BGE skim analysis pipeline

Download data from SciLifeLab

Have to install dds cli: https://scilifelabdatacentre.github.io/dds_cli/installation/

login first: **dds auth login**

Username =
Password =

OTP from email

List files in a project: **dds data ls --project snpseq00629**

Download all the data using: **dds data get --get-all --project snpseq00629**

Add user by: **dds user add --project snpseq00629 --role Researcher someone@email.com**

logout: **dds auth logout**

From <https://scilifelabdatacentre.github.io/dds_cli/auth/>

Generate sample list from folders

Run: 1_folder2csv_trim.py
cd into directory you want the output CSV file first, then run

```
python 1_folder2csv_trim.py path/to/run/folder
```

output is CSV with three columns called "[mainfolder]_folder2csv_out.csv"

ID	forward	reverse
Subfolder name, trimmed by first 15 characters to remove "Sample_" and project name. = BOLD Process ID	Full path to the forward reads	Full path to the reverse reads

Note: batch 1 had BOLD sample ID not Process ID so I've manually updated this column

---

Download hierarchical NCBI taxID for each specimen

Need biopython installed: **pip install biopython**

Start with BOLD download, using taxonomy tab of spreadsheet
Add Process ID to sheet so that sample ID and Process ID are in same sheet and save as a CSV file
Run sample2taxid.py

```
python 2_sample2taxid.py input_file.csv output_file.csv
```

Script searches for taxID starting with Species, if no match then Genus, then Family, then Order

two output files:
output_file.csv = input file with columns added for taxID and match rank
[input filename]_unique_taxids.txt = deduplicated list of taxIDs ready for go batch script

---

Make the skim2phylo sample sheet

Need pandas installed: **pip install pandas**

This combines the taxIDs with the sample ID and paths for the reads to make a csv file formatted for *skim2phylo* pipeline

Script: 3_makeSKIMsamples.py and the output CSV files generated above

```
python 3_makeSKIMsamples.py run_folder2csv_out.csv sample2taxid_out.csv
```

output is [input filename]_merged.csv
the key innovation here is that the output CSV now has paths for GetOrganelle references for each sample

ID	forward	reverse	seed	gene
BSNHM002-24	Path…	Path…	refs/177658/seed.fasta	refs/177658/gene.fasta
BSNHM003-24			refs/177627/seed.fasta	refs/177627/gene.fasta
BSNHM004-24			refs/177860/seed.fasta	refs/177860/gene.fasta

The paths will need to be made relative, not absolute, otherwise singularity falls over. I use find and replace in excel to go
from:
/gpfs/nhmfsa/bulk/share/data/mbl/share/workspaces/groups/genomics-collections/BGE/….
to:
../../ BGE/….

Run Go Fetch (batch script)

Install go fetch from: https://github.com/o-william-white/go_fetch

Input is the [input filename]_unique_taxids.txt generated above

Update the location of the go_fetch script, currently set to:
/home/benjp/software/go_fetch/go_fetch.py

If using array make sure to set value in go_fetch.sh script to match # unique taxIDs.
I use array but only 1 at a time to stop overwhelming the api.

The output is a folder, within each is a subfolder for each taxID and the corresponding seed.fasta and gene.fasta files for each sample

Check Go Fetch results

Sometimes it falls over. This script checks the taxID folders all have gene.fasta and seed.fasta files and prints a list of those that don't have either file.

```
python 4_skipped_gofetch.py path/to/folder
```

Output = skip_gofetch.txt

Rerun gofetch on this text file and update relevant folders before moving onto skim2phylo

Move the go-fetch reference data subfolders into the …/skim2phylo/refs folder

Run skim2phylo

Install from here: https://github.com/o-william-white/skim2phylo

Inputs:

- Samples.csv = [input filename]_merged.csv (make sure relative paths used)
- Config YAML (specify the samples filename)
- sbatch file specifying cluster stuff (specify the config filename)

GetOrganelle references should be in a folder called "refs" with subfolders named by taxID (the output of GoFetch batch script)

Note: if starting from a fresh pull of skim2phylo you need to manually add the "trim poly g" option for fastp:

```
"""
    if [ {fastp_dedup} == True ]; then
        fastp --in1 {input.fwd} --in2 {input.rev} \
            --out1 {output.fwd} --out2 {output.rev} \
            --html {output.html} --json {output.json} \
            --trim_poly_g \
            --dedup \
            --thread {threads} &> {log}
    else
        fastp --in1 {input.fwd} --in2 {input.rev} \
            --out1 {output.fwd} --out2 {output.rev} \
            --html {output.html} --json {output.json} \
           --trim_poly_g \
            --thread {threads} &> {log}
    fi
    """
```