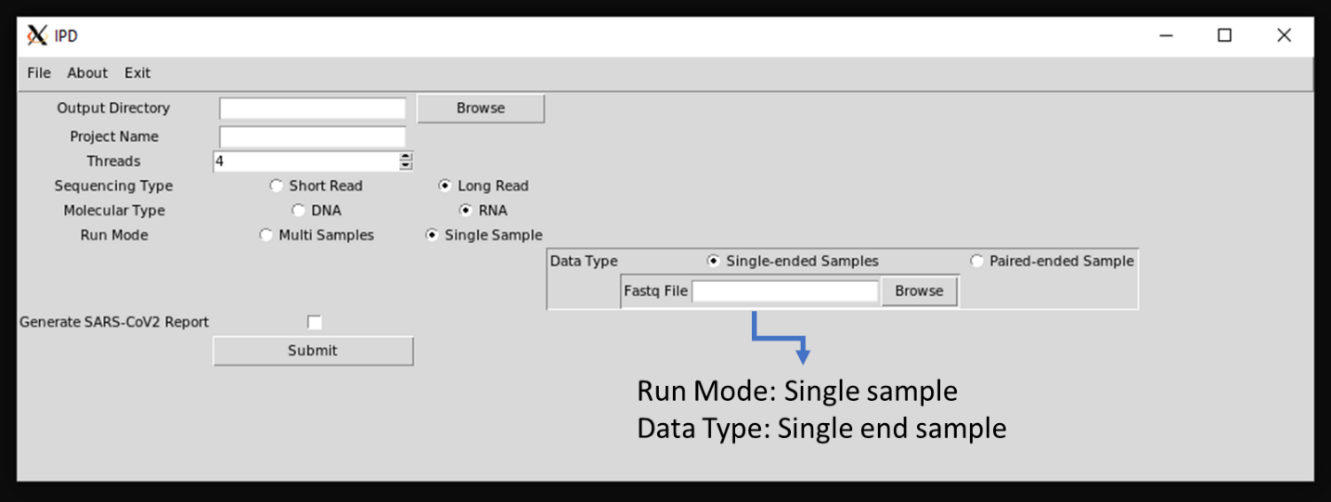
SampleName <tab> testfile\_r1.fastq.gz <tab> testfile\_r2.fastq.gz

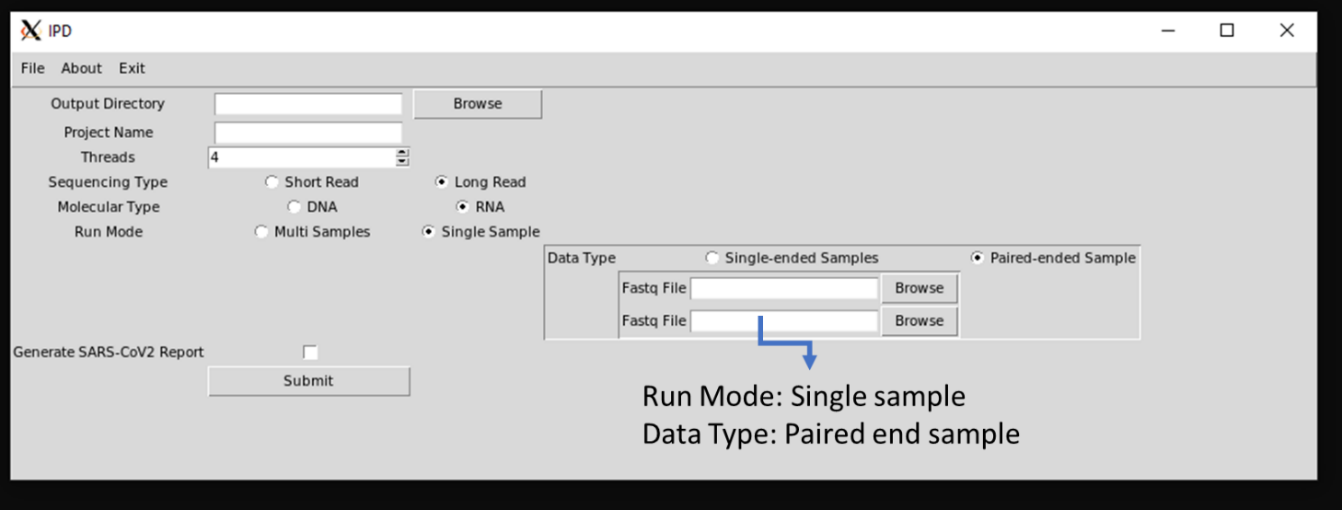
**Example Sample File format for single-end data:**

SampleName <tab> testfile\_r1.fastq.gz



For Single sample run mode there is further two options for the data type, paired-end and single-end. It enables the user to browse the fastq/fastq.gz input files. Project name will be used as the prefix for all the output files in this case.



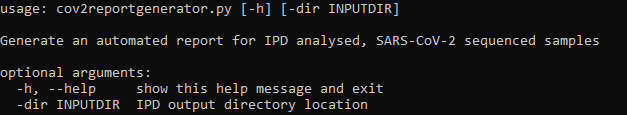


SARS-CoV2 report generation

Sars-Cov2 report generation script enables to user to visualize the coverage and expression of SARS-CoV2 in the sample. Further it provides a detailed summary of the reads, and assign clades based on the SARS-CoV2 variant profile. It required the output directory post IPD run. The code accesses the required files and generate a HTML report in the same output directory as Output.html (Refer Page 6 of 6 for sample IPD report)

Command:

$python3 cov2reportgenerator.py -dir <output directory>



1. IPD output

Sample HTML Report is attached below.

It has 4 Sections:

1. Basic Alignment Statistical summary: It includes total reads, aligned reads and read length of each sample in the project.
2. Per Base Coverage for SARS-CoV2: The read depth of each base of SARS-CoV2 genome is calculated and log2 of the reads is taken and sample-wise plots are generated
3. Relative Abundance: Stack-bar plot illustrates the relative abundance of Human, Pathogen, SARS-CoV2 and unaligned reads for each sample. The FPKM values of SARS-CoV2 are plotted in the adjacent bar plot.
4. Novel SARS-CoV2 Variants: Annotated variants not present in the IPD SARS-CoV2 vcf-database used are tabulated.
5. Variant Based SARS-CoV2 Clade Assignment: Based on the mutational profile of the sample’s clade assessment is done and tabulated in the last section of the report.

Apart from the HTML SARS-CoV2 report, IPD generates other tabulated output which are as follow:

1. Finalcount.tsv: it contains raw feature counts, length and FPM (Fragment per million) of all 1060 pathogen included in the database.
2. Final\_anno.vcf: It contains the annotated variants for all the pathogens present in database.
3. Sample\_assembledcontigs/ final.contigs.fa: It contains the assembled Contigs

\*\*\*\*IPD report on next page\*\*\*\*

