

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

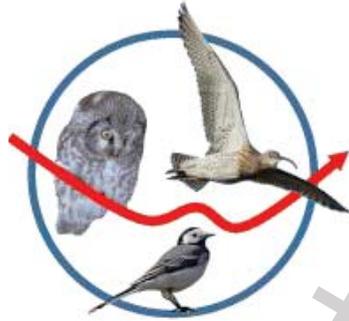
Diego Brambilla (LnU)
Daniel Lundin (LnU)

Uppsala

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



Systematic monitoring data



Biotelemetry (tracking) data



BioAtlas

Natural history collection data



Molecular biodiversity data



Metabarcoding
Metagenomics



python™

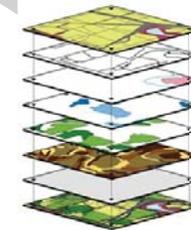


ALA4R...



OpenSci

Mirroreum: R online



Spatial
portal



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**

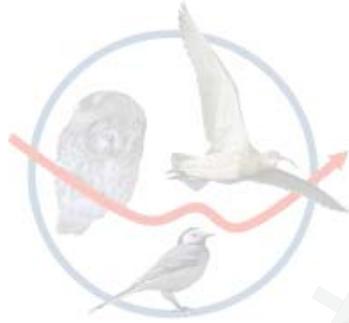


eDNA in the SBDI BioAtlas

- 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.**



Systematic monitoring data



Biotelemetry (tracking) data



Natural history collection data



Molecular biodiversity data



Metabarcoding
Metagenomics

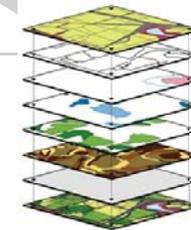


ALA4R...



OpenSci

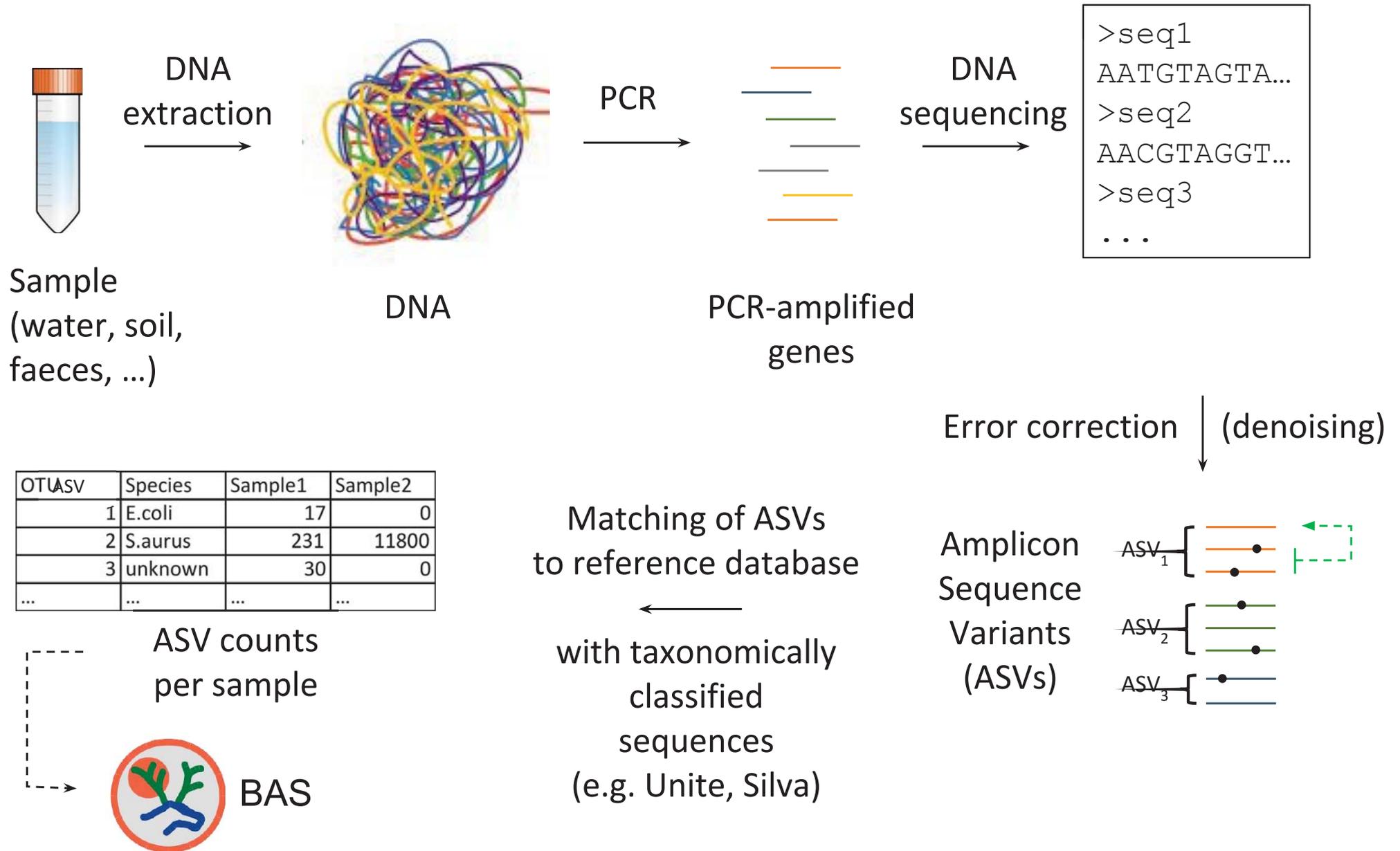
Mirroreum: R online



Spatial
portal



Metabarcoding (marker-gene amplicon sequencing)





Overview of molecular data flows and BAS-MOL deliverables

1. Guide for simplified archiving of raw reads & metadata



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..	<i>Bacillus</i>
2	16S	AATGCT..	...	CTGAT..	<i>Cytophaga</i>

2. Pipeline for denoising and taxonomic annotation of ASVs

Reference databases



ASV table

ASV	Barcode	Taxon	Sample 1	Sample 2	...
1	CTGCGAT..	<i>Bacillus</i>	24	32	...
2	CTGACAT..	<i>Cytophaga</i>	654	712	...

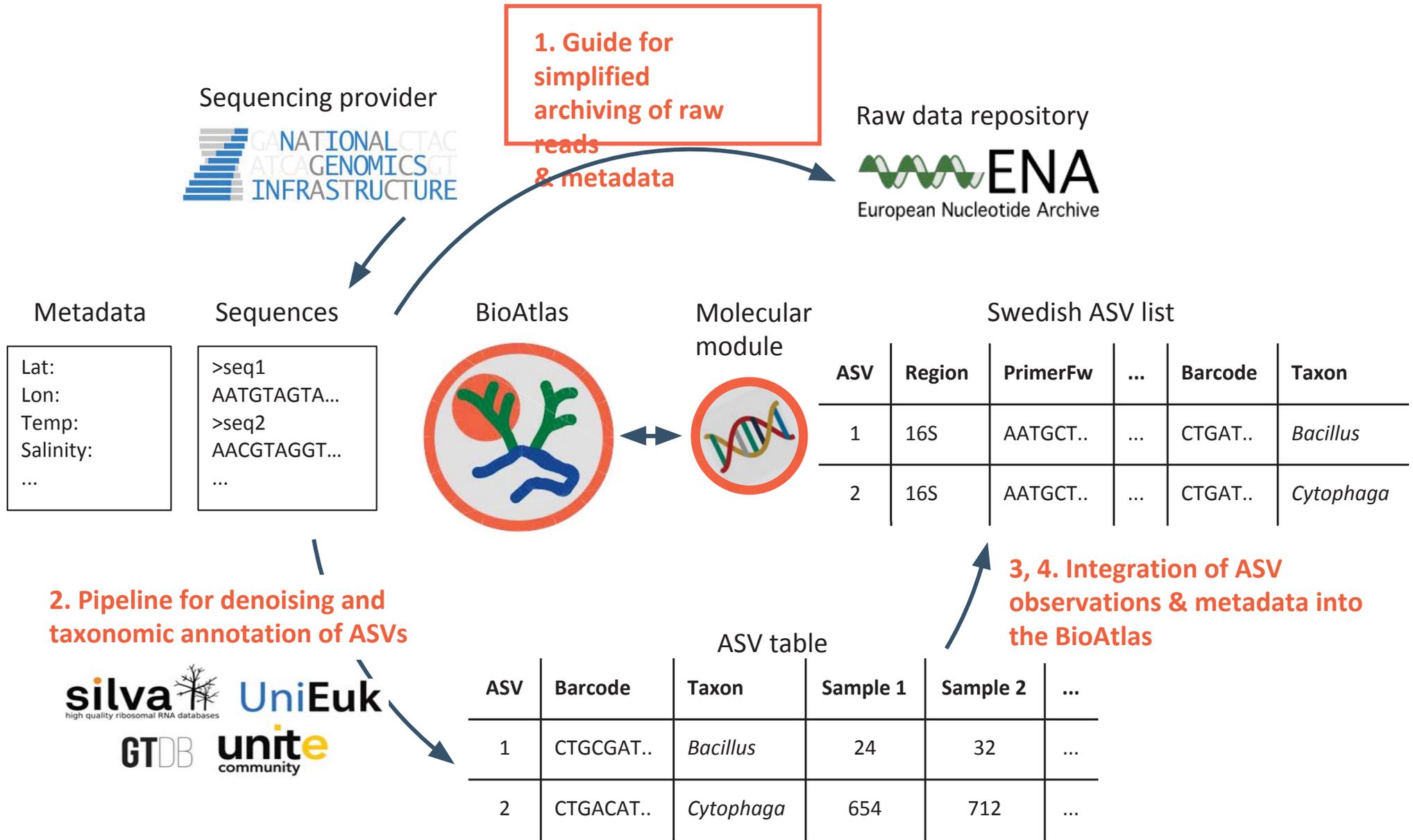
3, 4. Integration of ASV observations & metadata into the BioAtlas



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





Guide for submission of metabarcoding data to

ENA

https://bioatlas.github.io/mol-data/ena-metabar.html



Apps Privat SU Scilife BAS GDocs GBIF ALA BAS Docker VBox Git ENA

» Other Bookmarks

Biodiversity Atlas Sweden:
Molecular Data

Search docs

Submitting metabarcoding data to ENA

Preparation for submission

Step 1: Prepare data and metadata

Step 2: Register with ENA

Interactive submission

Post-submission editing

Docs » Submitting metabarcoding data to ENA

[View page source](#)

Submitting metabarcoding data to ENA

This is a guide on how to submit sequence reads from environmental samples to the [European Nucleotide Archive \(ENA\)](#), provided by the [Biodiversity Atlas Sweden \(BAS\)](#) project. Our guide is largely a summary of [ENA's own extensive instructions](#), with added pointers on issues specific to submission of metabarcoding data, as well as on more general matters that may confuse first-time contributors. While ENA provides [three different routes for submission](#), we describe interactive submission only.

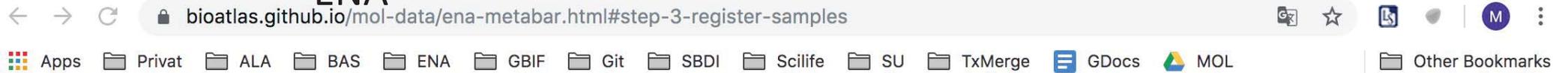
Preparation for submission

Step 1: Prepare data and metadata

In ENA, raw sequencing output from a next generation platform, including e.g. base calls and per-base quality scores, is called *reads* and is accepted in [FASTQ, CRAM or BAM format](#). Before submission, make sure that sequencing adapters have been removed (*trimmed*), and that reads have been assigned to their sample of origin (*demultiplexed*). In addition, gather all the information (*metadata*) you have about how, when and where you acquired the samples and generated the reads, as well as any contextual (environmental or clinical) data that was collected during sampling (see [ENA's metadata model](#)).



Guide for submission of metabarcoding data to ENA



Step 3a: Select sample checklist

The ENA sample checklists are partly overlapping sets of attributes (or data fields) that can be used to describe samples, and by selecting one of these you enable your sample metadata to be validated for correctness during submission. For environmental and organismal (host-associated) samples, alike, we recommend using one of the *Environmental Checklists* and, among these, to select the alternative from the *Genomic Standards Consortium (GSC) MixS checklists* (described [here](#) by GSC) that provides ~~the most specific match to your sampled environment~~, for example:

Sampled environment	Recommended checklist
Air or general, above-ground, terrestrial	GSC MixS air
Epi- or endophytic (e.g. leaf, root)	GSC MixS plant associated
Epi- or endozoic (e.g. spider gut, animal skin)	GSC MixS host associated
Fresh- or seawater	GSC MixS water
Human gut / oral / skin / vaginal	GSC MixS human gut / oral / skin / vaginal
Human non- gut / oral / skin / vaginal	GSC MixS human associated
Sediment	GSC MixS sediment
Soil	GSC MixS soil

Note that most GSC MixS checklists have similar setups of mandatory and recommended attributes, i.e. differ mainly in terms of which optional attributes can be added and validated during

- Preparation for submission
- Interactive submission
 - Step 1: Log in to submission portal
 - Step 2: Register study
 - Step 3: Register samples
 - Step 3a: Select sample checklist
 - Step 3b: Add sample attributes
 - Step 3c: Create spreadsheet template
 - Step 3d: Edit spreadsheet structure
 - Step 3e: Add sample metadata
 - Step 3f: Upload spreadsheet
 - Step 3g: Review and submit sample data
 - Step 4: Prepare and upload read files
 - Step 5: Submit sequence reads
 - Step 6: Submit to production service
- Post-submission editing



Guide for submission of metabarcoding data to

ENA

bioatlas.github.io/mol-data/ena-metabar.html#step-3e-add-sample-metadata

Apps Privat ALA BAS ENA GBIF Git SBDI Scilife SU TxMerge GDocs MOL Other Bookmarks

Preparation for submission

Interactive submission

Step 1: Log in to submission portal

Step 2: Register study

Step 3: Register samples

Step 3a: Select sample checklist

Step 3b: Add sample attributes

Step 3c: Create spreadsheet template

Step 3d: Edit spreadsheet structure

Step 3e: Add sample metadata

Step 3f: Upload spreadsheet

Step 3g: Review and submit sample data

Step 4: Prepare and upload read files

Step 5: Submit sequence reads

Step 6: Submit to production service

Post-submission editing

- **Some attributes should be selected from ontologies.** To increase searchability, some attribute values should be selected from designated ontologies, which are formal specifications of terms used in certain contexts, and of how these terms relate to each other. You can browse or search the latest versions of ontologies used in ENA submission using the [EMBL-EBI Ontology Lookup Service](#). You can also use the following direct links to find valid terms for mandatory or recommended attributes in a GSC MixS checklists:

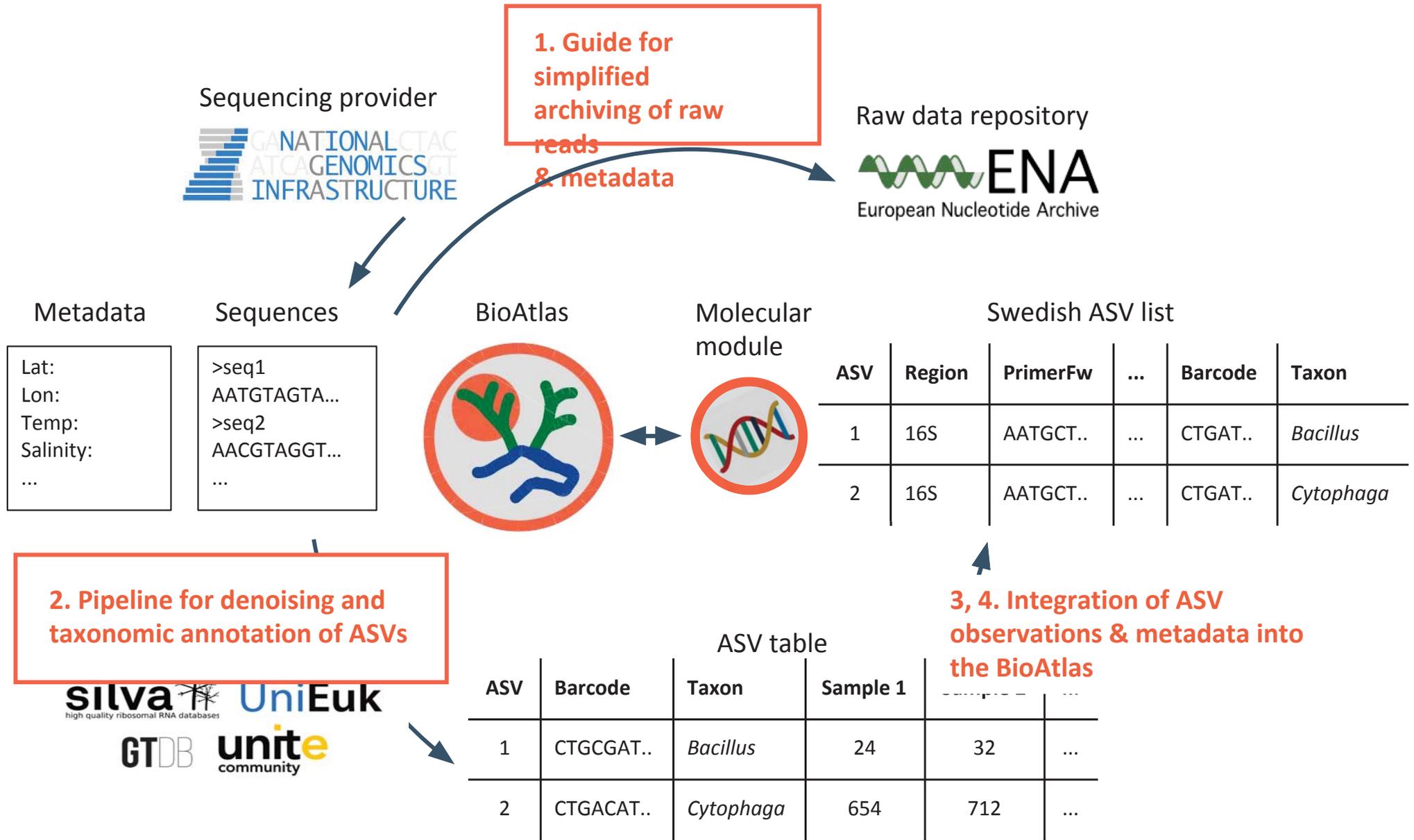
Checklist	Ontology-linked attribute	ENA description, in brief
[All]	environment (biome)	Broad ecological context, e.g. desert, taiga, deciduous wood
[All]	environment (feature)	More local environment, e.g. harbor, cliff, or lake
[All]	environment (material)	Material displaced by sample, or in which sample was emb

In the linked ontology tree views, click the plus sign next to a blue-shaded branch to show all instances of that term, and continue downwards until you find the most specific term that accurately describes your data. It is good practice to then register the term together with ontology acronym and accession, e.g: *marine pelagic biome* (ENVO:0100023).

- **Environmental attributes of host-associated samples are ambiguous.** As stated above, a spider may in a sense be the environment from which a host-associated sample derives, but as the external environment also may be of interest here, we suggest that you interpret *environment (biome)* and *environment (feature)* the same way as for non-host-associated samples, and use the most specific instance of *organic material* (ENVO:01000155) for the *environment (material)*

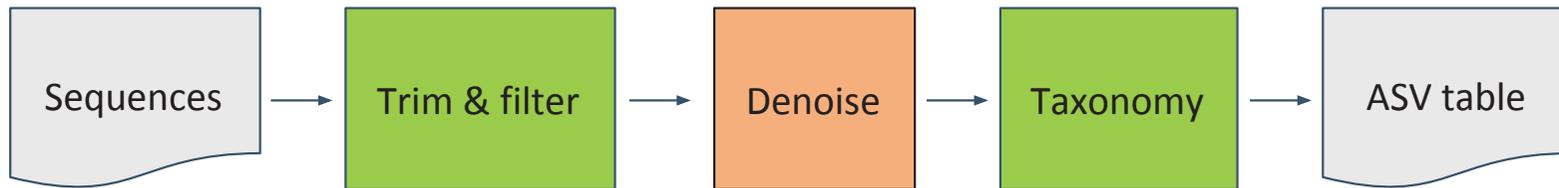


Overview of molecular data flows and BAS-MOL deliverables





Ampliseq: An automated tool for amplicon processing



- DADA2 Denoising > Amplicon 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





Overview of molecular data flows and BAS-MOL deliverables

1. Guide for simplified archiving of raw reads & metadata



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..	<i>Bacillus</i>
2	16S	AATGCT..	...	CTGAT..	<i>Cytophaga</i>

2. Pipeline for denoising and taxonomic annotation of ASVs

Reference databases



ASV table

ASV	Barcode	Taxon	Sample 1	Sample 2	...
1	CTGCGAT..	<i>Bacillus</i>	24	32	...
2	CTGACAT..	<i>Cytophaga</i>	654	712	...

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



Which taxonomy to use in the BioAtlas?

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

Reference databases



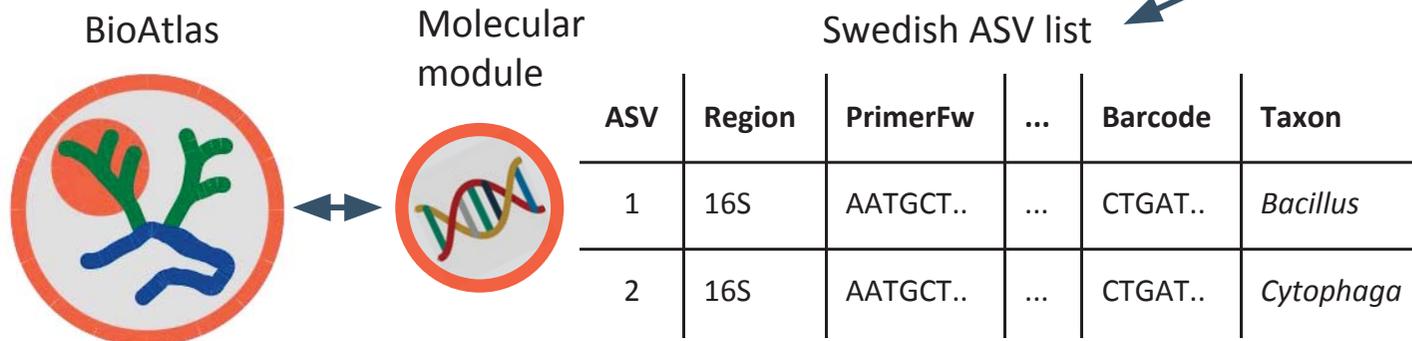


How (and why) make sequences available & searchable?

- 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

 European Nucleotide Archive



Adding data to BioAtlas

Searching data in BioAtlas (BLAST + http request?)

Exporting data from BioAtlas