

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



The image shows three overlapping screenshots of web interfaces related to sequencing data:

- Top Screenshot:** The National Library of Medicine (NIH) National Center for Biotechnology Information (NCBI) SRA (Sequence Read Archive) search page. It features a search bar with 'SRA' entered and a 'Search' button.
- Middle Screenshot:** A Zenodo repository page for a dataset titled 'jfy133/Hominid_Calculus_Micro... Evolution: C...'. The page includes a 'Getting Started' section with links for 'How to Submit', 'How to search and download', 'How to use SRA in the context of Zenodo', and 'Submit to SRA'. A file list is visible below, showing folders like '00-documentation', '01-00-calculus', '01-01-calculus', '02-calculus_micro', '02-calculus_individualscom', '02-calculus_individualscom', '03-human', '04-samples', '05-16s_exts', '05-MALT_tax', and '05-MALT_tax'.
- Bottom Screenshot:** The European Nucleotide Archive (ENA) homepage. It features a search bar for 'text search terms' and 'accession'. Below the search bar, there are navigation links for 'Home', 'Submit', 'Search', 'Rulespace', 'About', and 'Support'. A yellow banner recommends subscribing to the ENA-announce mailing list. A section for SARS-CoV-2 data submissions provides contact information for support and mentions a 'Drag-and-Drop Data Submission Service (currently in Beta)'. At the bottom, there is a large teal button with an upload icon and the text 'Submit'.







Is ancient metagenomic data FAIR*?

- Findable
- Accessible
- Interoperable
- Reproducible



Imagen: Australian National Data Services

Ancient metagenomic data

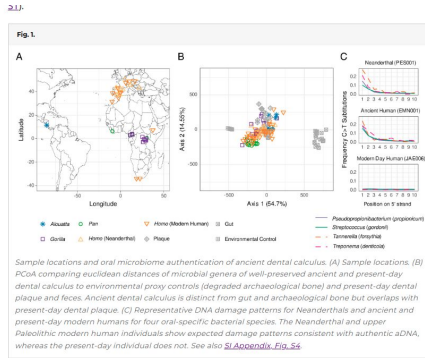
- Findable: 
 - ...not so much...
- Accessible 
 - Mostly! all on sequence archives
- Interoperable: 
 - Standardised file formats.
- Reproducible:
 -  Mostly (see interoperable)

*Wilkinson (2016) Sci. Data.

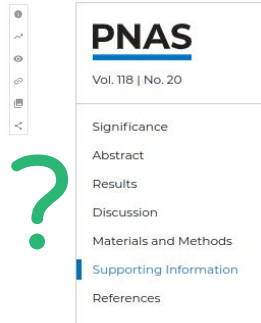


Why is getting data difficult?

- 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



Here, we investigate the structure, function, and core microbial members of the human oral microbiome within an evolutionary framework, seeking to determine whether a core microbiome can be defined for each African hominid group, whether the core is phylogenetically coherent, and whether some members of the core are specific to certain host groups. We test whether the oral microbiome of hominids reflects host phylogeny, finding that



Supporting Information

Appendix (PDF)

DOWNLOAD 17.25 MB

Dataset_S01 (XLSX)

DOWNLOAD 177.92 KB

Dataset_S02 (XLSX)

DOWNLOAD 2.24 MB

Dataset_S03 (XLSX)

DOWNLOAD 34.70 KB

Search or jump to... Pull requests Issues Marketplace Explore

jfy133 / Hominid_Calculus_Microbiome_Evolution

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 2 branches 4 tags Go to file Add file Code Gitpod About

jfy133 Merge pull request #1 from jfy133/war... e8e75d5 on 20 Oct 2021 193 commits

- 00-documentation backup Changed file permissions (apparently), and added... 2 years ago
- 02-scripts backup Add Bootstrapping analysis and table. 2 years ago
- 04-analysis Add Bootstrapping analysis and table. 2 years ago
- 05-images removed floating numbers from R44 2 years ago
- 06-additional_data_files Remove leftover UDG treated LIB blank in screen... 2 years ago
- gIgnore Re-add missing alluviums for single pop alluvium 2 years ago
- LICENSE Initial commit 3 years ago
- README.md Update README.md 8 months ago

Additional supplementary information and code, data repository for Fellows Yates, J.A. et al. (2021) PNAS. DOI: <http://dx.doi.org/10.1073/pnas.2021655118>

Readme MIT license 8 stars 3 watching 1 fork

Releases 4 Citation information update on 20 May 2021 + 3 releases

Evolution of the Hominid Calculus Microbiome

DOI: [10.5281/zenodo.4775267](https://doi.org/10.5281/zenodo.4775267)



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



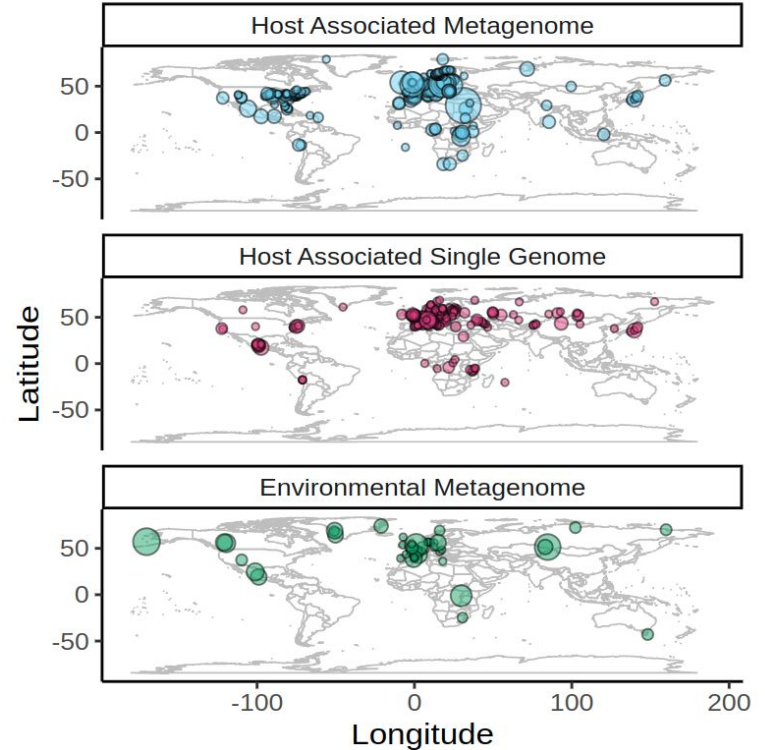
Host associated single-genome



Environmental metagenome

Geographic locations of samples

Updated: 2021-11-05



License: CC-BY 4.0. Source: AncientMetagenomeDir



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

- 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](https://doi.org/10.5281/zenodo.4003826) pypi package **7.7.7** docs **passing**

AMDirT

AMDirT: **A**n**ancient** **M**etagen**ome** **D**ir Toolkit

BETA



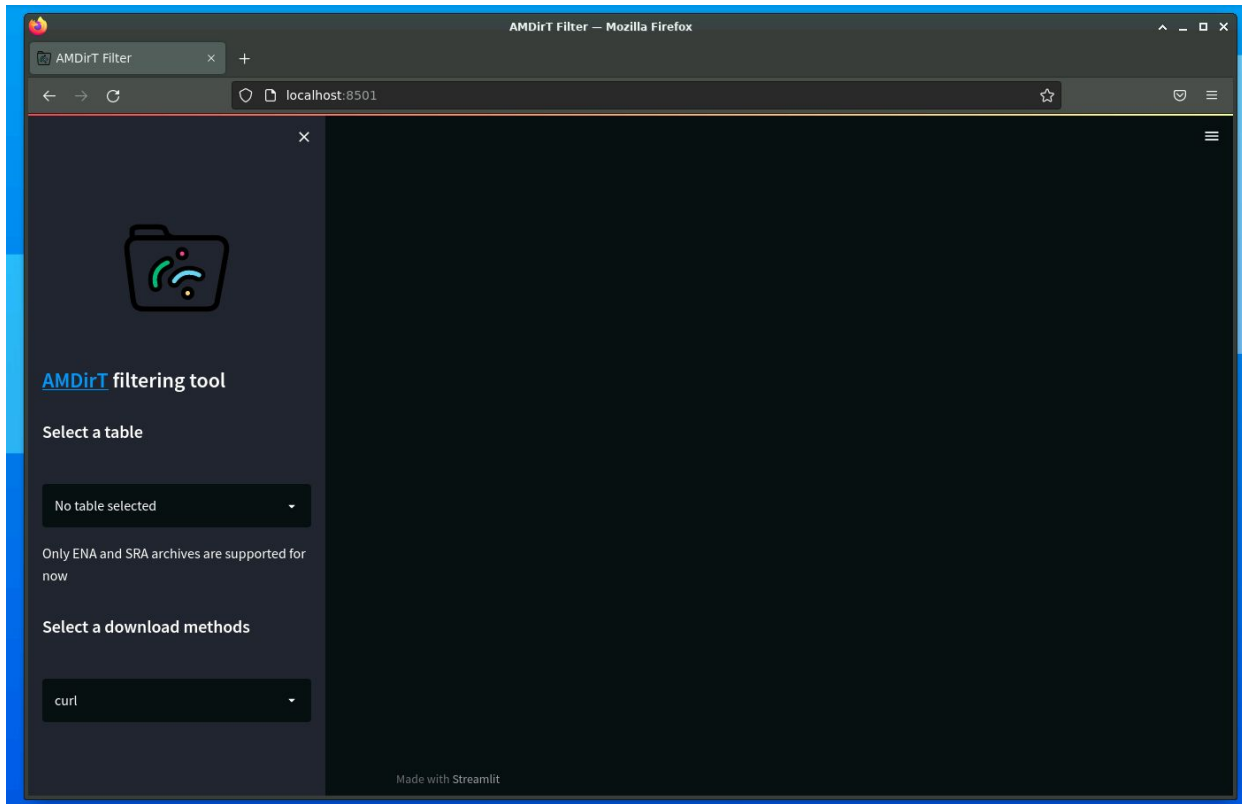
AMDirT: open

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

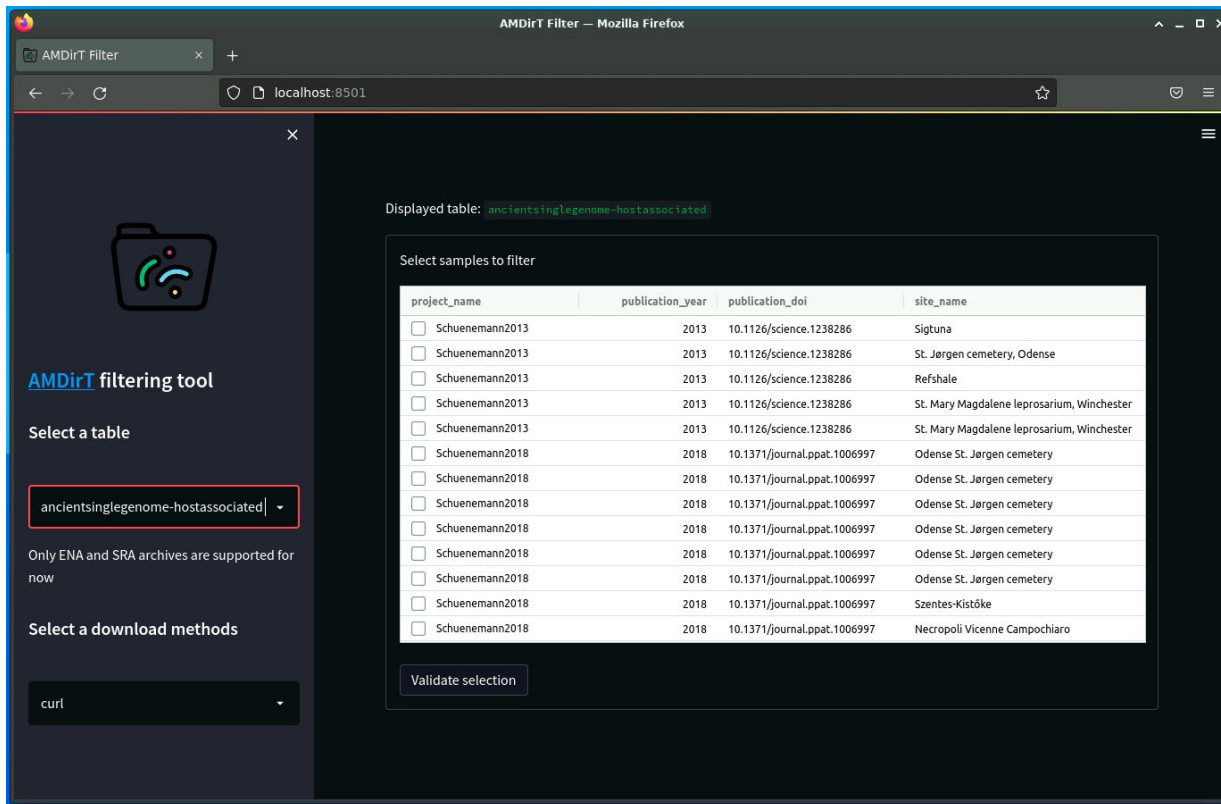
This will open a web browser for you!



AMDirT: first page



AMDirT: select table



AMDirT Filter — Mozilla Firefox

localhost:8501

AMDirT filtering tool

Select a table

ancientsinglegenome-hostassociated

Only ENA and SRA archives are supported for now

Select a download methods

curl

Displayed table: `ancientsinglegenome-hostassociated`

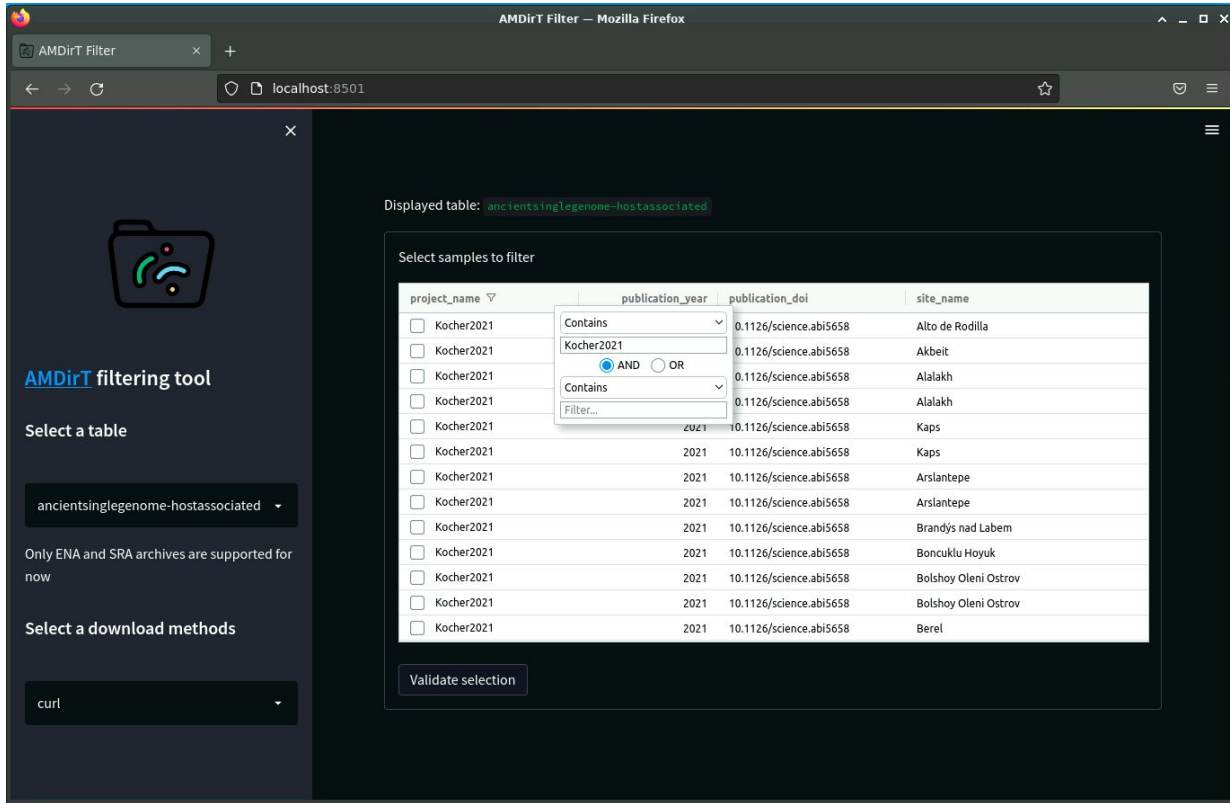
Select samples to filter

project_name	publication_year	publication_doi	site_name
<input type="checkbox"/> Schuenemann2013	2013	10.1126/science.1238286	Sigtuna
<input type="checkbox"/> Schuenemann2013	2013	10.1126/science.1238286	St. Jørgen cemetery, Odense
<input type="checkbox"/> Schuenemann2013	2013	10.1126/science.1238286	Refshale
<input type="checkbox"/> Schuenemann2013	2013	10.1126/science.1238286	St. Mary Magdalene leprosarium, Winchester
<input type="checkbox"/> Schuenemann2013	2013	10.1126/science.1238286	St. Mary Magdalene leprosarium, Winchester
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Odense St. Jørgen cemetery
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Odense St. Jørgen cemetery
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Odense St. Jørgen cemetery
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Odense St. Jørgen cemetery
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Odense St. Jørgen cemetery
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Odense St. Jørgen cemetery
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Odense St. Jørgen cemetery
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Szentes-Kistóke
<input type="checkbox"/> Schuenemann2018	2018	10.1371/journal.ppat.1006997	Necropoli Vicenne Campochiaro

Validate selection



AMDirT: filter by publication



Displayed table: `ancientsinglegenome-hostassociated`

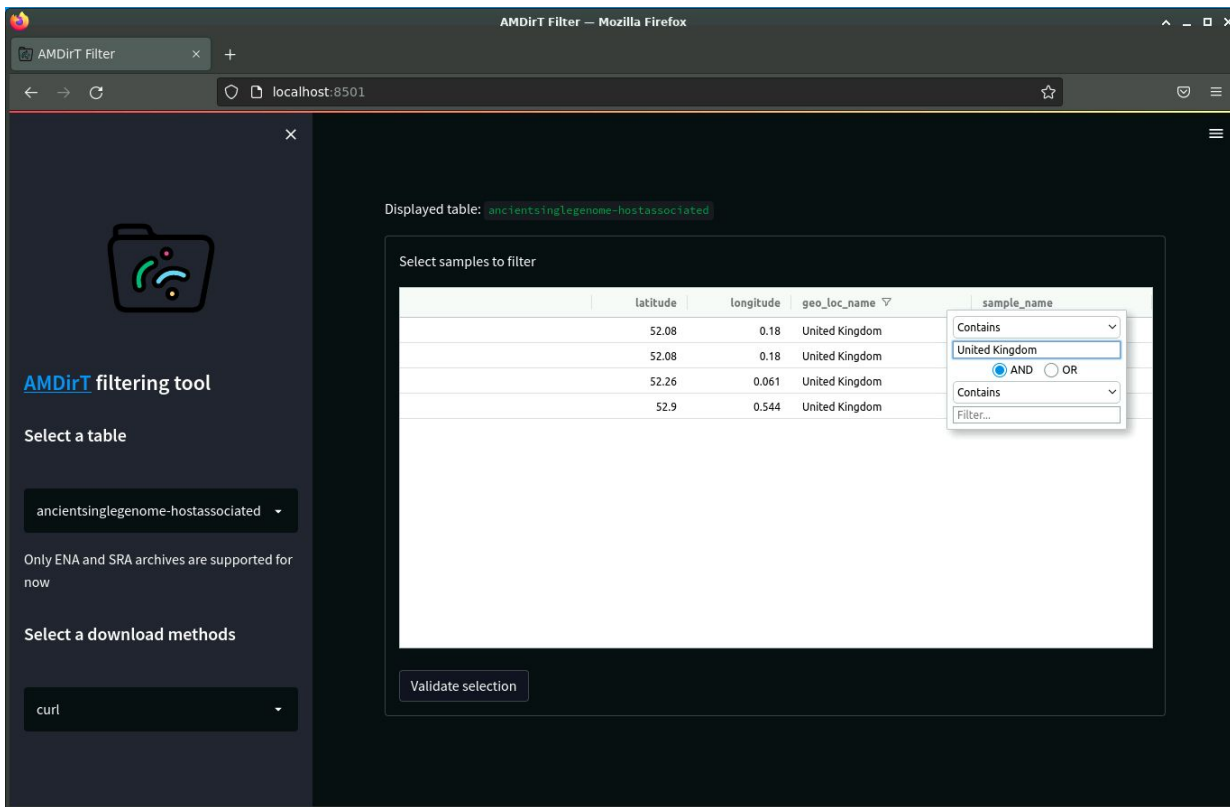
Select samples to filter

project_name	publication_year	publication_doi	site_name
<input type="checkbox"/> Kocher2021		0.1126/science.abi5658	Alto de Rodilla
<input type="checkbox"/> Kocher2021		0.1126/science.abi5658	Akbeit
<input type="checkbox"/> Kocher2021		0.1126/science.abi5658	Alalakh
<input type="checkbox"/> Kocher2021		0.1126/science.abi5658	Alalakh
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Kaps
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Kaps
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Arslantepe
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Arslantepe
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Brandýs nad Labem
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Boncuklu Hoyuk
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Bolshoy Oleni Ostrov
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Bolshoy Oleni Ostrov
<input type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Berel

Validate selection



AMDirT: filter by country



AMDirT Filter — Mozilla Firefox

localhost:8501

Displayed table: `ancientsinglegenome-hostassociated`

Select samples to filter

latitude	longitude	geo_loc_name	sample_name
52.08	0.18	United Kingdom	
52.08	0.18	United Kingdom	
52.26	0.061	United Kingdom	
52.9	0.544	United Kingdom	

Contains
United Kingdom
 AND OR
Contains
Filter...

Validate selection

Select a table

ancientsinglegenome-hostassociated

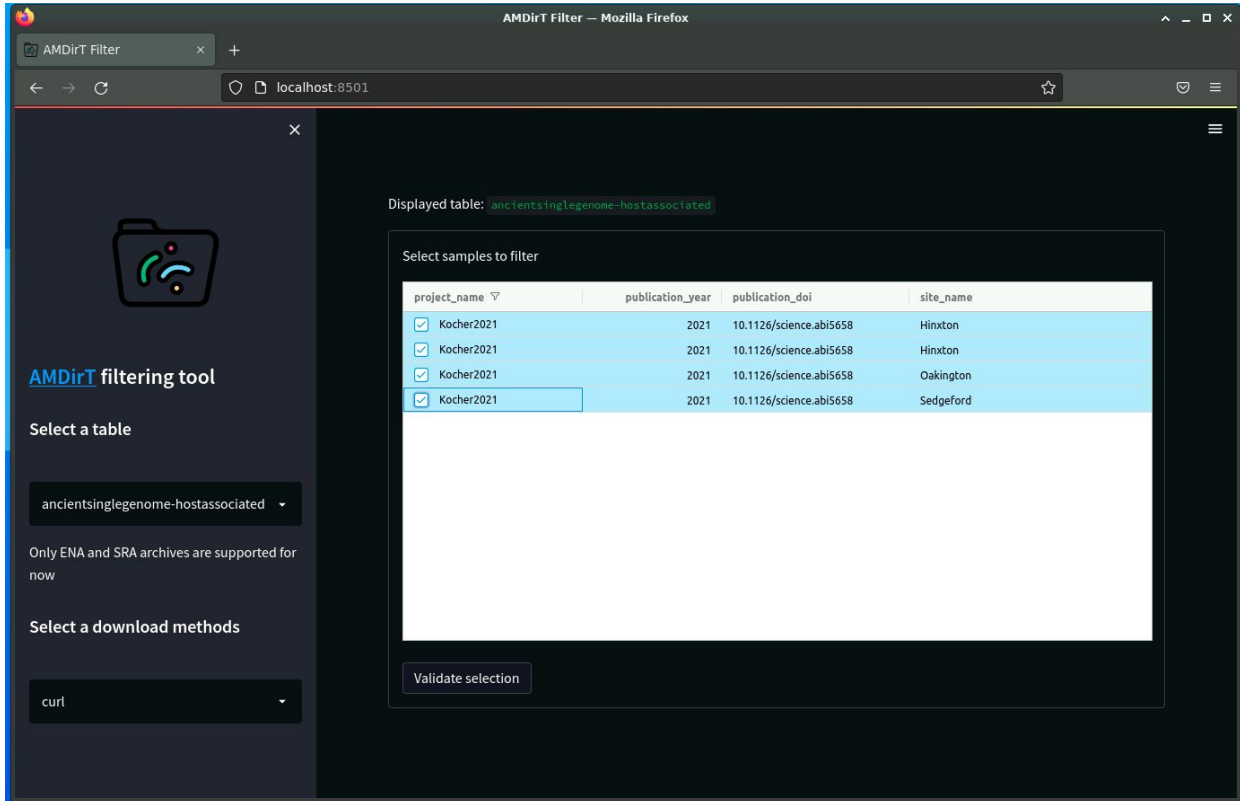
Only ENA and SRA archives are supported for now

Select a download methods

curl



AMDirT: validate selection



AMDirT Filter — Mozilla Firefox

localhost:8501

AMDirT filtering tool

Select a table

ancientsinglegenome-hostassociated

Only ENA and SRA archives are supported for now

Select a download methods

curl

Displayed table: `ancientsinglegenome-hostassociated`

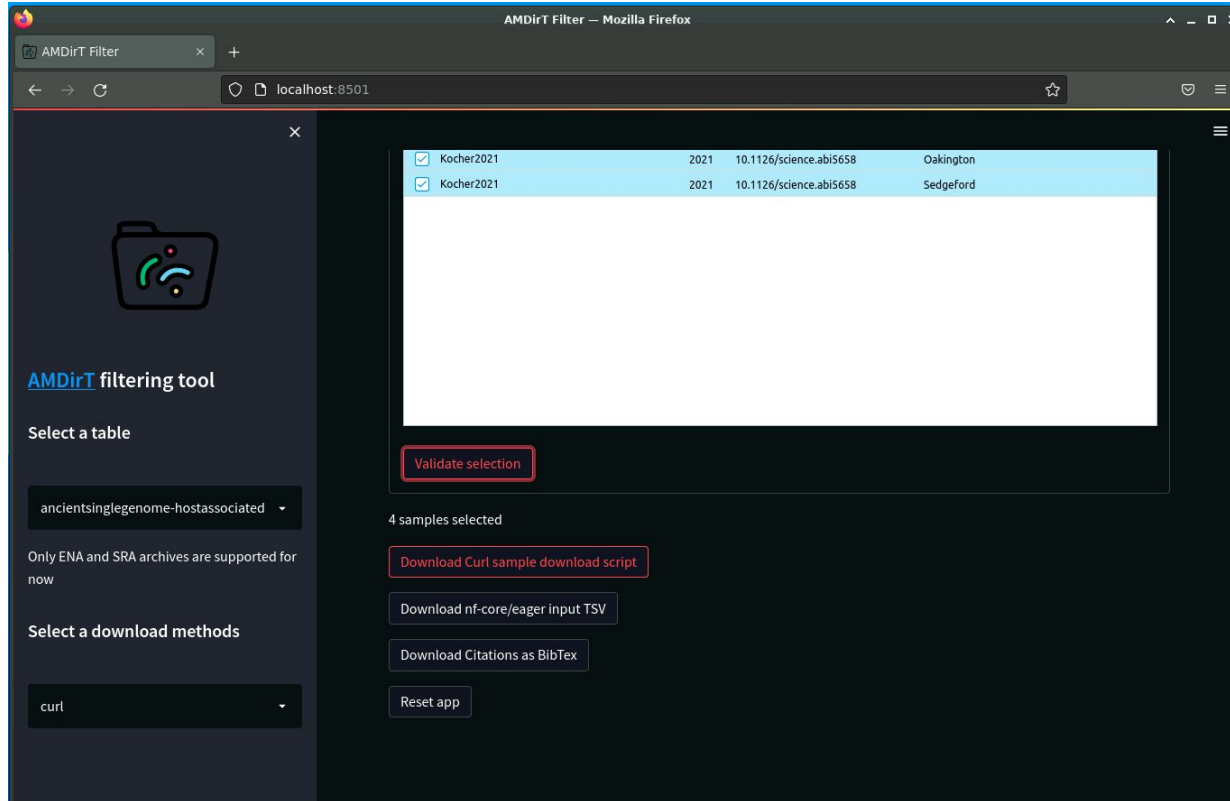
Select samples to filter

project_name	publication_year	publication_doi	site_name
<input checked="" type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Hinxton
<input checked="" type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Hinxton
<input checked="" type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Oakington
<input checked="" type="checkbox"/> Kocher2021	2021	10.1126/science.abi5658	Sedgeford

Validate selection



AMDirT: download download script



<input checked="" type="checkbox"/>	Kocher2021	2021	10.1126/science.abi5658	Oakington
<input checked="" type="checkbox"/>	Kocher2021	2021	10.1126/science.abi5658	Sedgeford

4 samples selected

Download Curl sample download script

Download nf-core/eager input TSV

Download Citations as BibTex

Reset app

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!)

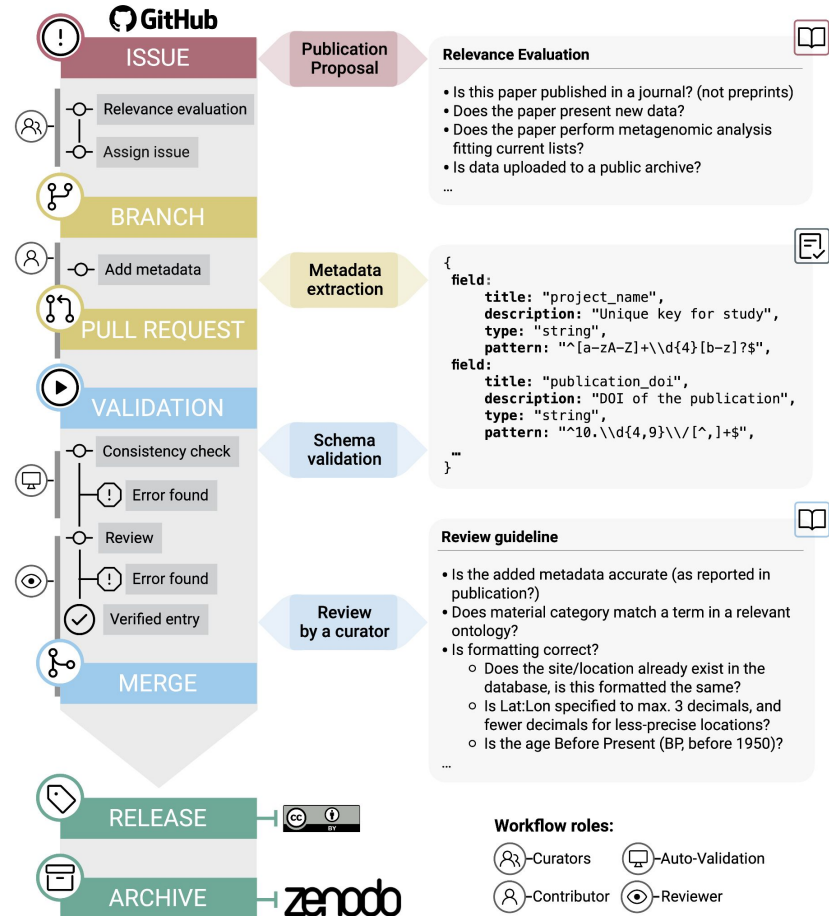
```
$ cd ~/Downloads  
$ cat ancientMetagenomeDir_eager_input.csv
```



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

