björn.biolab_release()

This is a draft document containing instructions on how to use *biolab_release()* - an early-stage extension to *björn* to facilitate post-processing of SARS-CoV- 2 genomic sequence data from biolab for public release via GISAID.

Installation

Pre-requisites

- familiarity with Python and the command-line interface (e.g. Terminal on MacOS)
- Install Windows Subsystem for Linux (only if you are running on Windows Operating System)
    - Follow these instructions for installation
- Open Ubuntu
- Using the Ubuntu terminal, install Anaconda environment management tool using the code below.

```
wget https://repo.anaconda.com/archive/Anaconda3-2020.11-Linux-x86_64.sh
```

- Clone the code repository (Install Bjorn on Ubuntu using the below code)

```
git clone https://github.com/andersen-lab/bjorn.git
```

- Open the repository

```
cd bjorn/
```

- Change to the biolab_genomics branch of the repository

```
git checkout biolab_genomics
```

- Install the Anaconda environment

```
conda env create -f env/linux.yml -n bjorn
```

Update

- Open Ubuntu application from Windows search bar
- Open bjorn directory

```
cd bjorn/
```

- Activate the bjorn environment

```
conda activate bjorn
```

- Update bjorn code for biolab

```
git pull origin biolab_genomics
```

Usage

- Activate the environment (the image below shows a screenshot of how the environment gets activated)

```
conda activate bjorn
```

- Open biolab_config.json using a text editor

```
code biolab_config.json
```

- Specify the required parameters for the release run
    - fasta_hub : the folder in Windows that contains the input sequences in separate FASTA files
    - meta_hub : the folder in Windows that contains the input metadata in Excel file. Please see the
    - next section for a description of the expected format of the metadata
    - results_hub : the folder in Windows that contains the output from bjorn, to be used for upload purposes
    - the remaining parameters are self-explanatory and usually do not need to be changed
- Save biolab_config.json and close the file
- Ensure that the required FASTA files are stored inside the folder specified under fasta_hub in biolab_config.json
- Ensure that the required metadata Excel file is stored inside the folder specified under meta_hub in biolab_config.json
- Run biolabs release

```
python3 src/biolabs_release.py --out-dir release_output_[YYYY-MM-DD] --
date [sequencing date] --metadata [metadata file name] --coverage [minimum
coverage threshold e.g. 80] --depth [minimum depth threshold e.g. 200]
```

- Locate the results in your Windows OS inside results_hub path
- Load the alignment file inside the msa folder using Geneious Prime, or any alternative alignment viewer
- Open the spreadsheet file named suspicious_mutations.csv
- Perform manual inspection on the sequence alignment
- Save the cleaned sequence alignment in FASTA format
- append _clean to the filename e.g. 2021 - 06 - 10 _release_aligned_clean
- save in the same msa folder within the results_output_hub
- Convert the alignment into an unaligned FASTA (i.e. concatenated FASTA file)
- Return to the Ubuntu window and press ENTER to proceed to the final step

N.B.: you may need to press ENTER twice to proceed *bjorn* should automatically generate an unaligned FASTA file from the clean alignment file inside the same *msa* folder (e.g. _2021 - 06 - 10 *release_clean*), which is ready for upload to GISAID along with the associated metadata file

- Preparing metadata
- Open the release metadata file and delete all rows under the header inside the Submissions sheet
- Open the raw metadata file and copy all rows under the header
- Paste the rows under the header inside the release metadata file Submissions sheet
- Fix format of Collection date by copying all values and pasting them inside the Date Format Correction sheet
- Copy the fixed dates and paste back into the Collection date column using Paste values option (see screenshot below)
- Change the values under FASTA filename column to the name of the final concatenated FASTA file
- Upload the cleaned concatenated FASTA and the associated Excel file to GISAID via the bulk upload channel

Input Metadata Format

In order for bjorn to work as expected, the input metadata file must be created in a specific format. Metadata needs to be stored in an Excel file containing three sheets:

1. Instructions
    - This contains the default GISAID template for submission
2. Submissions
    - This contains the metadata in the format required by GISAID
3. Coverage
    - This contains QC metrics for each sequence and should look like this biolab_metadata

Additional useful commands

- To identify the mutations present in raw FASTA files, we can run the biolab_fasta_ 2 _mutations supplementary script
- Place all FASTA files inside the FASTA_ 2 *explore folder found in FASTA* 2 _mutations_bjorn

- Open the Ubuntu window and activate bjorn

```
cd bjorn
conda activate bjorn
```

Run the following command

```
python3 src/biolab_fasta_2_mutations.py --out-dir mutations_output_[YYYY-MM-DD] --date [sequencing date]
```

- The mutations results can then be found inside the FASTA_ 2 *mutations_bjorn folder under the pre-specified mutations_output*[YYYY-MM-DD] folder