Submission of data to PITDB

We welcome submission of results PIT (proteomics informed by transcriptomics) studies for inclusion in PITDB. This document explains how to submit them. The main steps are (a) preparation of PIT results files, (b) upload of files to PITDB, (c) validation of submission by the PITDB team, (d) addition of results to the database. If you would like to make a pre-submission enquiry, please email `submissions@pitdb.org`.

Step 1: Preparing files for upload

PIT involves the analysis of a given sample by both RNA-seq and proteomic mass spectrometry followed by integration of the acquired data. The results of a PIT study therefore including data of different type, which need to be provided in a consistent way. To be compatible with PITDB, the data much be processed in the prescribed way, resulting in the files described below. Our strong recommendation for producing results in the right format is to use our PIT software pipeline, which is available via GitHub[1].

Experiment and sample description files

`experimentInfo.tsv`: This is a tab separated variable (TSV) file with three columns, namely Title, Description and Publication, and a row for the experiment. The first row of the file contains the column heading. Title column is the title for the experiment and Description column should have a brief description about the project. Publication column should provide reference information for the publication associated with the experiment.

`sampleInfo.tsv`: This is a TSV file with two columns, namely SampleName and Description. The first row of the file contains the column headers. Each row should describe a sample. It is essential that all associated data files for the sample (see below) have the sample name, e.g. if sample name is `sample1`, then the TGE FASTA file should be called `sample1.fasta`, the associated GFF3 file should be called `sample1.GFF3` and so on.

Data folders

`transcripts`: This folder contains FASTA files for transcript sequences that produced at least one ORF that was supported by mass spectrometry peptide evidence. There should be one transcript FATSA file per sample. The first word (all characters before the first space) from the sequence header will be stored in PITDB as the transcript name. Example: If `>transcript1 len=600 ENST10002334` is the sequence header, then `transcript1` will be the transcript name.

`AminoAcids-or-ORFs-orTGEs`: This folder should contain FASTA files, one per sample, containing identified ORFs (TGEs). Sequence ID has to start with the transcript ID followed by `|` and then ORF ID. For example, if transcript `transcript1` produces two open reading frames, their sequence headers will start with `transcript1|orf1` and `transcript1|orf2`. Any other information related to the TGE needs to be added after inserting a space.

`PSMs-Peptides-ORFs`: This folder contains two commas separated values (CSV) files per sample, one for peptide spectrum match (PSM) exported using mzIdentML-lib and the other is protein export using the same tool. The PSM export file should be named as `sample1.csv` and the protein export file should be called `sample1+prt.csv`.

[1] https://github.com/bezzlab/TGEClassification

`Variations-proVCF`: This folder contains two CSV files per sample, one is for polymorphisms and the other is for isoforms. Isoform proVCF is optional at the moment. For a sample named `sample1`, polymorphism proVCF should be named as `sample1.vcf` and the isoform proVCF is named `sample_isoforms.vcf`.

`GFF3`: This folder contains two GFF3 files per sample, one with the mapping of all the transcripts from the transcript folder and the other is for all the peptides identified for a sample. For a sample named `sample1`, the transcript GFF3 file is named as `sample1.gff3` and the peptide one as `sample1_peptide.gff3`.

`Summary`: A TSV file containing a summary of TGEs. Each row represents a TGE from the sample. For a sample named `sample1`, the files in this folder should be named as `sample1.tsv`.

Step 2: Submitting data to PITDB

All the necessary files should be combined together into a single zip file. This can then be submitted for inclusion in PITDB by clicking the Submission button in the top right of the PITDB web interface. You will be prompted to enter your name and email address (for future correspondence) and to upload the zip file. Clicking Submit will send the submission to the PITDB team.

Step 3: Validation of submission

Your submission will be acknowledged by the PITDB team, and then inspected to ensure that the necessary files are present, and in the correct format. This is a manual process, so turnaround times may vary. If there are any problems with the submission, you will be notified and assisted in correcting these.

Step 4: Addition of results to the database

When we satisfied that all the files are present and correct, we will add them to the database and email you when they are ready to view.