SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Lecture 2C: AncientMetagenomeDir

James A. Fellows Yates & Megan Michel

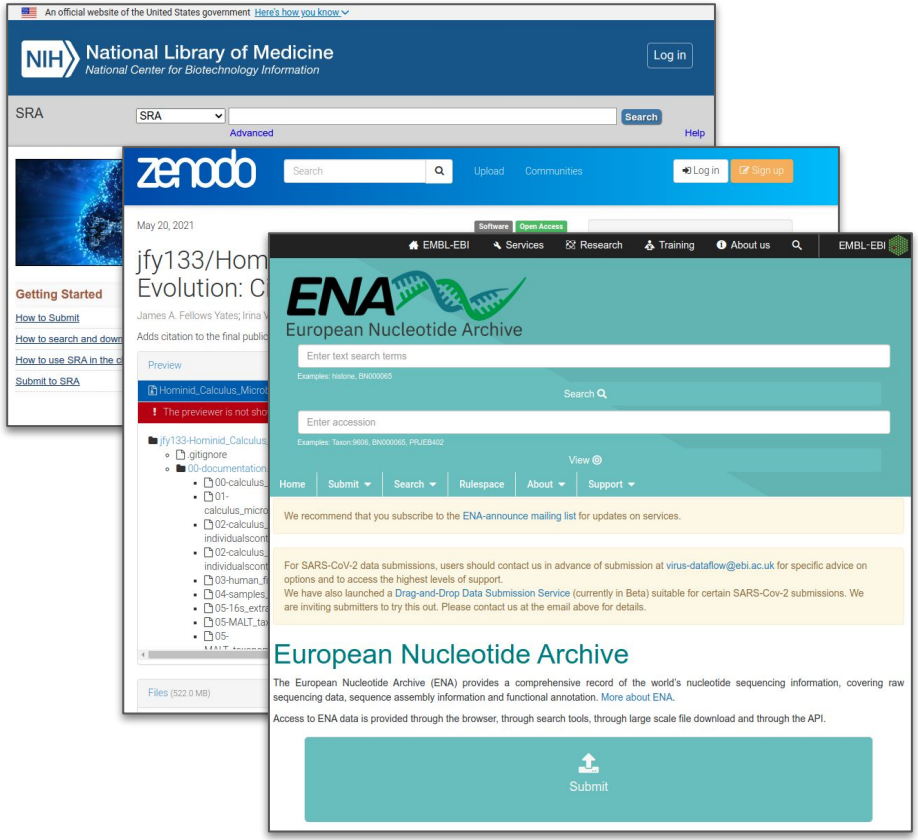
Where to get data?

Why get and use public data

- Expand sample sizes (stronger statistical power)
  - (possibly) improve robustness of microbiome differences
  - Add more nuance to phylogenetic trees
- Comparative data (preservation assessment)
  - How does your data compare to others?
- Meta-analyses
  - New ideas and projects!

Where to get sequencing data?

- Main
  - NCBI SRA
  - EBI ENA
- Others
  - Institutional FTP
  - Domain specific databases
  - Zenodo
  - Figshare
  - Github



Is ancient metagenomic data FAIR*?

- Findable
- Accessible
- Interoperable
- Reproducible

Imagen: Australian National Data Services

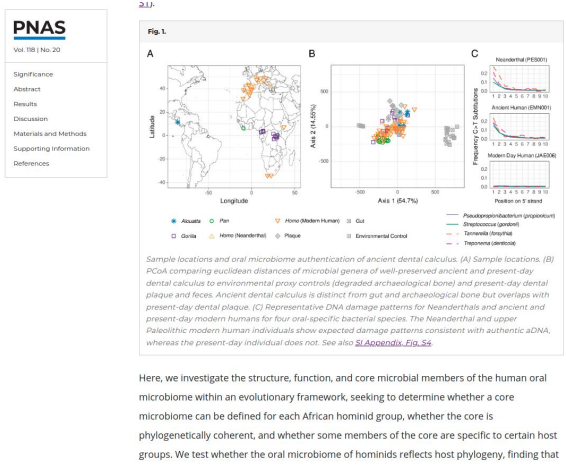
Ancient metagenomic data

- Findable: ❌
  - ...not so much...
- Accessible ✅
  - Mostly! all on sequence archives
- Interoperable: ✅
  - Standardised file formats.
- Reproducible:
  - ➖ Mostly (see interoperable)

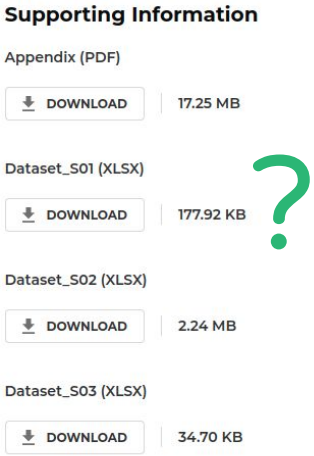
*Wilkinson (2016) Sci. Data.

Why is getting data difficult?

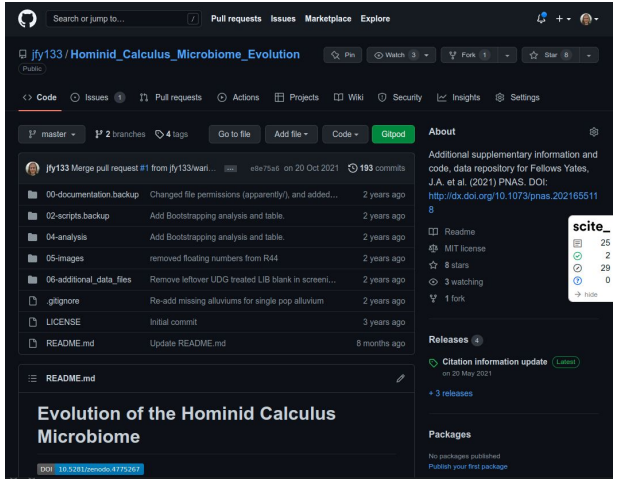
- Metadata in different places (main text/SI/external databases)
- Each author reports different types of metadata
- Each author reports metadata in different formats
- Each database has its own metadata types/formats



?



?



Ancient
Metagenome
Dir

What is AncientMetagenomeDir?

AncientMetagenomeDir is a

community curated resource of lists of all published shotgun-sequenced ancient metagenome or microbial genome-level enriched samples.

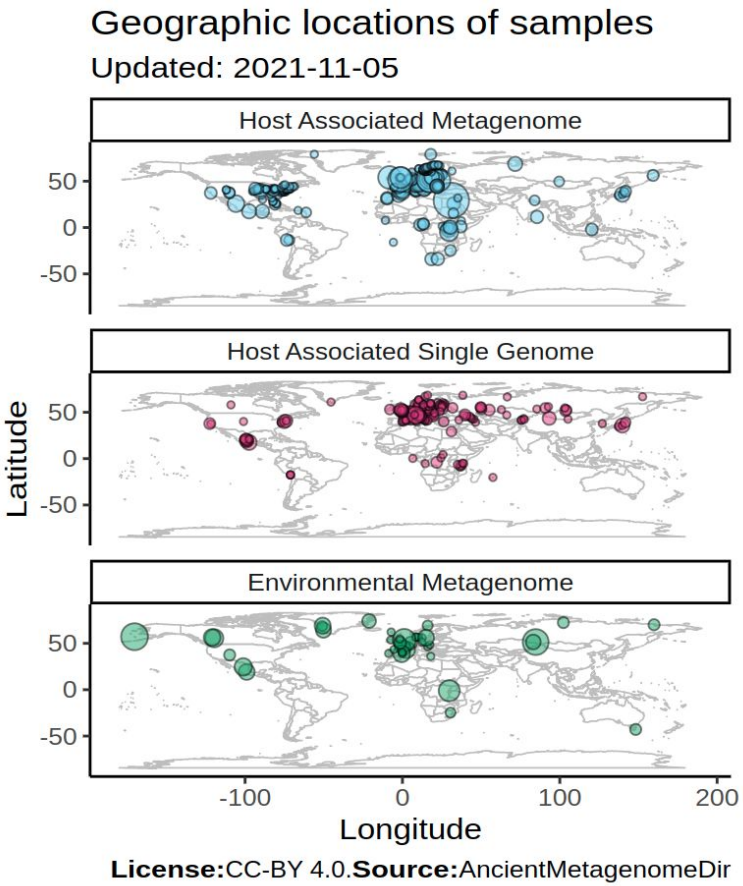
It is primarily meant to act as a reference guide to help point researchers toward any relevant public data for comparative analysis. It is hoped it will help researchers track growth and development of the field of ancient metagenomics over time.

Focus areas: ancient…

Host associated metagenome

Host associated single-genome

Environmental metagenome



What does it look like?

546 lines (546 sloc) | 94 KB

Edit on HackMD | Raw | Blame

Search this file...

	project_name	publication_year	publication_doi	site_name	latitude	longitude	geo_loc_name
2	Warinner2014	2014	10.1038/ng.2906	Dalheim	51.565	8.84	Germany
3	Warinner2014	2014	10.1038/ng.2906	Dalheim	51.565	8.84	Germany
4	Weyrich2017	2017	10.1038/nature21674	Gola Forest	7.657	-10.841	Sierra Leone
5	Weyrich2017	2017	10.1038/nature21674	El Sidrón Cave	43.386	-5.328	Spain
6	Weyrich2017	2017	10.1038/nature21674	El Sidrón Cave	43.386	-5.329	Spain
7	Weyrich2017	2017	10.1038/nature21674	Spy Cave	50.48	4.674	Belgium

- Download TSV & open in spreadsheet program
- Filter for period/area of interest & use accession

What does it contain?

A set of TSV files with **standardised** column fields. Covers:

- Publication
- Sample Name
- Sample Type
- Age
- Geographic Location
- Archive Accessions IDs (not data!)

Coming soon!

- Library level metadata!
  - 2 v 4 colour chemistry?
  - damage removal?
  - all reads? Mapped only?
- Download preparation tool!
  - in-browser filtering/selection
  - generates download scripts
- Standardised reporting sheet
  - MINAS project!

README.md

DOI 10.5281/zenodo.4003826 | pypi package ?.?.? | docs passing

AMDirT

BETA

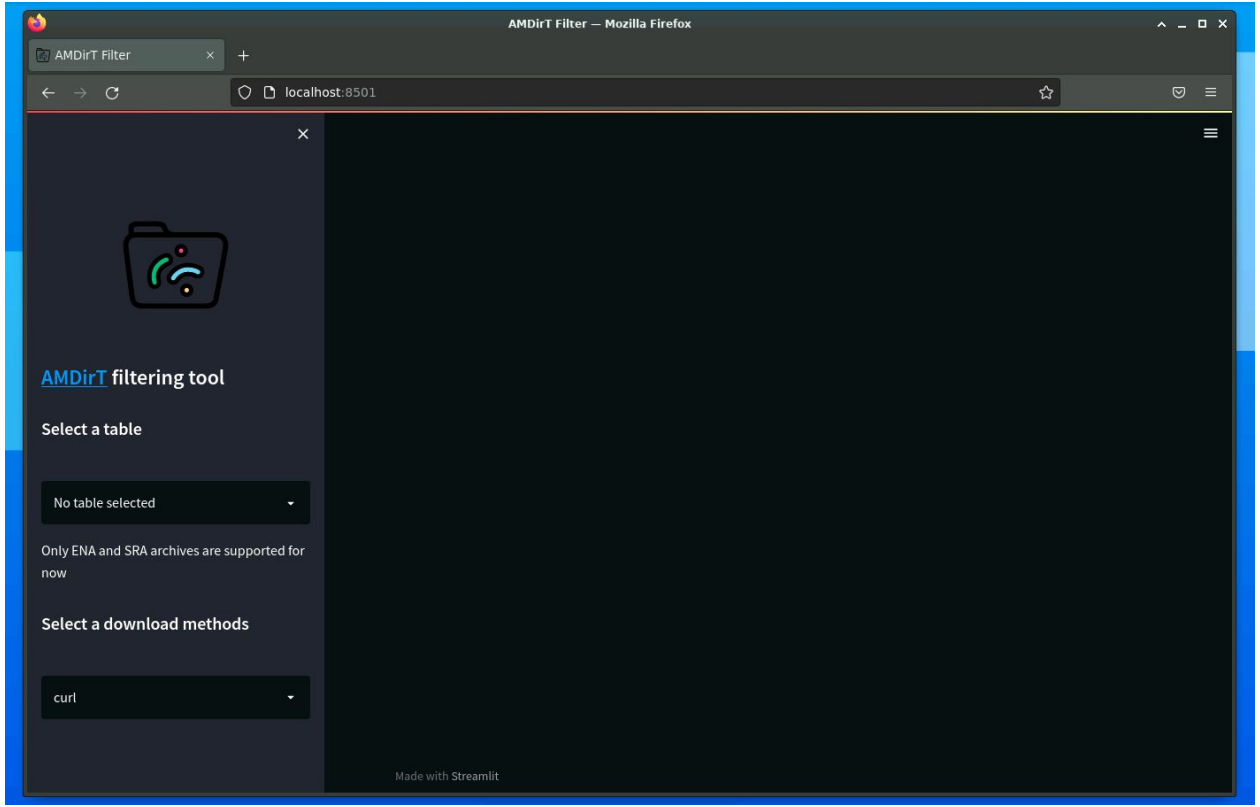
AMDirT: AncientMetagenomeDir Toolkit

AMDirT: open

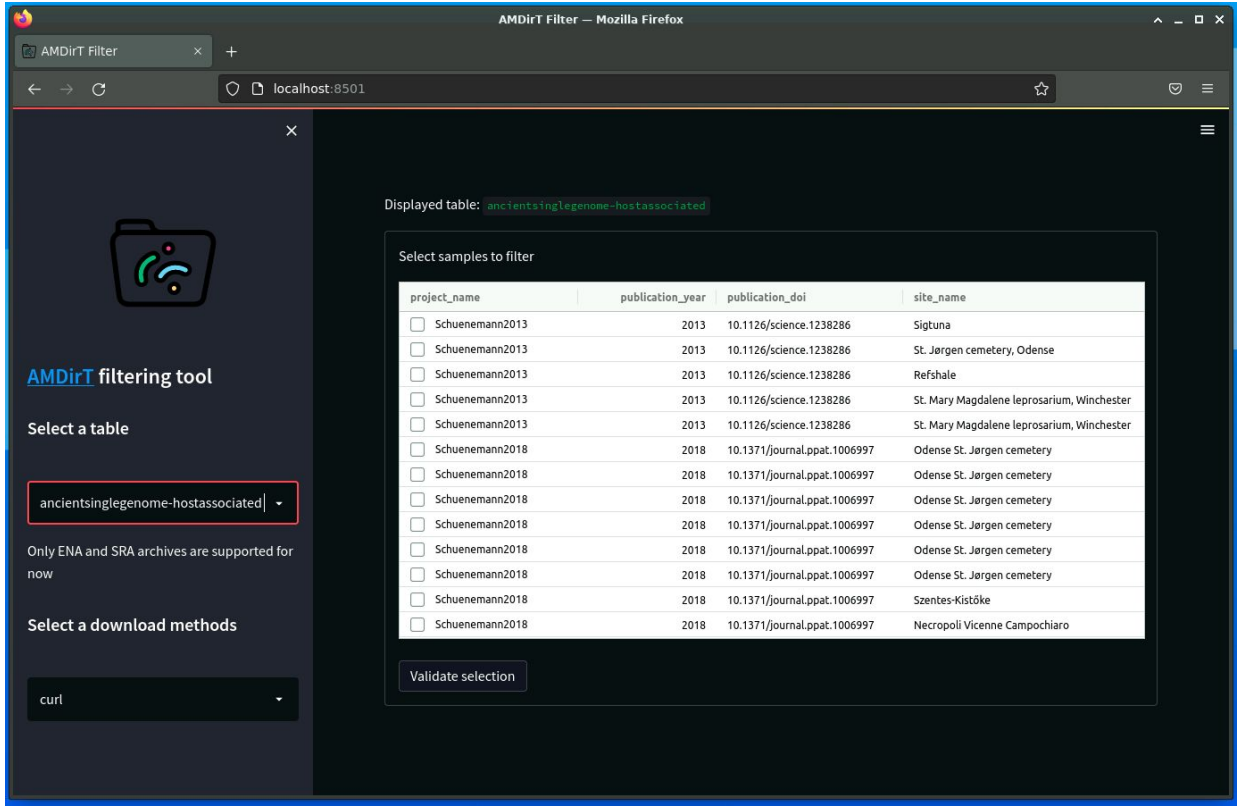
```
$ cd /vol/volume/2c-introduction-to-ancientmetagenomedir
$ conda activate git-eager
$ AMDirT filter
```

This will open a web browser for you!

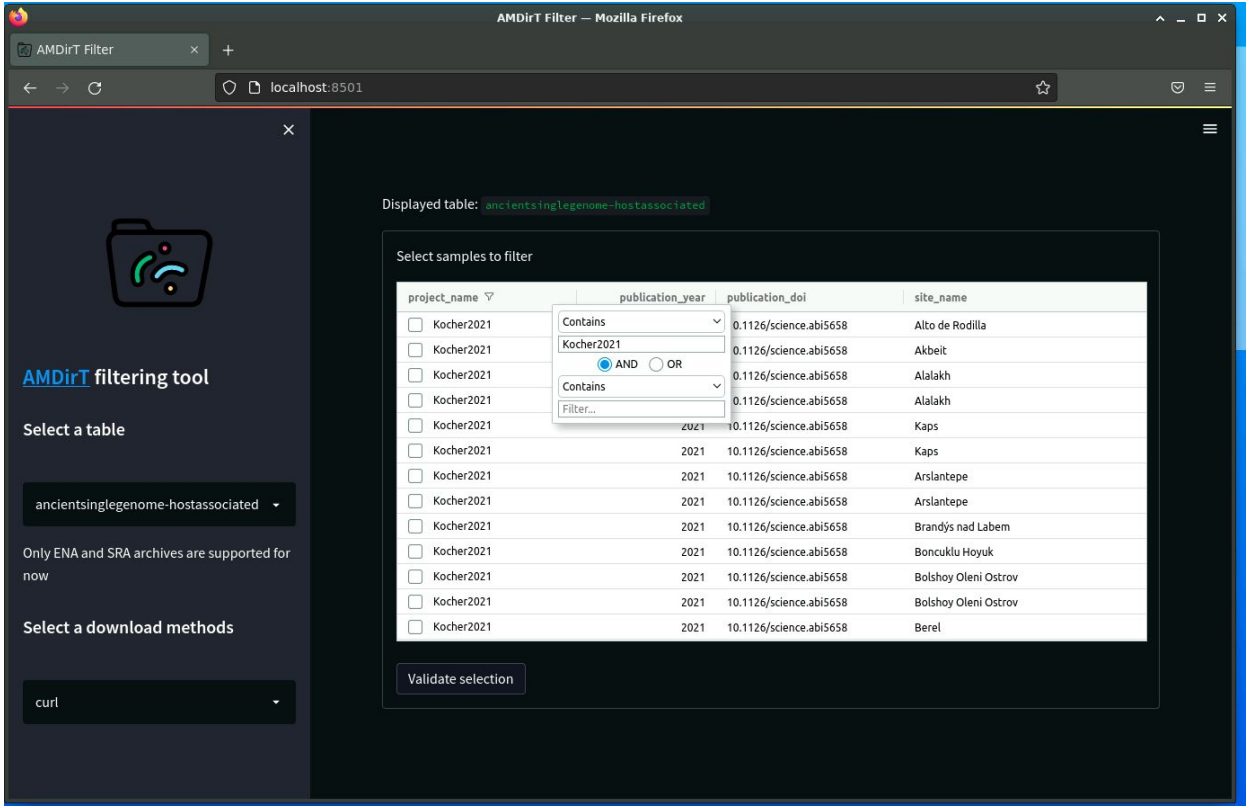
AMDirT: first page



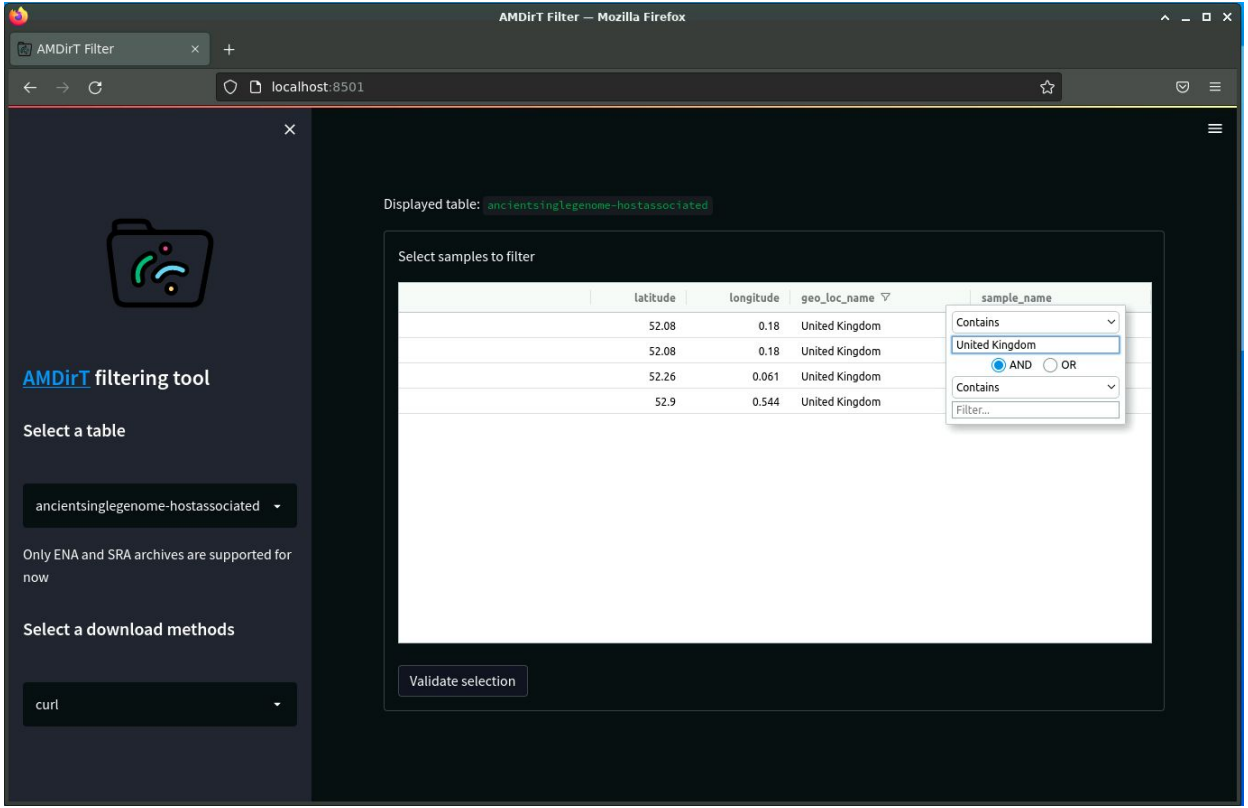
AMDirT: select table



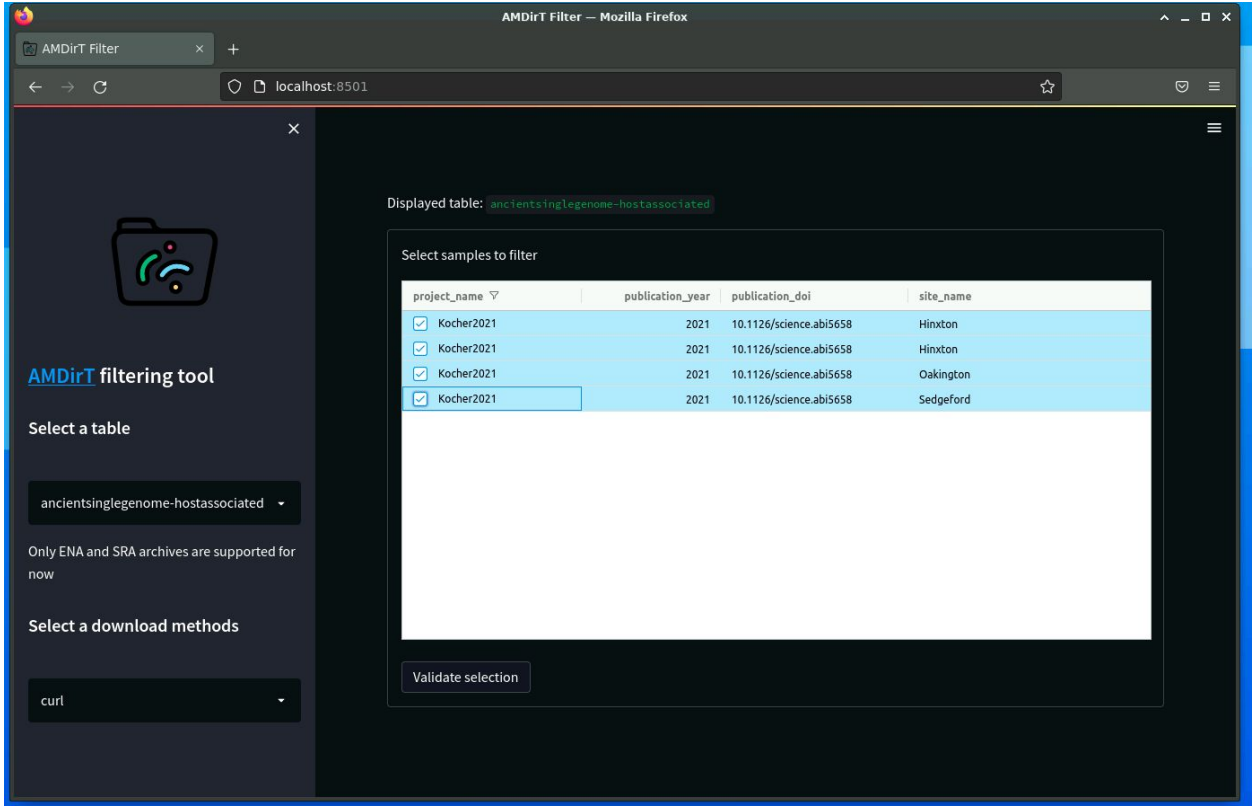
AMDirT: filter by publication



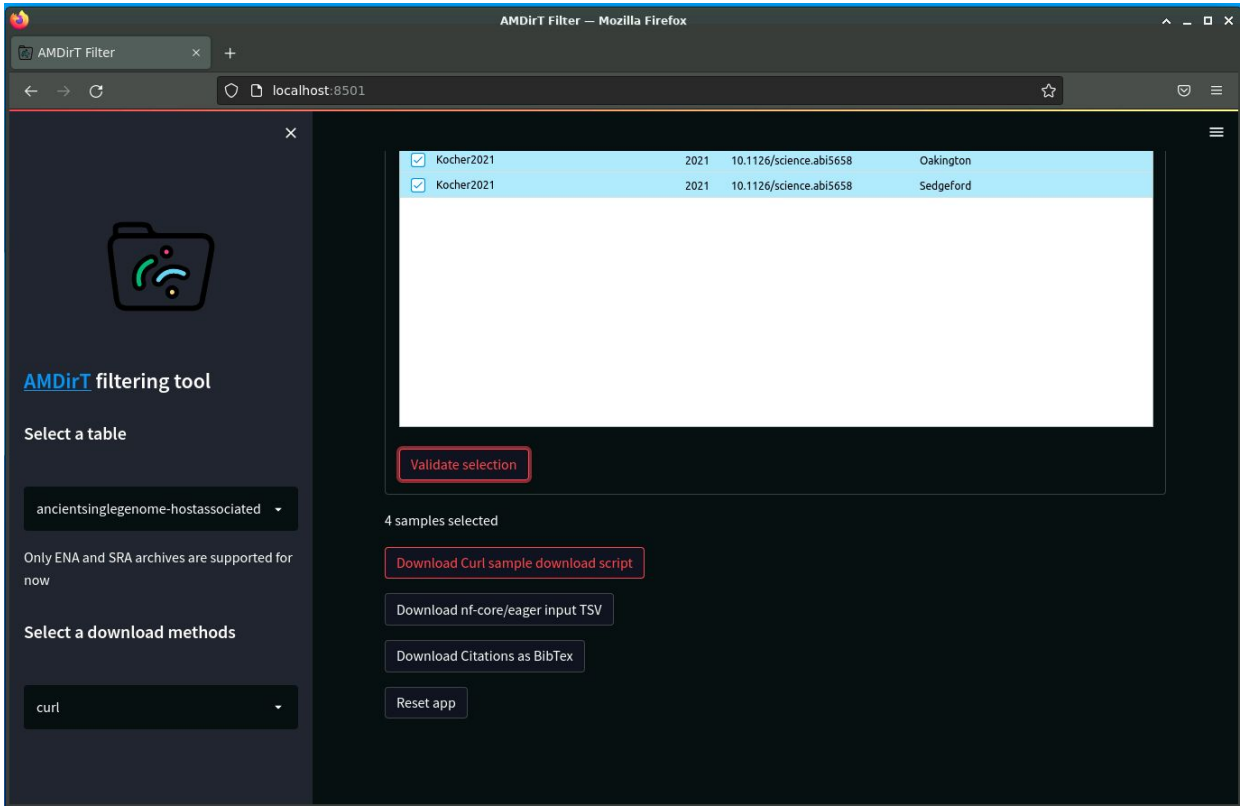
AMDirT: filter by country



AMDirT: validate selection



AMDirT: download download script



Download:

- Curl
- nf-core/eager
- bibtex

AMDirT: To exit

- Close tab in web browser
- On terminal: press `ctrl + c` on keyboard to exit!

Task

- What is inside the download script?
- What does it do? What does output look like?

```
$ cd ~/Downloads
$ ls
$ cat ancientMetagenomeDir_curl_download_script.sh
<...>
$ bash ancientMetagenomeDir_curl_download_script.sh
$ zcat ERR6053618.fastq.gz | head -n 8
```


AMDirT: references

How to correctly cite downloaded data?

Check the BibTex file!

```
$ cd ~/Downloads
$ cat ancientMetagenomeDir_citations.bib
```


AMDirT: references

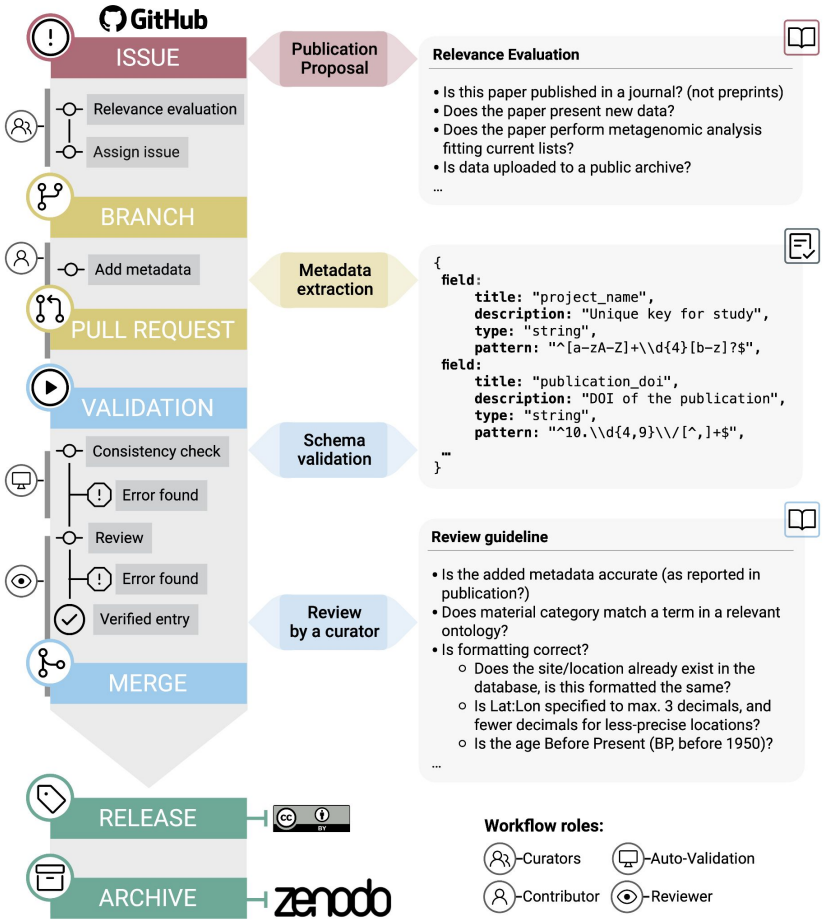
How to process the data (already!)

```
$ cd ~/Downloads
$ cat ancientMetagenomeDir_eager_input.csv
```

Relevance to Git(Hub)?

All hosted, and maintained on GitHub!

- Submissions created on branches
- Submissions via pull requests
- Reviews!



git practise!

Your task

1. Make a copy the jfy133/AncientMetagenomeDir repo to your account
   a. https://github.com/jfy133/AncientMetagenomeDir
2. Download the **copied** repo to your compute node
3. Change to the `dev` branch
4. Modify 'ancientsinglegenome-hostassociated_samples.tsv'
   a. Click here to get the data to copy in to the *end* of the TSV file
5. Send back to Git(Hub)
6. Request adding changes to the **original** repo

Recap

- Reporting of metadata messy! Consider when publishing your own work!
  - Use AncientMetagenomeDir as a template
- Use AncientMetagenomeDir and AMDirT (beta) to rapidly find public ancient metagenomic data
- Contribute to AncientMetagenomeDir with git (!)
  - Community curated!

Solves the problem of

…disjointed and inconsistent critical information needed data reuse, spread over main text, supplement, data upload etc.

Puts it in one place and makes it easy to filter, explore, and identify relevant data for *your* work