Contents

The SRA_RUNFINDER

There is a huge amount to be potentially gained by re-analysing publicly available RNA-seq transcriptomic data much of which is deposited at the SRA (https://www.ncbi.nlm.nih.gov/sra). Currently, the process of submitting sequences to the repository are difficult and time consuming. For the meta data, there are many different web forms to be filled in and many objects are referenced by redundant names. Consequently, what exactly is put in each column is research group dependent. Further, many groups do not seem to amend their deposits when corresponding papers are published. This means that on the SRA site, you can often go round in circles trying to find information about runs. Further, the downloaded run fastq files from the SRA do come with the appropriate run metadata.

For many tasks, the raw sequence data is required for reanalysis. These are big files of several GB and remapping them is a computer intensive process. Here we have made an application that can make it easier to peruse the available runs, allowing for careful selection of available runs with all the information on them readily accessible including links to papers not on the SRA site retrieved via entrez.

When picking runs, you may for example want to choose just one run of a repeat of three similar runs but choose the one with the most spots. Further, you may want to hone into the exact genotype that you want. This is a crucial stage as later analysis can be thrown by mis-picking of runs at this stage.

For many of the runs, the submitter has specified the forward and reverse strands. However, although this has been scraped and the data available in this application, due to the complexity of form filling in, it is prudent not to rely on this annotation. Instead, once the runs are downloaded from the SRA site one run per project should be mapped to a transcriptome using Salmon to check for the strandedness https://github.com/COMBINE-lab/salmon).

Example here used data from the SRA on *Aspergillus Fumigatus* runs downloaded May ‘22. Programmes are also included to form new databases of other sequence runs.

To search for useful sequence runs in Aspergillus fumigatus using SRA_RUNFINDER.py

```
$ python SRA_RUNFINDER.py
```

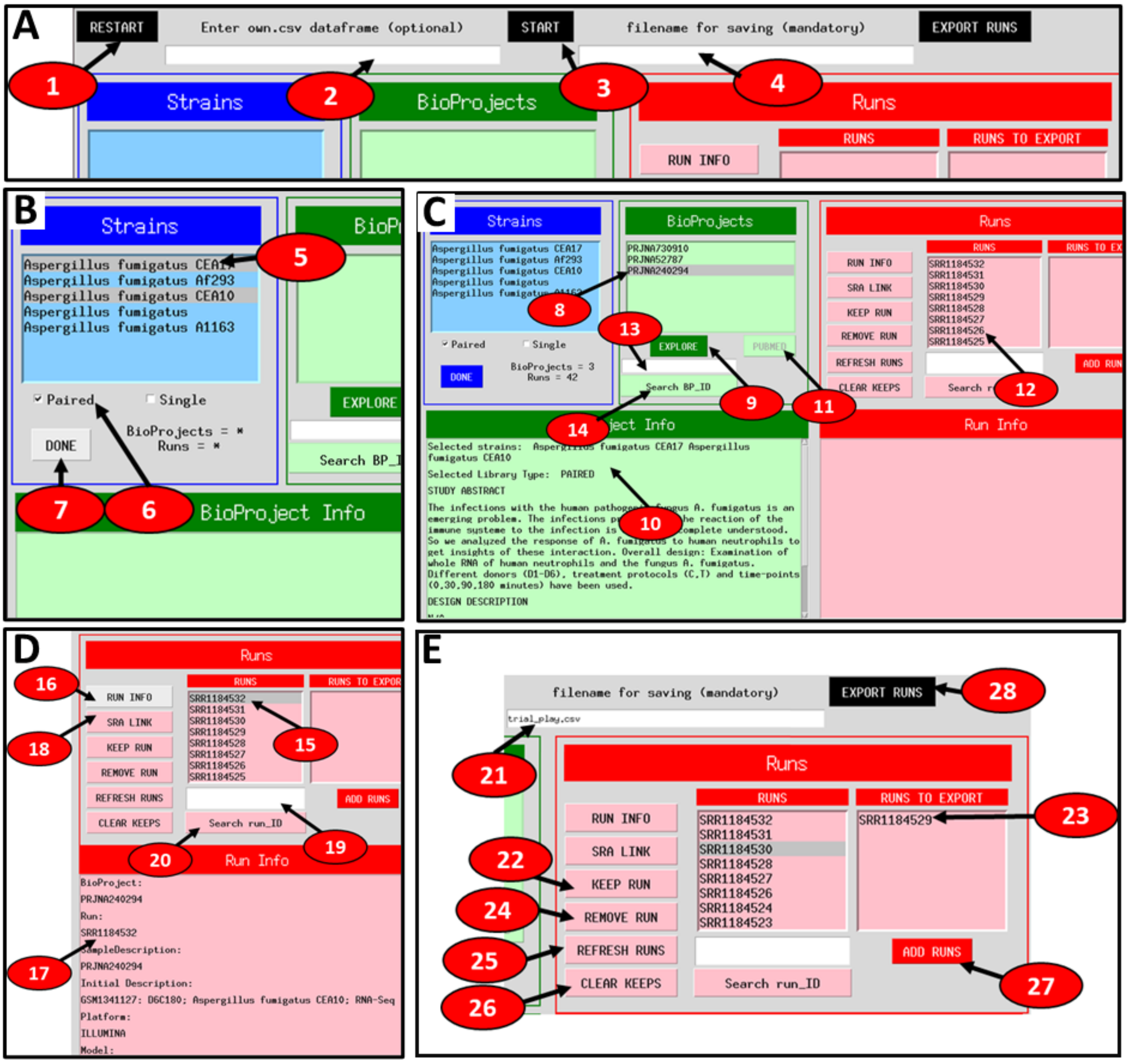
Requires:

SRA_RUNFINDER.py (webbrowser, tkinter, pandas, BeautifulSoup, os, sys)

Combined_pmc.csv (or a cropped version)

 Outputs:

Cropped version of Combined_pmc.csv with only chosen runs



```
Aspergillus fumigatus CEA17
Aspergillus fumigatus Af293
Aspergillus fumigatus CEA10
Aspergillus fumigatus
Aspergillus fumigatus A1163
```

```
PRJNA730910
PRJNA52787
PRJNA240294
```

```
Selected strains:  Aspergillus fumigatus CEA17 Aspergillus
fumigatus CEA10

Selected Library Type:  PAIRED

STUDY ABSTRACT

The infections with the human pathoge  fungus A. fumigatus is an
emerging problem. The infections p      he reaction of the
immune systeme to the infection is      complete understood.
So we analyzed the response of A. fumigatus to human neutrophils to
get insights of these interaction. Overall design: Examination of
whole RNA of human neutrophils and the fungus A. fumigatus.
Different donors (D1-D6), treatment protocols (C,T) and time-points
(0,30,90,180 minutes) have been used.

DESIGN DESCRIPTION
```

```
SRR1184532
SRR1184531
SRR1184530
SRR1184529
SRR1184528
SRR1184527
SRR1184526
SRR1184525
```

```
BioProject:
PRJNA240294
Run:
SRR1184532
SampleDescription:
PRJNA240294
Initial Description:
GSM1341127: D6C180; Aspergillus fumigatus CEA10; RNA-Seq
Platform:
ILLUMINA
Model:
```

```
trial_play.csv
```

```
SRR1184532
SRR1184531
SRR1184530
SRR1184528
SRR1184527
SRR1184526
SRR1184524
SRR1184523
```

```
SRR1184529
```

Figure1: SRA sequence run examiner.

(A)

(1) "RESTART" button – resets everything by re-opening application.
(2) To retrieve a previously saved *.csv* enter the file name.
(3) Starts application.
(4) Enter a filename to save output – **mandatory**.

(B)

(5) This brings up the strains that can be selected. Often the exact ones have not been defined by the submitters. Multiple can be selected.
(6) Choose Paired and or Single read sequences
(7) When chosen press "Done"

(C)

(8) This brings up the available BioProjects in the "BioProjects" window. These can be selected individually.
(9) To explore information press "EXPLORE".
(10) This brings up relevant information in the "BioProject Info" window.
(11) If either a PMID access provided by submitter, or a PMC retrieved via entrez-direct the "PUBMED" button will be highlighted via which the paper's abstract/paper can be accessed. If there are multiple entries only the first is accessed.
(12) The runs fitting the criteria are also now displayed in the "RUNS" list box.
(13) BioProjects IDs can also entered into the entry box', then searched for by pressing the "search BP_ID" key **(14)**.

(D)

(15) Runs can be selected.
(16) Clicking the "RUN INFO" button brings up information in the "Run Info" scroll text box **(17)**
(18) The link to this run's SRA information page, can be accessed by clicking the SRA button.
(19) As before runs can be searched by inserting their SRA run ids into the entry box and using the "Search run_ID" submit button **(20)**.

(E)

(21) Make sure a filename is entered into the saving box.
(22) Selected runs can be kept by clicking on the "KEEP RUN" button, which moves the run ID to the "RUNS TO EXPORT" list box **(23)** and removes it from the "RUNS" list box.
(24) Selected runs can also be discarded from the runs box by clicking "REMOVE RUN" option. This way the RUNS box can be cleared, and you can keep track of runs examined.
(25) The RUNS list box can be reset to the original by pressing "REFRESH RUNS".
(26) To clear the "RUNS TO EXPORT" list box, click "CLEAR KEEPS".
(27) If you are satisfied with the runs in the "RUNS TO EXPORT" list box, click the "ADD RUNS" button. This put the runs and associated information into the file specified in

(21). You can continue looking through projects and adding runs to the database. It does not matter if the same run is added more than once.

(28) When you have finished click the "EXPORT RUNS" button. The file is then de-duplicated. To relook through this file at selections made or add more to it, simply restart the programme, enter the location of your file and press start and repeat the processes.

To build a new database for SRA_RUNFINDER.py

STEPS 1-6 show how to make a tailor made data frame. To use supplied *A. fumigatus* one, jump in at **STEP 7.**

STEP 1: Get required TRANSCRIPTOMIC runs meta data from SRA

```
bash grab_SRA.sh fumigatus
```

Use a single **whole word** that should be in the runs you want such as "fumigatus" for *Aspergillus fumigatus*. Search appears to be case insensitive. An excess of runs can be picked at this stage as runs will be filtered down progressively. (This may take a few minutes).

Requires:

sra-tools   2.11.0 and entrez-direct 16.2 Pre-requisite:

```
>>> conda install -c bioconda sra-tools
>>> conda install -c bioconda entrez-direct
```

Outputs:

SraRunTable.txt

STEP 2: Extract run SRA Transcriptomic STUDY IDs on the command line

```
$ python get_SRA_STUDY_2.py -i SraRunTable.txt
```

Requires:

get_SRA_STUDY_2.py (pandas, argparse)

SraRunTable.txt

Outputs:

SRA_STUDY.txt

STEP 3: Download associated Run metafiles (.csv) and xml files from SRA

```
$ bash run_info_retrieval.sh SRA_STUDY.txt
```

Requires:

sra-tools   2.11.0 and entrez-direct 16.2 Pre-requisite:

```
>>> conda install -c bioconda sra-tools
>>> conda install -c bioconda entrez-direct
```

bash_run_info_retrieval.sh

SRA_STUDY.txt

Outputs:

SRA_STUDY associated run data as csv and xml files

STEP 4: Extract and merge data from the run info csv files and scraped xml files

```
bash merge_SRA_INFO_1.sh SRA_STUDY.txt fumig
```

Requires:

merge_files_SRA2.py (pandas, BeautifulSoup, argparse)

merge_SRA_INFO.sh

SRA_STUDY.txt

cleaning_combined_1.py (pandas, argparse)

A short key phrase should be supplied to restrict the data frame to only have runs that have the phrase in the "ScientificName". Here we have used “fumig” to get all *Aspergillus fumigatus* runs.

Output:

cleaned_combined.csv cleaned dataframe

bioproject.txt extracted Bioproject Ids

The studies are entered somewhat randomly into different columns of the data frame by researchers. A short key phrase should be supplied to restrict the dataframe to only have runs that have the phrase in the "ScientificName". Here we have used “fumig” to get all *Aspergillus fumigatus* runs.

- Extra data is scraped from the XML file including submitted PUBMED IDs using merge_files_SRA.py .
- The files are then concatenated and the no longer required files deleted.
- The data frame in this file is then cleaned using cleaning_combined.py and the Bioproject Ids extracted

STEP 5: Searches PMC for associated references

```
$ bash extract_relevant_PMC.sh bioproject.txt  SRA_STUDY.txt
```

- Finds on PMC papers associated with BioProject Ids and merges into one file
- Finds on PMC papers associated with SRA Study Ids and merges into one file
- Requires:
    - bioconda entrez-direct
    - bioproject.txt
    - SRA_STUDY.txt
- Outputs bigfile1 and bigfile2

STEP 6: Add PMC references to data frame

```
$  python pmc_adding.py
```

Requires:

cleaned_combined.csv

bigfile1

bigfile2

pmc_adding.py (pandas, re, numpy)

Outputs:

Combined_pmc.csv