IPD

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.

Developed by: Dr. Amit Dutt Laboratory, ACTREC-TMC, Navi Mumbai, India
Web link:
Version: V 0.1.0
Developer and maintainer:
E-mail:
GUI Developer:
Documentation:

Pre-requisites 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 in python3 programming environment.

System Pre-requisites:

- Pip3
- Python3
- Conda

Python Packages:

- pysam
- Tkinter (required for GUI only)

Packages Required for automated report generation:

- Numpy
- Matplotlib
- Pandas
- SciPy

#This binary was compiled using Fedora and tested on Fedora/Red Hat OS
wget IPD.zip
unzip IPD.zip
cd IDP

#install it to system if you have sudo permission
bash install.sh

For GUI:

python3 IPD_gui.py

For Command Line interface:

python3 IPD_cli.py

Guide to use command-line interface of IPD

There are two modes in the IPD command-line interface based on the sequencing type selected.

python3 IPD_cli.py

```
usage: IPD_cli.py {long,short} ...

Sequencing_Type:
  {long,short}
```

python3 IPD_cli.py long -h

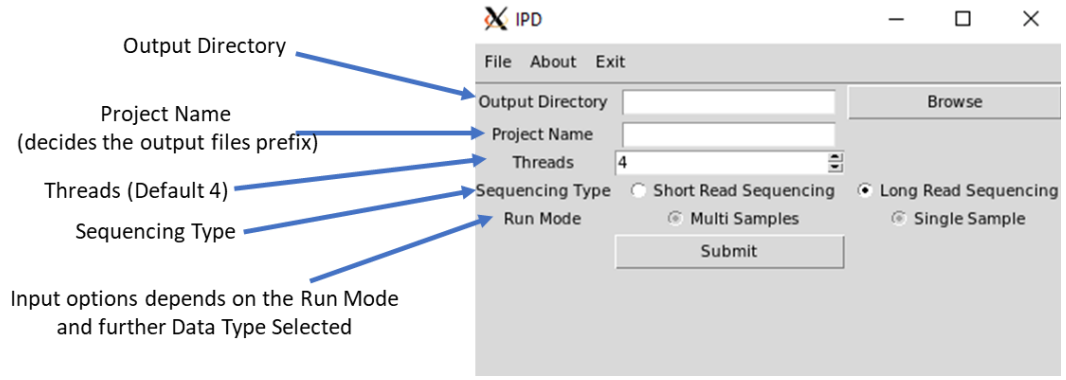
```
usage: IPD_cli.py long [-h] -projname PROJNAME [-thread THREAD] -outdir OUTDIR
                       -inputfastq INPUTFASTQ

optional arguments:
  -h, --help            show this help message and exit
  -projname PROJNAME    set project name
  -thread THREAD        set threads (default = 4)
  -outdir OUTDIR        set Output Directory
  -inputfastq INPUTFASTQ
                        Fastq file
```

python3 IPD_cli.py short -h

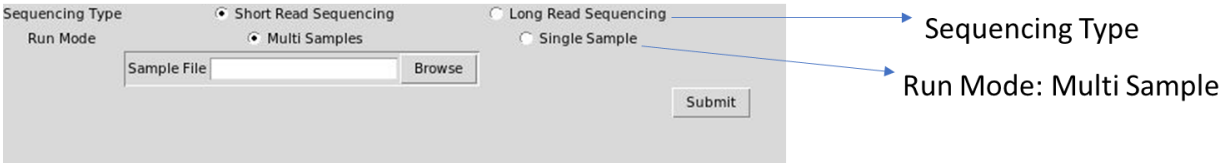
```
usage: IPD_cli.py short [-h] -projname PROJNAME [-thread THREAD] -outdir
                        OUTDIR -inputfastq_R1 INPUTFASTQ_R1 INPUTFASTQ_R1

optional arguments:
  -h, --help            show this help message and exit
  -projname PROJNAME    set project name
  -thread THREAD        set threads (default = 4)
  -outdir OUTDIR        set Output Directory
  -inputfastq_R1 INPUTFASTQ_R1 INPUTFASTQ_R1
                        Fastq file
```

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.



It has both multi-sample and single sample run mode. For Multi-sample run mode, user need to provide a sample info file containing tab separated fastq files (in case of Paired-end sample, tab separated R1 and R2 fastq file with path) and sample name. Sample name and Project name will be used as prefix for all the output files.



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 input files. Project name will be used as the prefix for all the output files in this case.

IPD output

Sample HTML Report: /link/to/report.html

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:

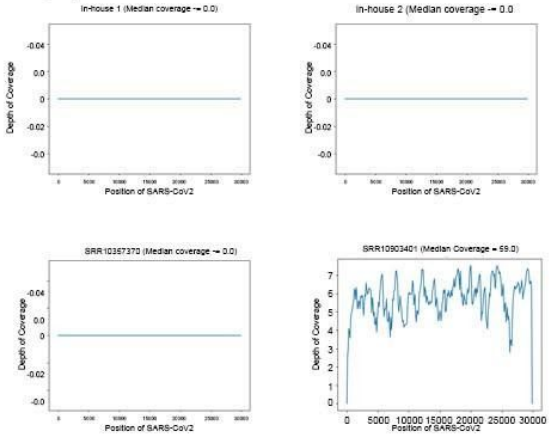
1. Finalcount.tsv: it contains all the raw and normalised counts of all 1060 pathogen included in the database.
2. Final_anno.vcf: It contains the annotated variants for all the pathogens present in database.

IPD Report

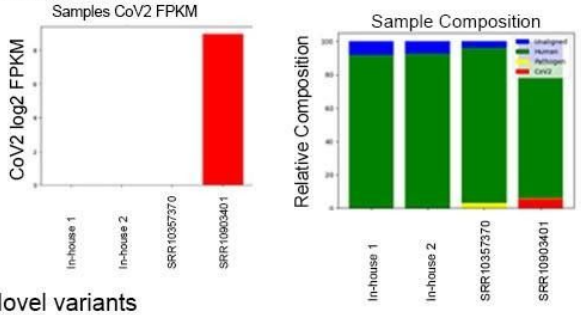
Sequence Statistics

	In-house 1	In-house 2	SRR10357370	SRR10903401
Total_Read	13387370	25911120	17608567	101109
Aligned_Reads	13168161	25580392	17471766	99176
Percent_Aligned_Read	98.36	98.72	99.22	98.09
Mean_Read_Length	104.25	149.13	149.13	120.03

Coverage plot



Relative abundance



Novel variants

Sample	Genome	Position	Reference	Altered	Consequence	Gene	Transcript	Protein Change
SRR10903401	NC_045512	24323	A	C	missense_variant	S	GU280_gp022	p.Lys921Gln

Variant based related strains

Sample	Number of Variants	Related Strains	GISAID code	Pangolin Lineage
SRR10357370	0	None	None	None
SRR10903401	2	EPI_ISL_436732-1	G	B.1
In-house 1	0	None	None	None
In-house 2	0	None	None	None