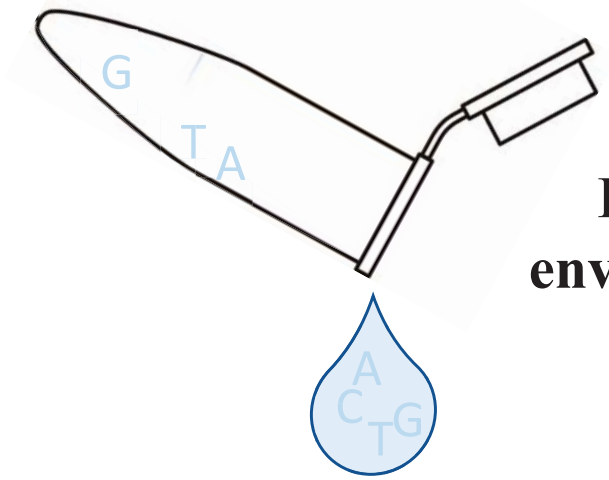


Integration of molecular data from environmental samples into the BioAtlas

Biodiversity Atlas Sweden is part of the **Swedish Biodiversity Data Infrastructure**, which aggregates Swedish biodiversity data and makes it freely available and usable online.

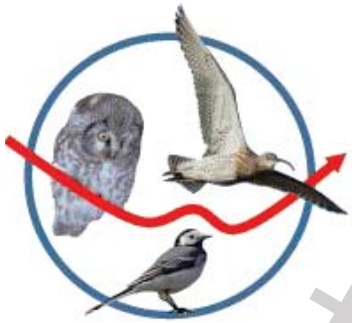
Stockholm
Maria Prager (SU/KI)
Anders Andersson (KTH)

Kalmar
Diego Brambilla (LnU)
Daniel Lundin (LnU)

Uppsala
Jeanette Tångrot (UU/UmU)
Anna Rosling (UU)

Biodiversity Atlas Sweden

Systematic monitoring data



Biotelemetry (tracking) data

BioAtlas

Natural history collection data

Molecular biodiversity data

Metabarcoding
Metagenomics

python™

Spatial portal

ALA4R...

Mirroreum: R online

eDNA in the SBDI
BioAtlas

Although massively parallel sequencing methods have revolutionized the collection of biodiversity data from environmental samples, metabarcoding data are rarely accessible for re-use.

To enable interpretation of fields and values not originally designed for environmental samples in 2018-2020 we will (1-4):

eDNA in the SBDI
BioAtlas

1. **provide a guide with specific pointers for metabarcoders,**
2. **supply a pipeline for automated processing of raw reads into denoised, taxon-annotated Amplicon Sequence Variants (ASVs) and**

eDNA in the SBDI
BioAtlas

3. **provide the necessary structures for integration of ASV observations into the Swedish BioAtlas aggregating biodiversity data and making them freely available on-line. Additionally, since**

eDNA in the SBDI
BioAtlas

4. **Species observations in the Bioatlas are currently mapped against the GBIF taxonomic backbone which unfortunately has poor coverage of some organisms, such as procaryotes, we thus aim to complement the GBIF taxonomic backbone with identifiers for Swedish ASVs and higher taxonomy from selected external databases.**

Biodiversity Atlas Sweden

Systematic monitoring data

Biotelemetry (tracking) data

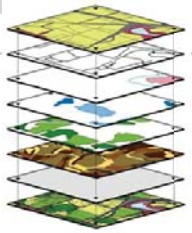
Natural history collection data

Molecular biodiversity data

Metabarcoding

Metagenomics

python™



Spatial portal

ALA4R...

Mirroreum: R online

Metabarcoding (marker-gene amplicon sequencing)

DNA extraction →

PCR →

DNA sequencing →

```
>seq1
AATGTAGTA…
>seq2
AACGTAGGT…
>seq3
...
```

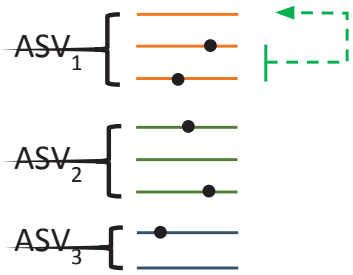
Sample (water, soil, faeces, …)

DNA

PCR-amplified genes

Error correction (denoising) ↓

Amplicon Sequence Variants (ASVs)



Matching of ASVs to reference database ←

with taxonomically classified sequences (e.g. Unite, Silva)

OTU/ASV	Species	Sample1	Sample2
1	E.coli	17	0
2	S.aurus	231	11800
3	unknown	30	0
…	…	…	…

ASV counts per sample

BAS

Overview of molecular data flows and BAS-MOL deliverables

1. Guide for simplified archiving of raw reads & metadata

Sequencing provider

NATIONAL GENOMICS INFRASTRUCTURE

Raw data repository

ENA
European Nucleotide Archive

Metadata

Lat:
Lon:
Temp:
Salinity:
...

Sequences

>seq1
AATGTAGTA...
>seq2
AACGTAGGT...
...

BioAtlas

Molecular module

Swedish ASV list

ASV	Region	PrimerFw	...	Barcode	Taxon
1	16S	AATGCT..	...	CTGAT..	*Bacillus*
2	16S	AATGCT..	...	CTGAT..	*Cytophaga*

2. Pipeline for denoising and taxonomic annotation of ASVs

Reference databases

silva high quality ribosomal RNA databases
UniEuk
GTDB
unite community

ASV table

ASV	Barcode	Taxon	Sample 1	Sample 2	...
1	CTGCGAT..	*Bacillus*	24	32	...
2	CTGACAT..	*Cytophaga*	654	712	...

3, 4. Integration of ASV observations & metadata into the BioAtlas

	Deliverable	2018	2019	2020	
D3.8 - A1	Submission service of sequencing- and metadata to INSDC databases				SU, UU, LnU, KTH, KI
D3.8 - A2	Service for mapping metabarcoding data to reference databases				SU, UU, LnU, KTH, KI
D3.8 - A3	Database of reference MOTUs for major Swedish biomes				SU, UU, LnU, KTH, KI
D3.8 - A4	Service for metagenome sequence annotation				SU, UU, LnU, KTH, KI
D3.8 - A5	Database with processed Swedish metagenome data				SU, UU, LnU, KTH, KI
D4.5 - A1	Integration of molecular data into ALA				SU, UU, LnU, KTH, KI

Overview of molecular data flows and BAS-MOL deliverables

1. Guide for simplified archiving of raw reads & metadata

Sequencing provider

NATIONAL GENOMICS INFRASTRUCTURE

Raw data repository

ENA European Nucleotide Archive

Metadata

Lat:
Lon:
Temp:
Salinity:
...

Sequences

>seq1
AATGTAGTA...
>seq2
AACGTAGGT...
...

BioAtlas

Molecular module

Swedish ASV list

ASV	Region	PrimerFw	...	Barcode	Taxon
1	16S	AATGCT..	...	CTGAT..	*Bacillus*
2	16S	AATGCT..	...	CTGAT..	*Cytophaga*

2. Pipeline for denoising and taxonomic annotation of ASVs

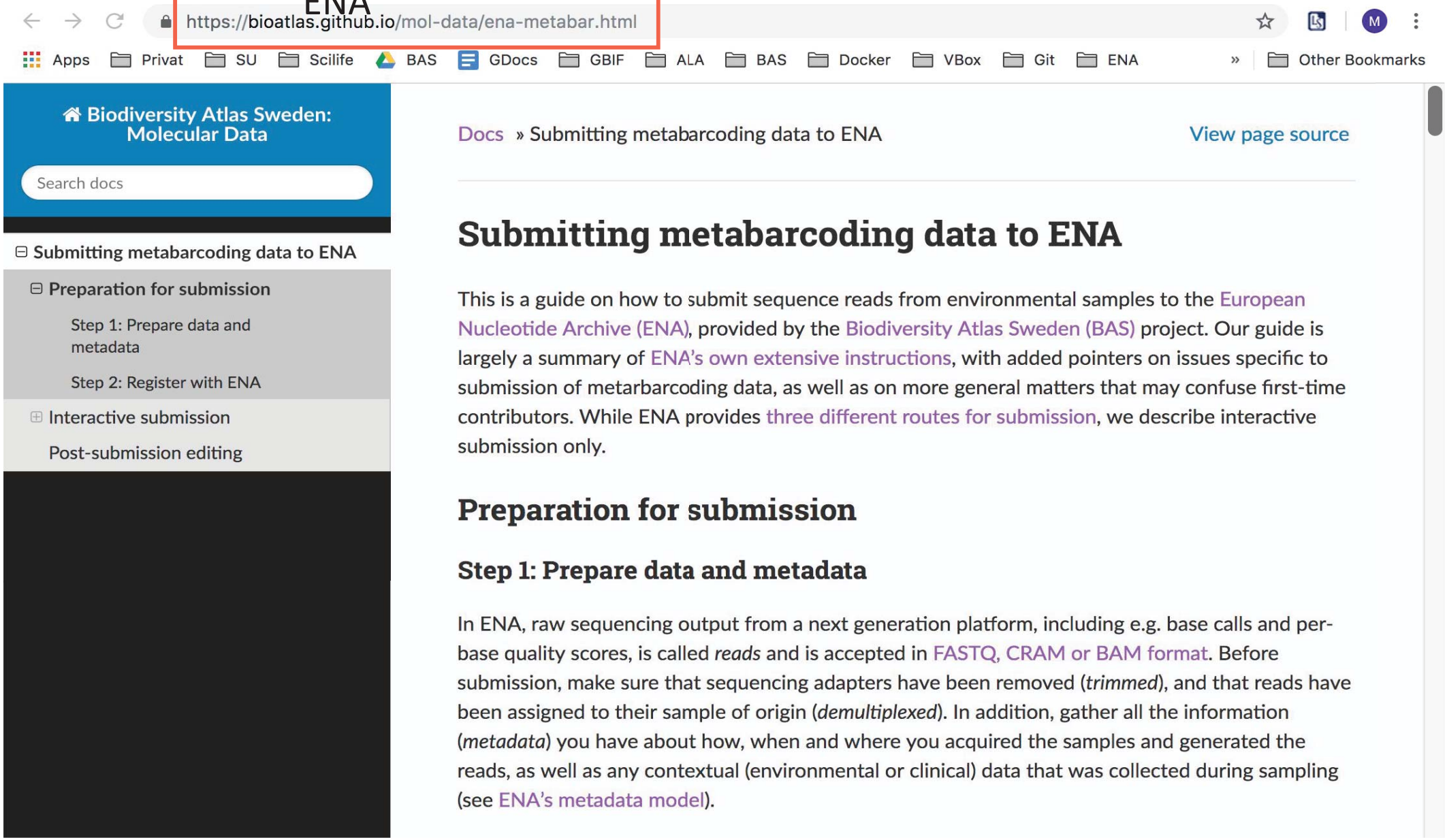
silva high quality ribosomal RNA databases · UniEuk · GTDB · unite community

ASV table

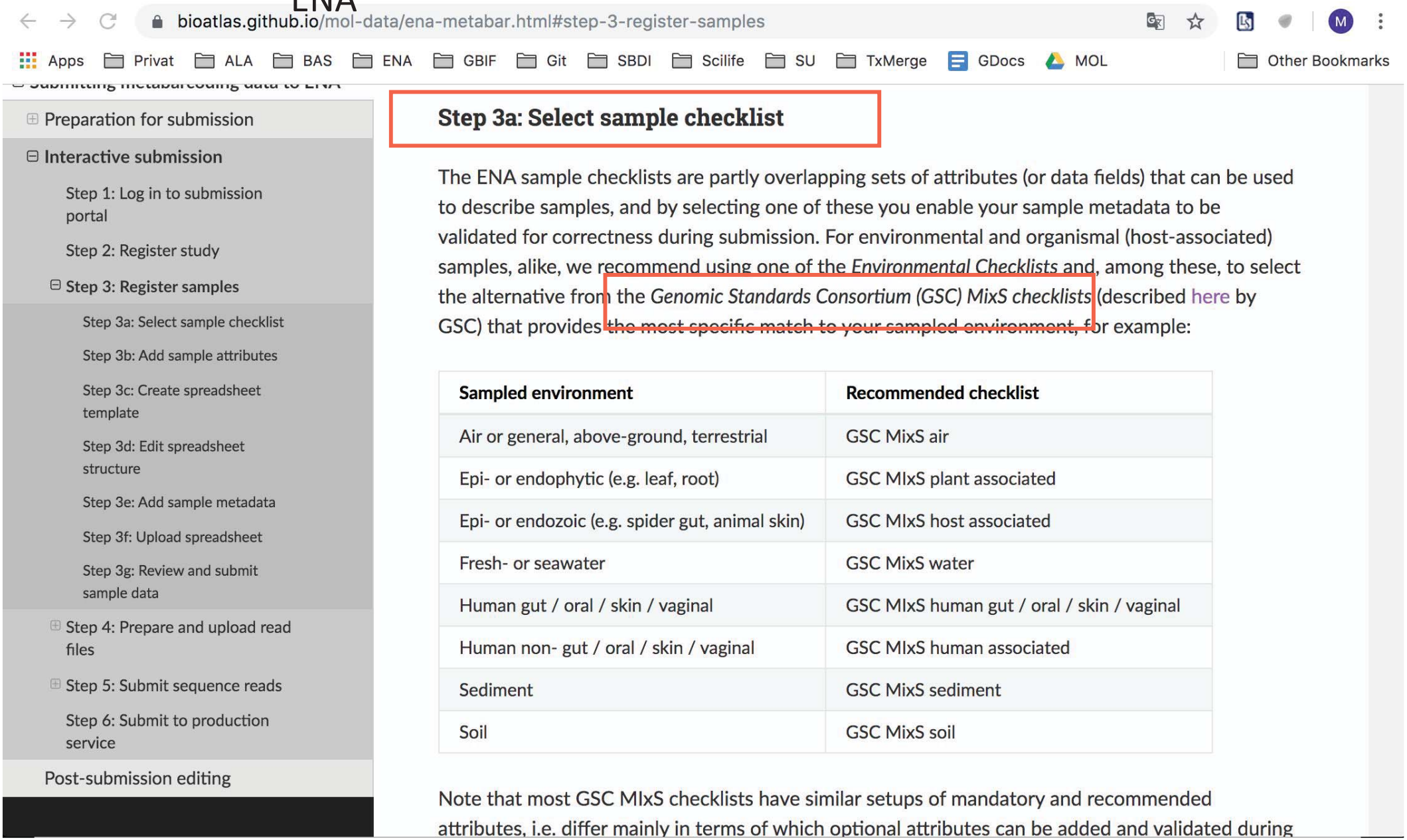
ASV	Barcode	Taxon	Sample 1	Sample 2	...
1	CTGCGAT..	*Bacillus*	24	32	...
2	CTGACAT..	*Cytophaga*	654	712	...

3, 4. Integration of ASV observations & metadata into the BioAtlas

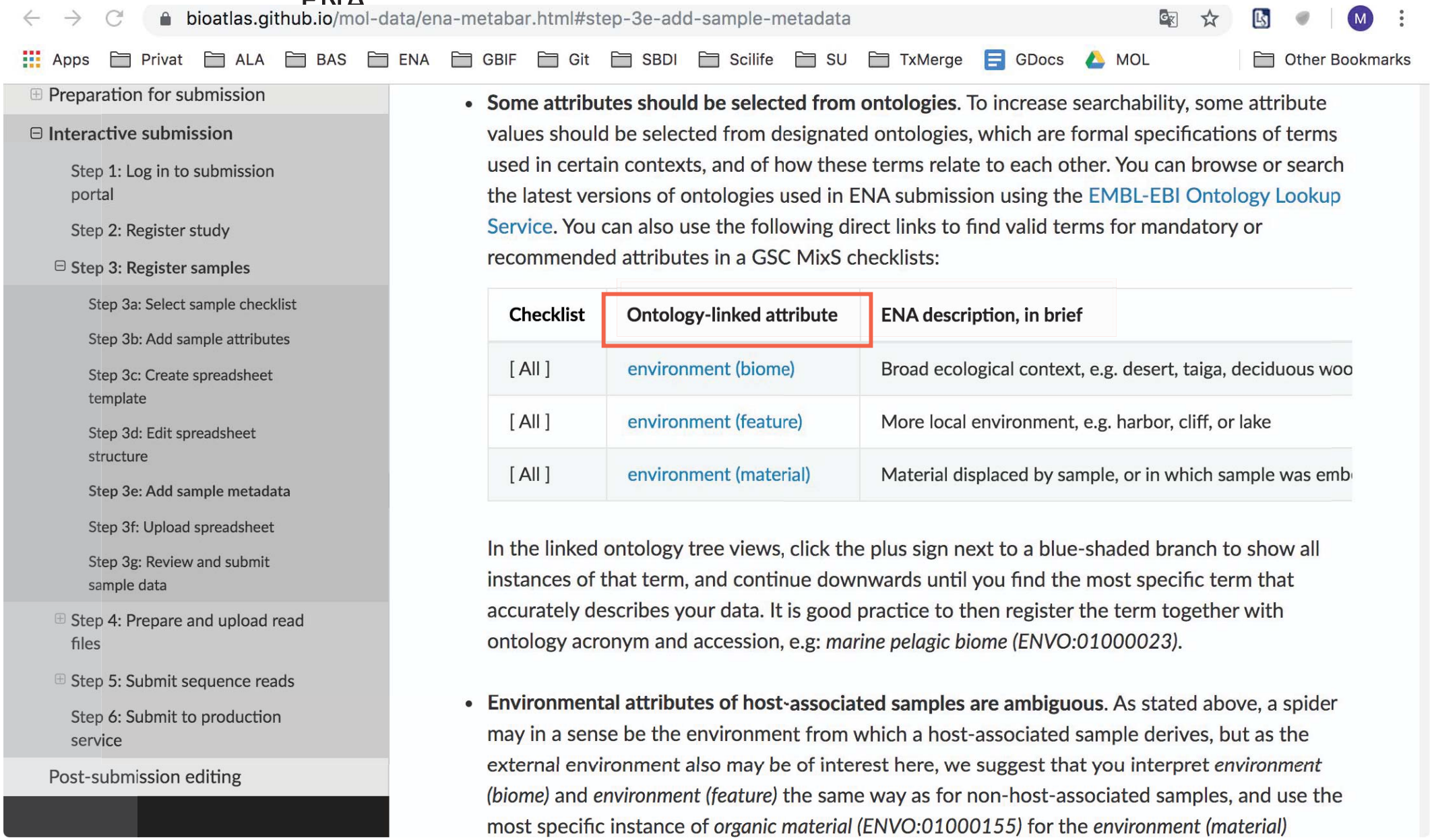
Guide for submission of metabarcoding data to ENA



Guide for submission of metabarcoding data to ENA



Guide for submission of metabarcoding data to ENA



Overview of molecular data flows and BAS-MOL deliverables

1. Guide for simplified archiving of raw reads & metadata

Sequencing provider

NATIONAL GENOMICS INFRASTRUCTURE

Raw data repository

ENA European Nucleotide Archive

Metadata

Lat:
Lon:
Temp:
Salinity:
...

Sequences

>seq1
AATGTAGTA...
>seq2
AACGTAGGT...
...

BioAtlas

Molecular module

Swedish ASV list

ASV	Region	PrimerFw	...	Barcode	Taxon
1	16S	AATGCT..	...	CTGAT..	*Bacillus*
2	16S	AATGCT..	...	CTGAT..	*Cytophaga*

2. Pipeline for denoising and taxonomic annotation of ASVs

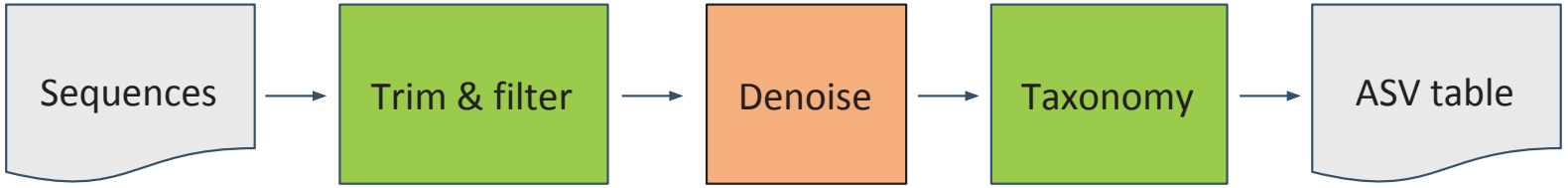
silva high quality ribosomal RNA databases UniEuk GTDB unite community

ASV table

ASV	Barcode	Taxon	Sample 1	...	...
1	CTGCGAT..	*Bacillus*	24	32	...
2	CTGACAT..	*Cytophaga*	654	712	...

3, 4. Integration of ASV observations & metadata into the BioAtlas

Ampliseq: An automated tool for amplicon processing



- DADA2 Denoising > Ampicon Sequence Variants (ASVs) - instead of clustering
  - Illumina MiSeq (v. 1.0) , PacBio (1.1)

- Taxonomy assignment using RDP classifier or IDTAXA

Organism group	Genomic region	Database	Pipeline v.
Prokaryotes	16S	SILVA or GTDB	1.0
Fungi	ITS	Unite	1.1
Protists	18S	UniEuk / PR2	1.x
Metazoa	COI	BOLD	1.x
Plants	rbcL, matK	BOLD	1.x

- Nextflow pipeline + Docker image > cluster compatible and reproducible

Parallel
Linear

Overview of molecular data flows and BAS-MOL deliverables

1. Guide for simplified archiving of raw reads & metadata

Sequencing provider

NATIONAL GENOMICS INFRASTRUCTURE

Raw data repository

ENA
European Nucleotide Archive

Metadata

Lat:
Lon:
Temp:
Salinity:
...

Sequences

>seq1
AATGTAGTA...
>seq2
AACGTAGGT...
...

BioAtlas

Molecular module

Swedish ASV list

ASV	Region	PrimerFw	...	Barcode	Taxon
1	16S	AATGCT..	...	CTGAT..	*Bacillus*
2	16S	AATGCT..	...	CTGAT..	*Cytophaga*

2. Pipeline for denoising and taxonomic annotation of ASVs

Reference databases

silva high quality ribosomal RNA databases
UniEuk
GTDB
unite community

ASV table

ASV	Barcode	Taxon	Sample 1	Sample 2	...
1	CTGCGAT..	*Bacillus*	24	32	...
2	CTGACAT..	*Cytophaga*	654	712	...

3, 4. Integration of ASV observations & metadata into the BioAtlas

Which taxonomy to use in the BioAtlas?

- How toprovide good taxonomic coverage & ASV resolution?
- Contributor's vs. 'Standardised' annotation?
- Single annotation vs. regular update (as reference databases grow)?

Reference databases

GTDB silva high quality ribosomal RNA databases UniEuk

international BARCODE OF LIFE unite community

GBIF

Taxonomy backbone

How (and why) make sequences available & searchable?

- No repository for processed ASVs, e.g. for prokaryotes
- ASV identifiers will (probably) be arbitrary
- Little support for sequences in GBIF / BioAtlas
- Solution: Molecular module + Swedish ASV list

Raw data repository

ENA
European Nucleotide Archive

BioAtlas

Molecular module

Swedish ASV list

ASV	Region	PrimerFw	...	Barcode	Taxon
1	16S	AATGCT..	...	CTGAT..	*Bacillus*
2	16S	AATGCT..	...	CTGAT..	*Cytophaga*

Adding data to BioAtlas

Searching data in BioAtlas (BLAST + http request?)

Exporting data from BioAtlas