

National Institute of Allergy and Infectious Diseases

Introduction to **METAGENOTE**

Webinar, March 24th, 2020

Mariam Quiñones, Ph.D.
Head, Science Support Section
Bioinformatics and Computational
Biosciences Branch/OCICB/NIAID



U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES NATIONAL INSTITUTES OF HEALTH NATIONAL INSTITUTE OF ALLERGY AND INFECTIOUS DISEASES BIOINFORMATICS @NIAID

METAGENOTE BROWSE USER GUIDE ABOUT Contact Us

METAGENOTE is a quick and intuitive way to annotate data from genomics studies including microbiome.

[Start Here!](#)

Why use METAGENOTE?

Annotate
Fully describe samples from which genomic sequences have been obtained.

Use Standards
Follow guidelines from the [Genomics Standards Consortium \(GSC\)](#) standards for ease of reproducibility.

Store & Search
Organize metadata into studies, projects and sample groups. Browse existing metadata.

Publish
Validate metadata and automatically publish to the [NCBI Sequence Read Archive \(SRA\)](#).

<https://metagenote.niaid.nih.gov/>

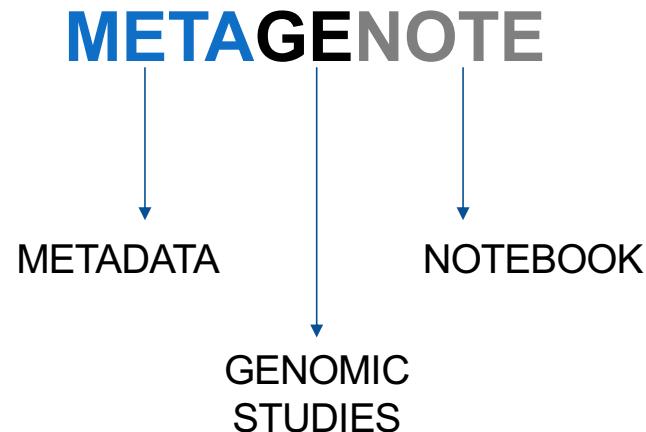
NIAID

Webinar agenda

1. What is **METAGENOTE**?
2. Importance of metadata annotation
3. Main features of **METAGENOTE**
4. Demo of annotation and submission to NCBI's SRA

1- What is METAGENOTE?

- It is web portal that greatly facilitates the annotation of samples from genomic studies and streamlines the submission process of sequencing files and metadata to the NCBI's SRA repository for public access.



What do we mean by **metadata** of a Genomic Sample?

The “**What**”, **How**”, “**Where**”, “**When**” of biological sample and processing



BioSample

- Host: *Macaca mulatta*
- Organism: gut metagenome
- Host body product: Stool
- Collection date: 16 Nov 2019
- Chemical administration: Vancomycin CHEBI:28001
- Environmental feature: animal house ENVO:00003040

SRA

- Sequencing methods: Illumina MiSeq

Why developing METAGENOTE?

“We need a way to annotate samples that is better than an Excel spreadsheet”



Dr. Yasmine Belkaid,
LPD and CHI/NIAID



National Institute of
Allergy and
Infectious Diseases

- ✓ Metadata for all studies should be kept in one system
- ✓ There should be unification in annotation
- ✓ It should be easy to make metadata public at time of manuscript submission as required by the NIH Data Sharing policy

NIAID

2- Why is it important to annotate and publish sample metadata?

To facilitate:

- ✓ Reproducible research
- ✓ Reuse of data
- ✓ Integration of multiple studies & meta-analyses

Where is METADATA normally stored?

Repositories such as: NCBI, ENA, MG-RAST, RAST

NCBI Example 1: WGS of *Borrelia miyamotoi* isolated from blood plasma

▼ Hide common fields

Example 2: miRNA-Seq from liver tissue of HIV patients after treatment with Simtuzumab

Example 3: WGS of a Dog genome

▼ Hide common fields										
AnalysisName	Phylo: PER00816_x									
Assay Type:	WGS									
AssemblyName:	GCF_000002285.3									
AvgSpotLen:	200									
BioProject:	PRJNA448733									
BioSample:	SAMN08873519									
BioSampleModel:	Model organism or animal									
Center Name:	NHGRI									
Consent:	public									
CoverageAll:	16.0582									
CoverageRatio:	1.6783218244964209									
CoverageX:	9.56801									
DATASTORE filetype:	sra									
DATASTORE provider:	ncbi									
Experiment:	SRX4041934									
InsertSize:	0									
Instrument:	Illumina HiSeq 2500									
LibraryLayout:	PAIRED									
LibrarySelection:	RANDOM									
LibrarySource:	GENOMIC									
Library Name	Organism	Phylo Results	ReleaseDate	age	biomaterial provider	breed	dev stage	sex	sub species	t
LoadDate	Canis lupus familiaris	Rottweiler	2018-05-10	missing	Dr. Susannah Sample, The Comparative Orthopaedic Research Laboratory, School of Veterinary Medicine...	Rottweiler	not applicable	male	familiaris	
MBases:	Canis lupus familiaris	Rottweiler	2018-05-10	missing	Dr. Susannah Sample, The Comparative Orthopaedic Research Laboratory, School of Veterinary Medicine...	Rottweiler	not applicable	male	familiaris	
MBytes:	Canis lupus familiaris	Rottweiler	2018-05-14	missing	Dr. Susannah Sample, The Comparative Orthopaedic Research Laboratory, School of Veterinary Medicine...	Rottweiler	not applicable	female	familiaris	
Organism	Canis lupus familiaris	Portuguese Water Dog	2018-05-16	missing	The Dog Genome Project, National Human Genome Research Institute, National Institutes of Health, Be...	Portuguese Water Dog	not applicable	female	familiaris	
Phylo Boc	Canis lupus familiaris	Portuguese Water Dog	2018-05-16	missing	The Dog Genome Project, National Human Genome Research Institute, National Institutes of Health, Be...	Portuguese Water Dog	not applicable	male	familiaris	
Phylo Res	Canis lupus familiaris	Portuguese Water Dog	2018-05-16	missing	The Dog Genome Project, National Human Genome Research Institute, National Institutes of Health, Be...	Portuguese Water Dog	not applicable	male	familiaris	
Platform:	Canis lupus familiaris	Portuguese Water Dog	2018-05-16	missing	The Dog Genome Project, National Human Genome Research Institute, National Institutes of Health, Be...	Portuguese Water Dog	not applicable	male	familiaris	
ReleaseD	Canis lupus familiaris	Portuguese Water Dog	2018-05-16	missing	The Dog Genome Project, National Human Genome Research Institute, National Institutes of Health, Be...	Portuguese Water Dog	not applicable	male	familiaris	
Run:	Canis lupus familiaris	Portuguese Water Dog	2018-05-16	missing	The Dog Genome Project, National Human Genome Research Institute, National Institutes of Health, Be...	Portuguese Water Dog	not applicable	female	familiaris	
SRA Sam	Canis lupus familiaris	Portuguese Water Dog	2018-05-16	missing	The Dog Genome Project, National Human Genome Research Institute, National Institutes of Health, Be...	Portuguese Water Dog	not applicable	male	familiaris	
SRA Study:	SRP144493									
Sample Name:	PER00816									
age:	missing									
biomaterial provider:	Waltham Science, Waltham on the Wolds, UK									
breed:	Yorkshire Terrier									
dev stage:	not applicable									
sex:	male									
sub species:	familiaris									

BioSample Model “Package”

- Provides template
- Enforces a minimum set of attributes

*Other Models include those developed by Genomics Standards Consortium (e.g. MIMS, MIMARKS, MIGS)

Published
metadata is
very
inconsistent,
we need to
do better!

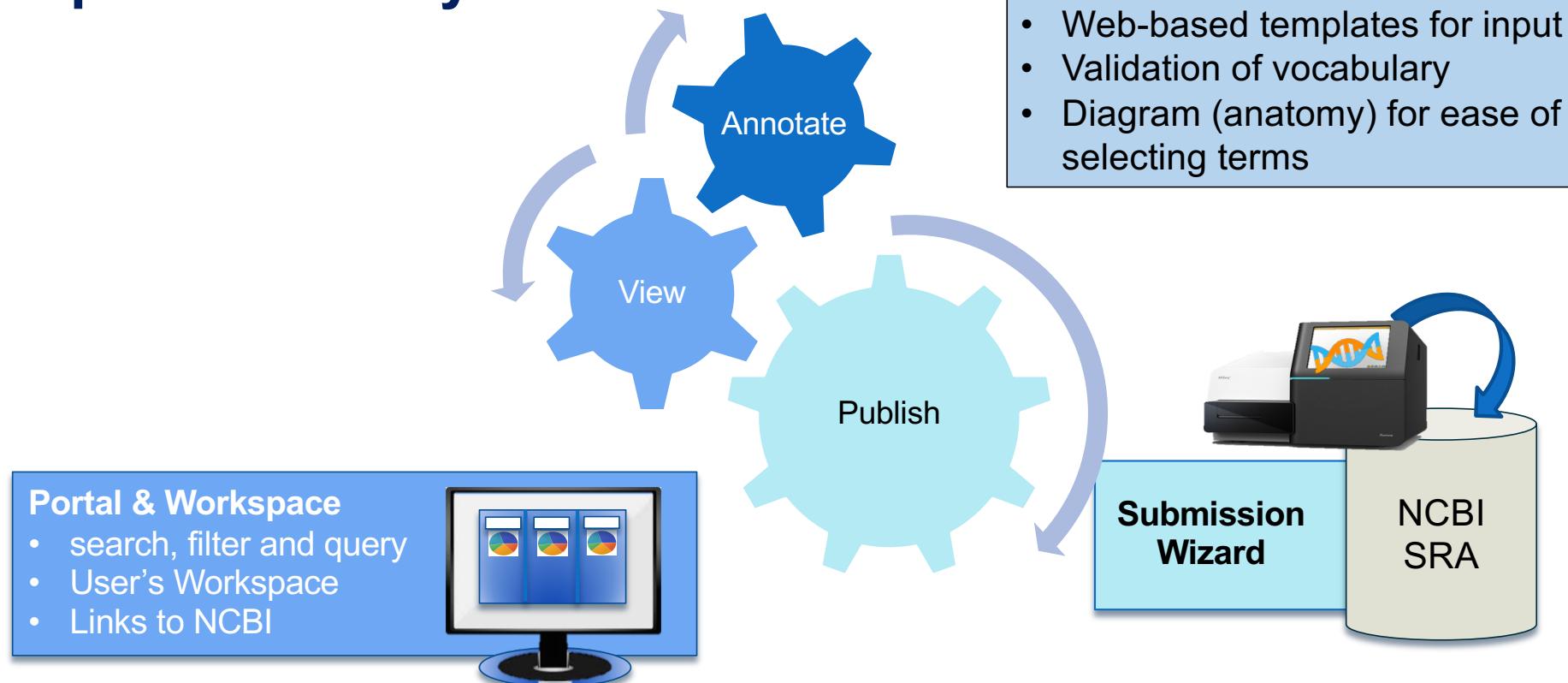
Table 1: Example of vocabulary used in BioSamples records

Package:	MIGS.ba.host-associated.5.0		MIMARKS.specimen.host-associated.5.0		Model.organism.animal.1.0	
total entries:	2598		19338		276262	
attribute:	host tissue sampled		host diet		developmental stage (required)	
number of responses	94		3616		179373	
metadata provided	27	intestine	358	herbivore	41112	adult
	16	gill	206	CHO	18309	missing
	8	Gill	199	therapeutic_hydrolyzed_protein_diet	16856	not collected
	6	spleen	191	Chow	16416	Adult
	5	nasopharynx	144	Artemia_salina	16209	not applicable
	5	Mammary gland	106	KD	5465	Juvenile
	4	Sputum	105	regulat_diet	4797	juvenile
	3	bronchoalveolar lavage	103	Unmedicated_swine_feed	2315	NA
	2	Tissue	91	LFD	2123	Embryo
	2	pharynx	90	purified_rat_diet_AIN-93M	2098	mixed
	2	head kidney	90	10_%_sucrose_solutio_n	1511	mixed stage
	1	Whole	88	HFD	1401	mature
	1	submandibular lymph node	88	finely_ground_Koi_food_(Foster_&_Smith,_Inc._Rhinelander,_WI)	1160	larva
	1	stomach	84	10_%_w/w_rice_bran_fermented_with_Bifidobacterium_longum_A TCC-55813	1077	MISSING
	1	skin lesion	79	10_%_w/w_heat-stabilized_rice_bran	1014	tadpole
	1	Pleural fluid	67	Termitomyces	988	adult pig
	1	periodontal pocket	64	HC	947	months postnatal

Problem and Proposed Solution

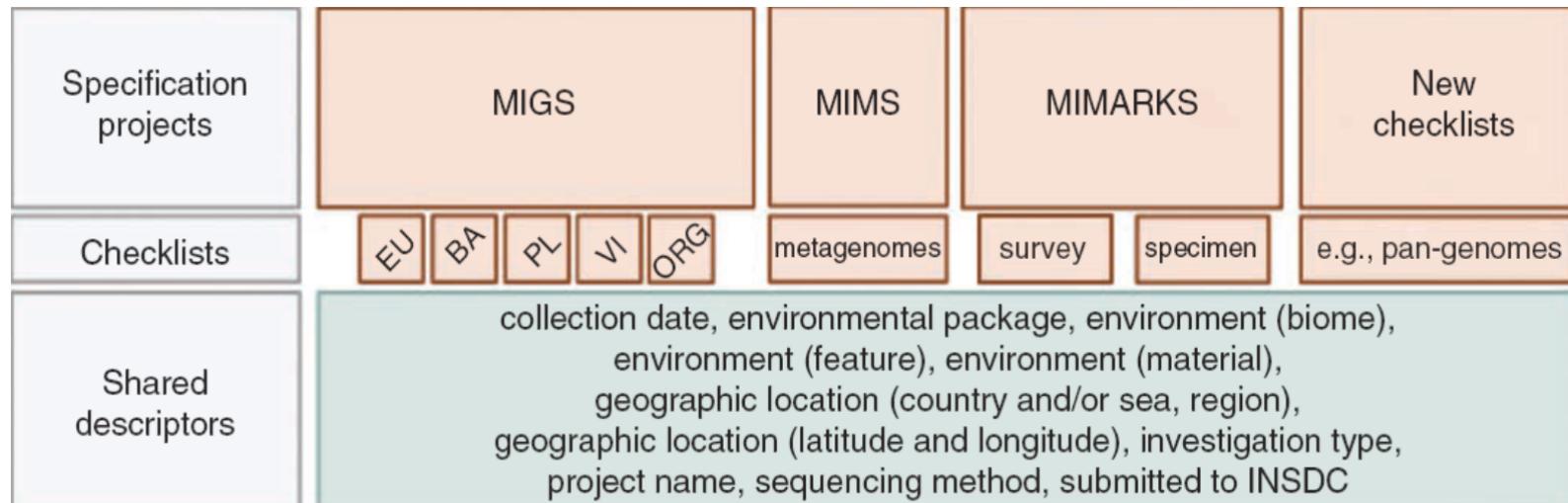
- Standards are crucial for cross-study comparisons but cumbersome to use
- Our goal is to build a system that streamlines sample metadata annotation while leveraging the existing ontologies and standards recognized by NCBI Sequence Read Archive
- We also aim to simplify the transfer of sequencing files by providing a fully web browser solution, → no more FTP!

3- METAGENOTE's top functionality



GSC: WIDELY USED STANDARD VOCABULARY FOR MICROBIOME SEQUENCE DATA

- MIMARKS - Minimum Information about a MARKer gene Sequence Project
- MIMS - metagenome
- MIGS - genome

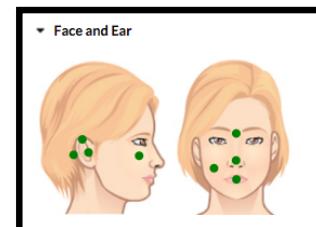
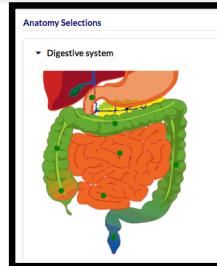
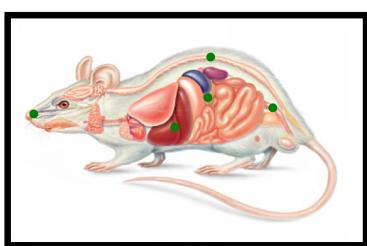


METAGENOTE's Notebook provide templates for various model types

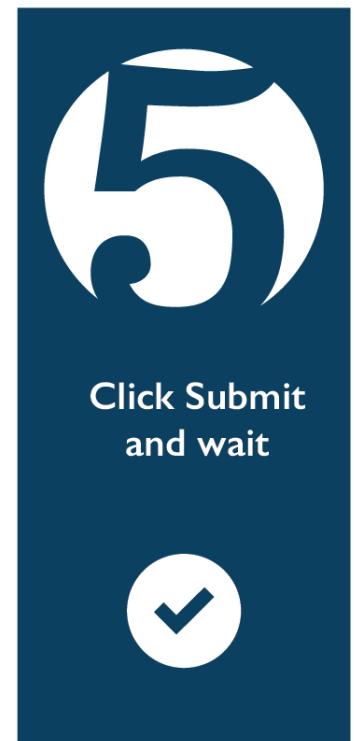
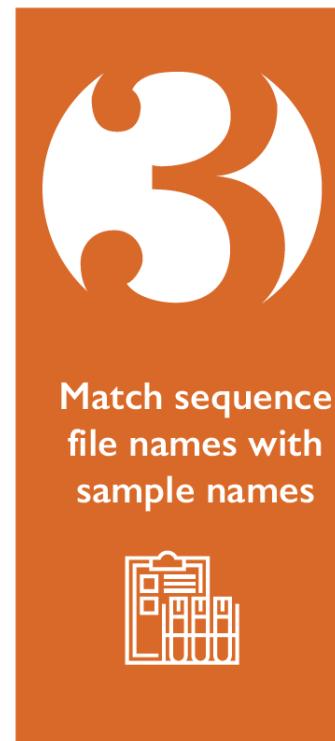
Select a template appropriate for your sample group, then annotate

The screenshot shows the METAGENOTE interface. At the top, there are six circular icons representing different model types: Microbiome (survey or WGS), Cultured bacteria/archaea, Virus, Human, Eukaryote, and Model organism (represented by a fly). To the right of these icons are three boxes: 'Model type' (containing 'Microbiome (survey or WGS)'), 'Ontologies' (containing 'Cultured bacteria/archaea'), and 'Description' (containing 'Virus'). A large blue arrow points downwards from this row to the main workspace. The workspace contains a table titled 'Samples' with columns for 'Sample Name*', 'Host Age', 'Host Body Mass Index', and 'Host Body Product'. The 'Host Body Product' column shows values like 'Feces FMA:64183' repeated for each sample. To the right of the table is a 'Descriptions' panel with a search bar and sections for 'Host Age' and 'Host Body Mass Index'. The 'Host Body Mass Index' section includes a definition: 'body mass index of the host, calculated as weight/(height)squared'. The overall interface has a clean, modern design with a blue and white color scheme.

"Host tissue samples" has an anatomy diagram



After annotation is complete, METAGENOTE provides users with a 5 steps Submission Wizard that moves metadata and files to NCBI's Sequence Read Archive (SRA)

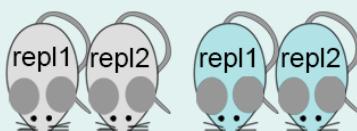


Anatomy of SRA submission

BioProject:
description
of research
project

BioSample:
description of
biological
samples

Project title: Transcriptome analysis of hepatotoxicity induced by botulin in mice Transcriptome of flowering plant Metagenome of chlorophyll-containing microbiome in Norwegian lake Mapping and manipulating E. coli transcriptome using antibiotics



Sample type:
Organism:



Plant sample
Fancypsis pretticus

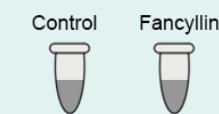


Metagenome or environmental sample
Lake water metagenome



Microbe sample
Escherichia coli

Sample name:



Other NCBI
databases

SRA:

metadata

sequence data files



Library_id: Cntr1 Cntr2 Botulin Illum Roche Light Dark Cntr Fancyllin

Title; Library: strategy, source, selection, layout; Platform; Instrument model; Design description; filetype; filenames

.bam, .fastq, .sff, .h5, fasta

<https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/>

DEMO

➤ Visit <https://metagenote.niaid.nih.gov> and register today!

1. Overview of interface and user's workspace
2. Basic annotation of 2 samples (human oral)
 - a) Use of ontology (Environmental material, tissue sampled)
 - b) Use of default row
 - c) Use of host body site diagram
3. Batch Import of annotations
4. 5-step submission to NCBI SRA
5. Guest submission

METAGENOTE's BCBB Team



Lewis Kim
Scrum master



David Liu
Lead developer



Conrad Shyu
Developer



Mariam Quiñones
Product Owner/SME

Contact us: metagenote@mail.nih.gov

Thanks!!

For additional questions, contact us at:
metagenote@niaid.nih.gov

**For other webinars and events,
visit the NIAID Bioinformatics Portal
<https://bioinformatics.niaid.nih.gov/events>**

Example of notification when submission was successful

Success!

Your submission has been sent to SRA.

Thank you for your submission. Your submission has been successfully forwarded to NCBI Sequence Read Archive (SRA). NCBI will evaluate to accept or reject your submission. We will notify you as soon as possible once we hear from NCBI.



Meanwhile, please feel free to contact us at METAGENOTE@nih.gov if you have any questions. Please refer to the METAGENOTE publish id below for this submission

METAGENOTE publish ID: R3HZ9BMT3VC

Thank you.

OK

Example of email notification when submission was successful

METAGENOTE Team <metagenote_submission@mail.nih.gov> 12:03 PM (4 minutes ago)   
to me ▾

Dear METAGENOTE User,

We are processing your submission of sample group **R3HZ9BMT3VC** to the Sequence Read Archive (SRA). This submission relates to the project entitled **Demo microbiome oral NIH**.

What to expect next:

1. If this is your first submission using email quinonesmwk@gmail.com, the NCBI Submission Portal team will contact you to **verify submission ownership**. Please follow the enclosed instructions to confirm ownership.
2. The **NCBI BioProject Submissions Staff** will email updates regarding the progression of your samples as additions to the BioSample database.

Updates from NCBI should follow within 24 hours. If you do not receive these notifications or have any questions, reply to this email or send us a message at metagenote_submission@mail.nih.gov.

Thank you for using METAGENOTE.

The METAGENOTE Team

Example of email notification when NCBI offers user the option to claim ownership

Submission ownership transfer

NN

NLM NCBI nobody

To: Quinones, Mariam (NIH/NIAID) [E]

Your submission by niaid-bcbb-tbp-srv successfully processed.

Important: You must take ownership of this submission in order to make updates to it. You will need to create or log into an NCBI account.

To continue, please click on this link:

<https://dsubmit.ncbi.nlm.nih.gov/invites/RvezGt7RWdR2rPeeuHn6h4vZp7S9vtlRYzfys24BekcZN6DRY8wWK6eMSwNvde3z/>

Please do not reply to this email.

Sincerely,
NCBI Submission Portal team