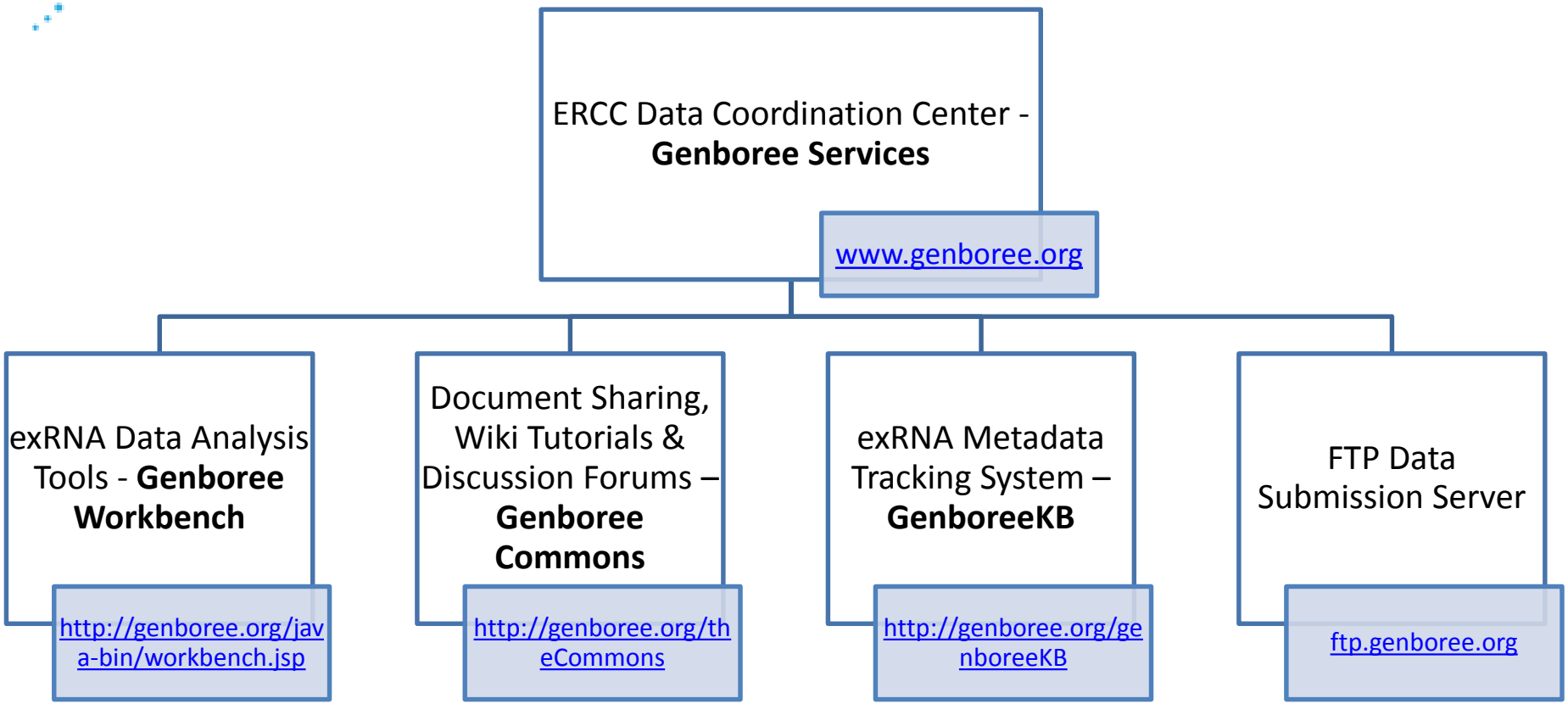
ERCC Data Submission Demo Series

Small exRNA Profiling Data & Metadata Submission

Sai Lakshmi Subramanian
William Thistlethwaite
Data Coordination Center (DCC)
Data Management and Resource Repository (DMRR)

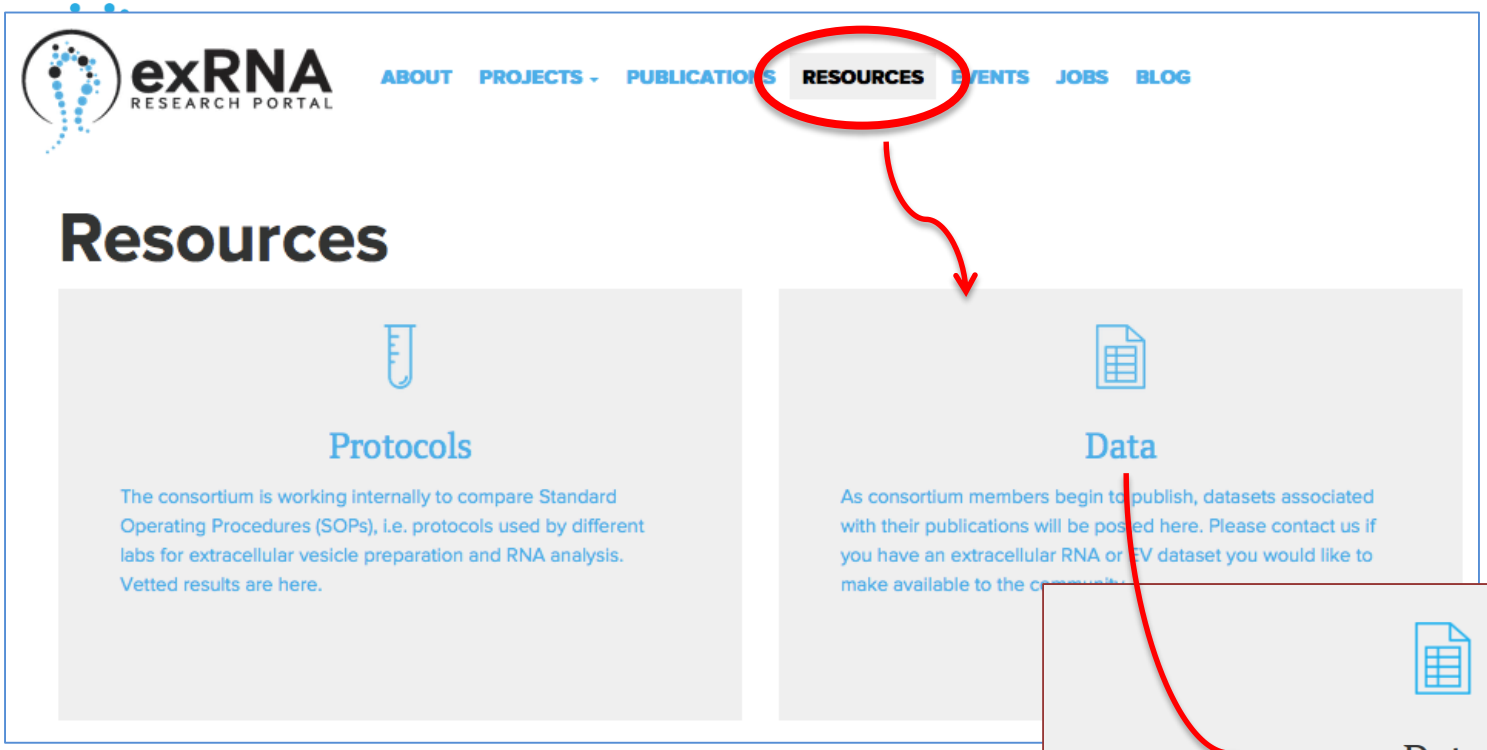
Wednesday, 9th September, 2015

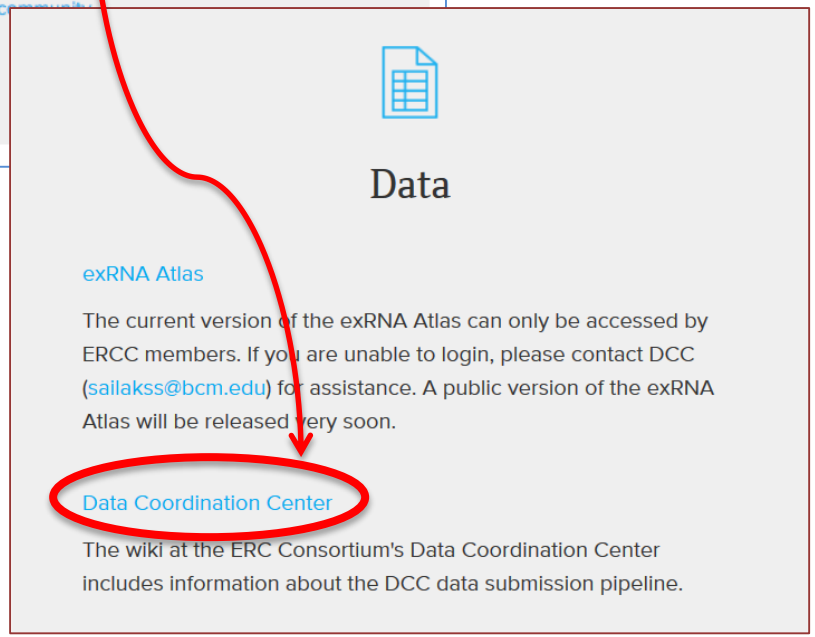
Summary of Genboree Services



Use the same user name and password for all of these Genboree Services

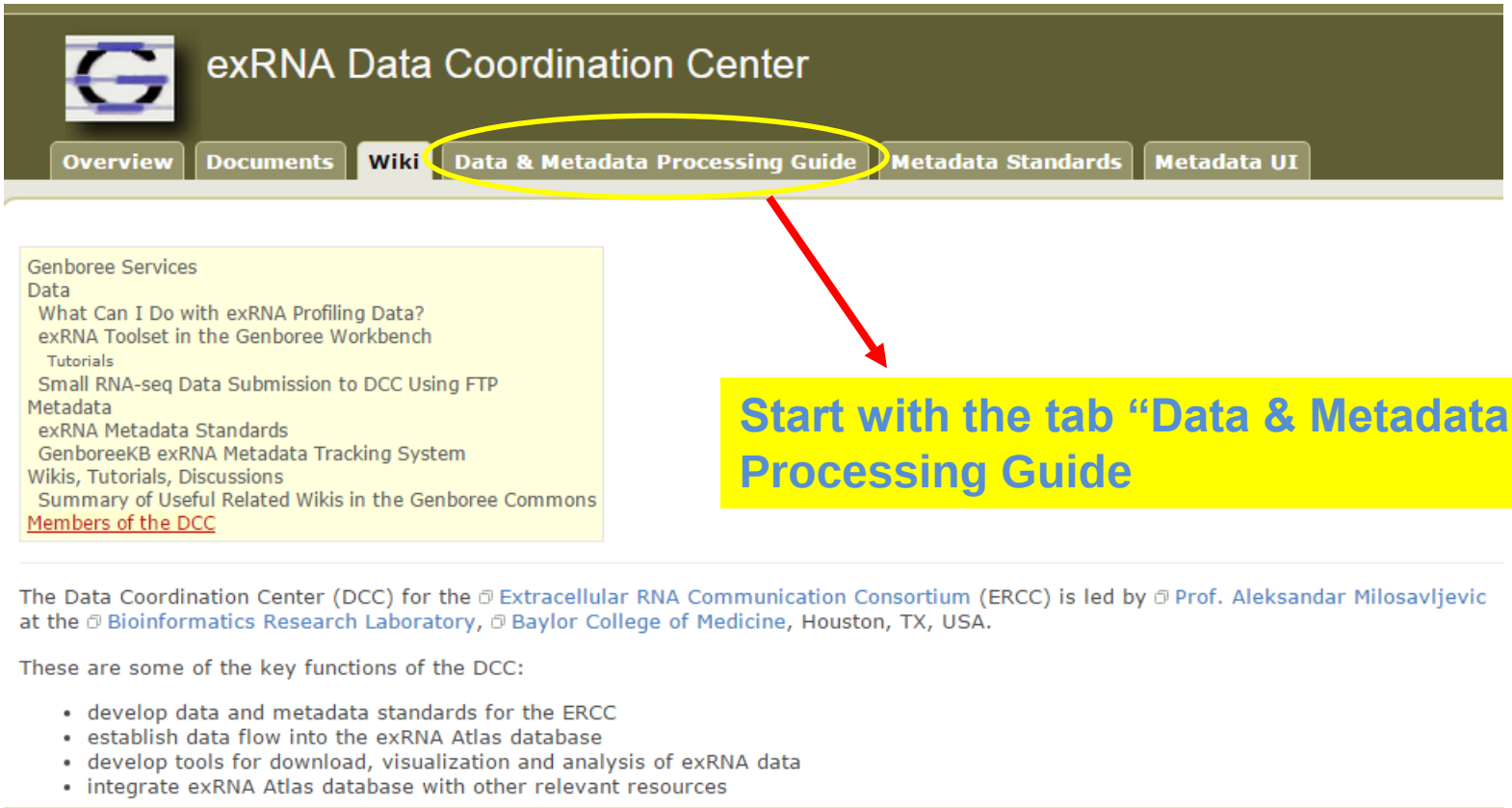
DMRR Documentation – Link from exrna.org





Wikis for Preparing Your Submission

Tutorials for preparing your FTP submission can be found at the following link:
http://genboree.org/theCommons/projects/exrna-mads/wiki



What can I do with exRNA Profiling Data?

If you have exRNA Profiling Data

Small exRNA-seq Data

Identify your category	Identify what you want to do with your data	Choose the type of submission for exceRpt small RNA-seq Pipeline	What is needed for Data Analysis?	What happens to my data after analysis?
Data from ERCC Funding	Data Submission to DCC	FTP Data Submission Pipeline	Data Files – FASTQ, SRA format	Deposited in exRNA Atlas
				Public domain Databases like dbGaP, GEO, SRA
			Metadata Files – Tab separated value format	
			Manifest file	
Other Data from any ERCC group	Data Analysis	Genboree Workbench	Data Files Only – FASTQ, SRA format	Private in your own Genboree Group, Database
Non-ERCC User	Data Analysis	Genboree Workbench	Data Files Only – FASTQ, SRA format	Private in your own Genboree Group, Database

Genboree FTP Data Submission Pipeline

Step 0: Create an FTP account
Step 1: Prepare your data archive
Step 2: Prepare your metadata archive
Step 3: Prepare your manifest file
Step 4: Upload your submission to the FTP server for processing
Step 5: View your results

exRNA Profiling Data - FTP Submission

http://genboree.org/theCommons/projects/exrna-mads/wiki/Creating_an_FTP_Account

Genboree FTP Server	• ftp.genboree.org
If You Have Data	• Email Sai Lakshmi Subramanian at the DCC (sailakss@bcm.edu) for an account on the FTP server.
Login	• Your Genboree user name
Password	• Your Genboree password
Upload directory	• A dedicated, unique and private directory named **"exrna-picode"** for your lab/group, shared only by your lab members. • Details will be provided when your FTP account is created.
How to upload data	• Use Unix/Linux/OSX Command Line or FileZilla FTP Client • **TIP**: Typing the FTP server address on the web browser will **not** allow you to upload files.

Files to Submit

http://genboree.org/theCommons/projects/exrna-mads/wiki/Data%20Submission%20to%20DCC%20using%20FTP#Files-Needed-for-Data-Submission

Data Archive File — Contains your input data files

Metadata Archive File — Contains metadata about your inputs

Manifest file — Contains details of your submission

All three files **must** have the same basic file name.

EXAMPLE:

- samples_data.zip
- samples_metadata.zip
- samples.manifest.json

NOTE: Remember to replace “samples” with a file name relevant to your submission.

Preparing Data Archive

http://genboree.org/theCommons/projects/exrna-mads/wiki/Prepare_your_Data_Archive

Data Archive File

Contains your input data files

- **FILE EXTENSION - _data.zip or _data.tar.gz**
- **FORMAT** – FASTQ/SRA format (can be compressed)
- **REQUIRED FILES - .fastq or .fastq.gz or .fastq.zip or .sra**
- **OPTIONAL FILES** – Spike in sequence file in FASTA format.
  - If you include a spike-in FASTA file in your data archive, make sure to add this in the "Settings" section of your manifest file: *"useLibrary": "uploadNewLibrary"*
- **No** folders are allowed in this archive. Should contain only FASTQ/SRA/FASTA files.

Creating an archive (.zip or .tar.gz)
http://genboree.org/theCommons/projects/exrna-mads/wiki/Creating_an_Archive

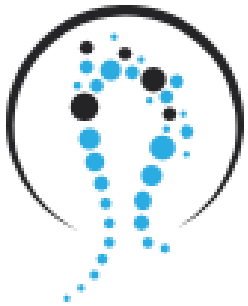
Preparing Metadata Archive

http://genboree.org/theCommons/projects/exrna-mads/wiki/Prepare_your_Metadata_Archive

Metadata Archive File

Contains metadata about your inputs

- **FILE EXTENSION - _metadata.zip or _metadata.tar.gz**
- **FORMAT** – All metadata files should be in tab separated value format
- **REQUIRED FILES - .metadata.tsv** files - Submission, Study, Run, Experiment(s), Biosample(s) and Donor(s) documents.



Genboree KnowledgeBase (GenboreeKB)

http://genboree.org/genboreeKB/projects/genboreekb-introduction/wiki

GenboreeKB → Mongo Database (Backend)
Redmine Plugin (UI)

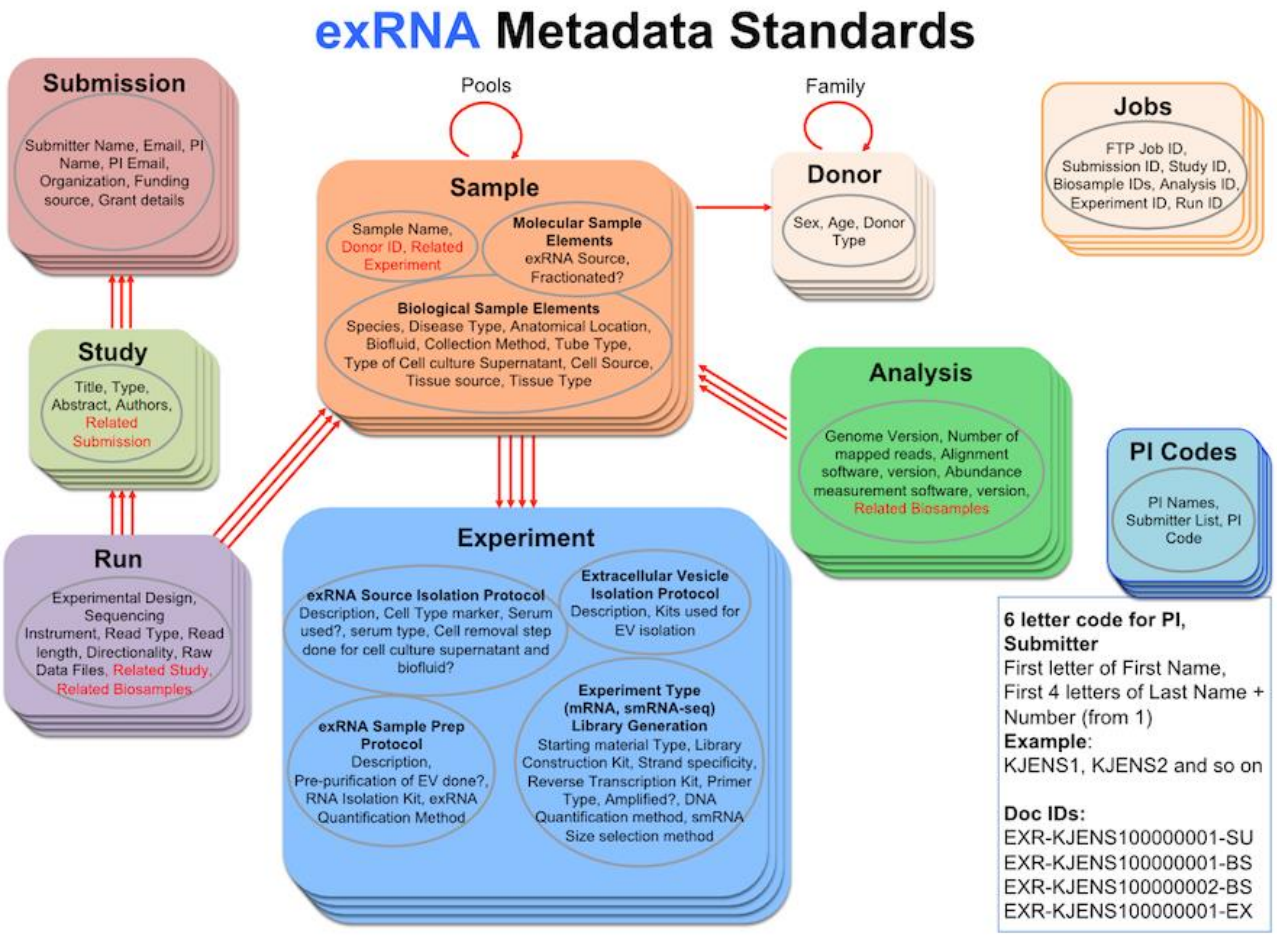
- Project-specific GenboreeKB
- Multiple Collections of Documents
- Each metadata collection has its own document data model
- Singly-Rooted Nested Collection of Properties
- **Data Model** - Defines "properties" and "property definitions"
- **Property Definitions** - Fields describing each property like "domain", "required", "identifier", "category", "description", etc.
- **Key Features**
  - Browse, Manage documents and models, Queries, Views, Bulk upload and download of documents, JSON/Tabbed formats
- Dynamic retrieval and validation of ontology terms from **Bioportal**

exRNA Metadata Standards

http://genboree.org/genboreeKB/genboree_kbs?project_id=exrna-metadata-standards

Submitting metadata is extremely IMPORTANT

- Your samples can be analyzed and compared only if metadata is available
- Reproducibility of experiments



Metadata Submission Guidelines

Submission	• Submit 1 **Submission** document – **UNIQUE** for each Submitter-PI-Grant combination – **REUSE** for submissions from same user/PI/grant - Do NOT provide duplicates.
Study	• Submit 1 **Study** document with a **unique** title and overall description for a single research initiative – This describes "why" you sequenced your samples (similar to NCBI BioProject) – **REUSE** for all samples sequenced under a single study.
Run	• Submit 1 **Run** document providing a list of data files that are part of the submission – **UNIQUE** for each FTP submission.
Experiment	• Submit 1 or more **Experiment** document(s). Ensure Biosample is linked with the correct Experiment document ID. **REUSE** for experiments that followed exact same protocols.
Donor	• Submit one or more **Donor** document(s) for all Biosamples that are part of this submission – **UNIQUE** for each Donor – **REUSE** for samples from same donor - Do NOT provide duplicates.
Biosample	• Submit 1 **Biosample** document for each data file submitted. Ensure the correct **Donor ID** is provided in each Biosample document.
Analysis	• Optionally submit 1 **Analysis** document – If not provided with submission, it will be generated and relevant fields along with Biosamples will be populated by the FTP submission pipeline – **UNIQUE** for each FTP submission.

exRNA Metadata Document Accession

EXR-({6 character PI CODE}[A-Za-z0-9]{6,})-[ST|SU|BS|AN|EX|RU|FL|DO]

Prefix **EXR**, followed by "-",
6 character PI Code, alphanumeric string with 6 or more characters with letters and numbers, followed by "-",
a suffix with two-letter code for each collection type.

Type	Example Accession
Biosample	EXR-KJENS12P3L78-BS
Donor	EXR-KJENS12P3L78-DO
Experiment	EXR-KJENS12P3L78-EX
Analysis	EXR-KJENS12P3L78-AN
Submission	EXR-KJENS12P3W78-SU
Run	EXR-KJENS12P3W78-RU
Study	EXR-KJENS12P3L78-ST

For PI Code: Refer to the **Master List of ERCC PIs:**
http://genboree.org/genboreeKB/genboree_kbs?project_id=exrna-metadata-standards&coll=PI%20Codes&doc=EXR-MASTERLIST-PI

exRNA Metadata Models, Templates, Examples

Available for download at:
http://genboree.org/theCommons/projects/exrna-mads/wiki/Prepare_your_Metadata_Archive#Download-Metadata-Models-Document-Templates-and-Example-Metadata-Documents

Schema	Description	Doc Template For Editing in Excel	User Submitted Metadata Example	Template in GenboreeKB UI
Biosamples TABBED Model	Detailed information about the sequenced sample, biofluid source, etc. Samples can be used in any number of experiments.	Biosample Template Multi-tabbed Format	EXR-LLAUR1NEBLIB41-BS	Biosample KB Template
Donors TABBED Model	Information about each individual donor who contributed biosamples.	Donor Template	EXR-LLAUR1NEBLIB41-DO	Donor KB Template
Studies TABBED Model	A study groups together experiments or analyses for public data release purposes.	Study Template	EXR-LLAUR1M4TD4M0N-ST	Study KB Template
Experiments TABBED Model	An experiment contains instrument and library preparation information and groups together one or more runs.	Experiment Template	EXR-LLAUR1M4TD4M0N-EX	small RNA-seq KB Template mRNA-seq KB Template
Analyses TABBED Model	An analysis contains secondary analysis results.	Analysis Template	EXR-LLAUR1M4TD4M0N-AN	Analysis KB Template
Submissions TABBED Model	Information about PI / submitter associated with submission.	Submission Template	EXR-LLAUR1M4TD4M0N-SU	Submission KB Template
Runs TABBED Model	A run contains sequencing reads submitted in data files.	Run Template	EXR-LLAUR1M4TD4M0N-RU	Run KB Template

Data Models

Example Templates for editing in Excel

User submitted metadata documents

GenboreeKB Document Templates

http://genboree.org/genboreeKB/

Username: Your Genboree (Commons) login name

Password: Your Genboree (Commons) Password

Email Sai (sailakss@bcm.edu) after logging in once, so I can add you to the exRNA Metadata Standards Project

Metadata File – Tabbed Format

OPTION 1

Download templates from tutorial and edit

Download the templates from the column **Doc Template For Editing in Excel** in the table and edit.

Fill in values in the second column in these template documents.

The **domain**, **required** and **description** columns are available in these templates for your reference.

It is sufficient to fill values for properties that have required column marked as TRUE or 1. Providing any additional metadata is optional.

TIP: File should have file extension **.metadata.tsv**.

OPTION 2

Generate a template using the GenboreeKB UI

The link in the column **Template in GenboreeKB UI** will open the template in the browser.

Simply add values to the required properties and other optional properties for which you have metadata.

All ontology terms are dynamically retrieved from Bioportal.

Save and download this template document, and then use it for creating multiple metadata files.

TIP: Provide the correct **PI code** in the document accession property.

Metadata File – Tabbed Format – Key Bits

The leading "-" or "--" and similar dashes in the first property column indicate the level of nesting of properties in the document.

The leading “*” or “*-” or any combination of star-dash indicates that the property is part of an item list. If you would like to add more than one value for an item list property, then copy-paste the row and add additional values.

Example: For the property **Authors** in the Study collection, if you would like to add 10 authors, then add 10 rows of ***- Author Name*** below the property ***Authors**.

TIP: If you do not have values for some properties in the document, DO NOT enter N/A or n.a. or na or NA or any of its variants.

TIP: You can delete any unused rows. Remember that you should maintain the level of nesting (i.e. the number of leading dashes and stars).

Metadata File – Nested Document Model

Tabbed Document Template in Excel

A	B	
#property	value	domain
Study		autoID(EXR, uniqAlphaNum, ST)
- Status		enum(Add, Modify, Hold, Cancel, Suppress, F
- Schema Version		float
- Title		string
- Type		enum(Whole Genome Sequencing, Metagen
-- Other Type		string
- Abstract		string
* Authors		[valueless]
*- Author Name		string
- Overall Design		string
- Notes		string
* References		[valueless]
*- PubMed ID		posInt
* Related Documents		[valueless]
*- Related Document		regexp(EXR-[A-Z0-9]{8}-[BS|RU|EX|AN|SU|
*-- DocType		enum(Biosample, Run, Experiment, Analysis,
*-- DocURL		url
* Aliases		[valueless]
*- Accession		string
*-- dbName		enum(SRA, GEO, DDBJ, ENCODE, dbGaP)
*-- URL		url

Nested Model in the UI

Model Tree

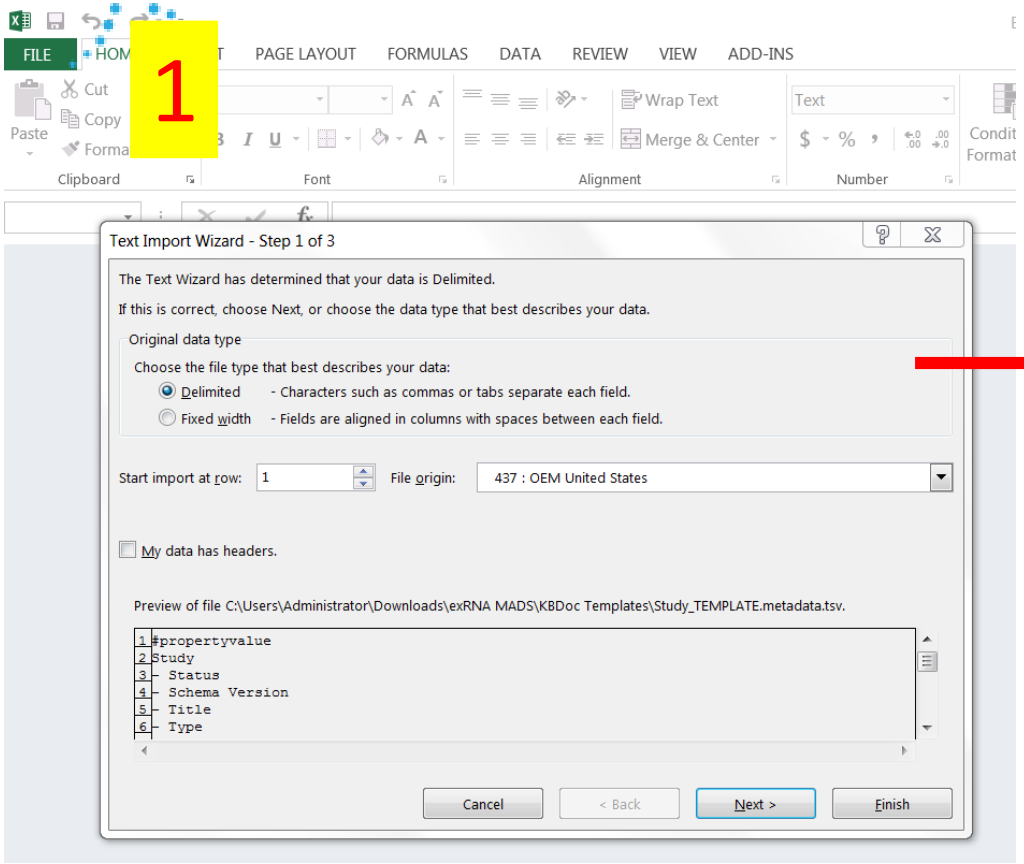
Name	Domain	Required
Study	autoID(EXR, uniqAlphaNum, ST)	true
Status	enum(Add, Modify, Hold, Cancel, Suppress, ...	true
Schema Version	float	
Title	string	true
Type	enum(Whole Genome Sequencing, Metagen...	true
Other Type	string	
Abstract	string	true
Authors	[valueless]	
Author Name	string	true
Overall Design	string	
Notes	string	
References	[valueless]	
PubMed ID	posInt	true
Related Documents	[valueless]	
Related Document	regexp(EXR-[A-Z0-9]{8}-[BS|RU|EX|AN|SU...	true
DocType	enum(Biosample, Run, Experiment, Analysis,...	
DocURL	url	
Aliases	[valueless]	
Accession	string	true
dbName	enum(SRA, GEO, DDBJ, ENCODE, dbGaP)	
URL	url	

Metadata Docs in Microsoft Excel –

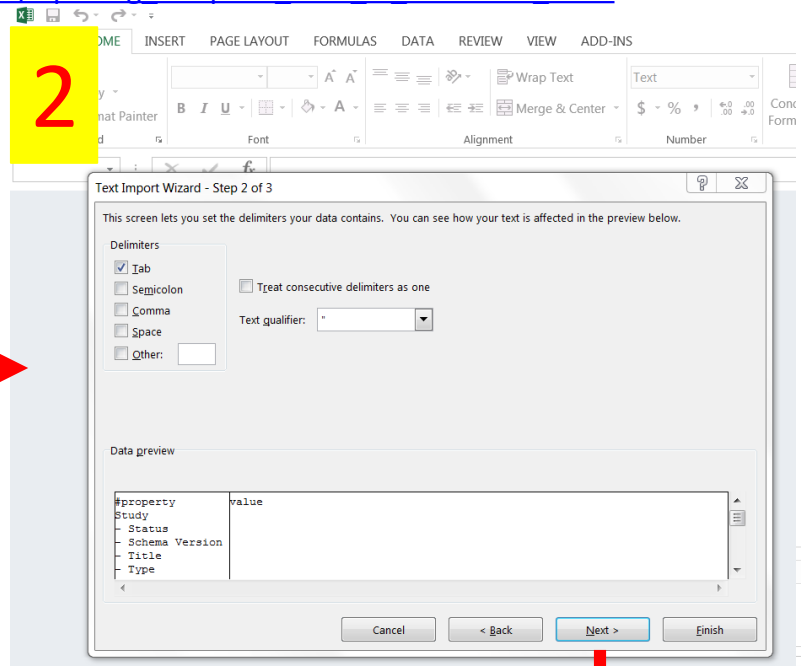
Preserve the Tree Structure

http://genboree.org/theCommons/projects/exrna-mads/wiki/Opening_template_docs_in_Microsoft_Excel

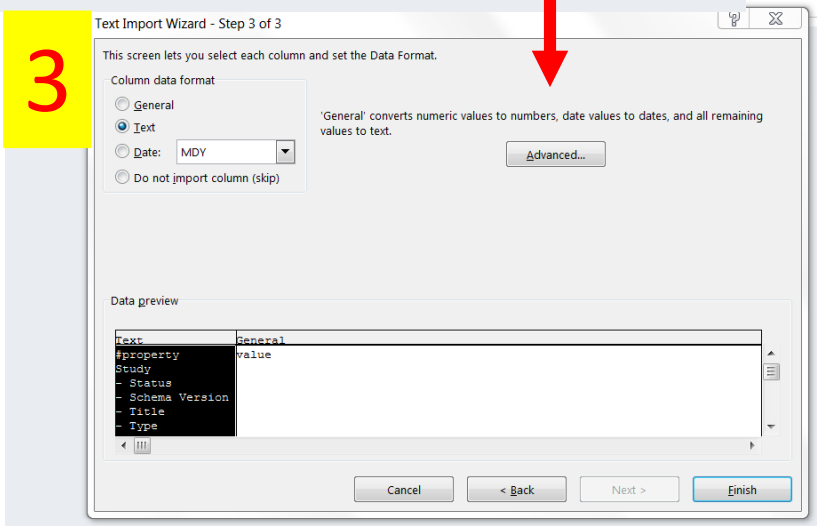
1



2



3



Saving Metadata Document as a tsv file:
http://genboree.org/theCommons/projects/exrna-mads/wiki/Saving_metadata_documents

Metadata File – Special Domains

http://genboree.org/theCommons/projects/exrna-mads/wiki/Prepare_your_Metadata_Archive#Special-Domains

enum Domain

- Certain fields only accept certain words as valid text.
- **EXAMPLE**: In a **biosample** metadata file, the "Status" field has a domain of ***enum(Add, Modify, Hold, …)***. This field will only accept certain words as valid (Add, Modify, Hold, etc.).
- In the UI, editing one of these fields results in a drop-down menu consisting of all valid choices.

Measurement Domain

- Certain fields only accept numeric values followed by the ***unit*** specified in the domain or other acceptable conversions of the defined unit.
- **EXAMPLE**: In a **biosample** metadata file, the **Age** property has the domain ***measurement(years)***. You can enter 10 years or 10 days or 10 months – all units related to years are acceptable but the unit will be converted before it is saved.
- http://genboree.org/theCommons/projects/exrna-mads/wiki/List_of_Units_supported_by_GenboreeKB

Valueless Domain

- Do not enter any value for these properties.
- **EXAMPLE**: In a **biosample** metadata file, you cannot attach a value to the "Donor" category.

Metadata File – Ontology Validation

bioportalTerm(s) Domain	Any bioportalTerm(s) domain requires a value that matches the correct ontology term.
EXAMPLE	For **Disease Type** in the **Biosamples** Collection, schema specifies this domain: bioportalTerms((DOID,http://purl.obolibrary.org/obo/DOID_4),(NCIT,http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C7057),(NCIT,http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C49651))
Validation	Any value entered for this property will be validated against the **NCIT** ontology in **Bioportal**.
TIP	Get correct term from correct ontology in Bioportal. (or) Use GenboreeKB UI to get correct ontology terms for all required fields.

Preparing Manifest File

Manifest file

Contains details of your submission

- **FILE EXTENSION - .manifest.json**
- **FORMAT** - JSON format
- **REQUIRED** - Genboree login name, group name, database name, list all files that are submitted, MD5 checksum, tool specific settings
- http://genboree.org/theCommons/projects/exrna-mads/wiki/Prepare_your_Manifest_File

Manifest File

REQUIRED File – Job will not be accepted if this file is not provided.

Templates are available in the Tutorial.

Data submitter should provide a valid Genboree user login.

Genboree Group name and Genboree database name are required for uploading pipeline outputs – **Group and Database must exist in Genboree**

Compute MD5 checksum of the data archive that you are submitting and include that value in the manifest file. See Tutorial for more info.

TIP: If you modify your archive, remember to recompute MD5 checksum and enter the new value in manifest file.

Data Upload to FTP Server

http://genboree.org/theCommons/projects/exrna-mads/wiki/Upload_Submission_to_the_DCC_using_FTP_Server

To upload data to the FTP server, use:

FTP client like FileZilla

Command line FTP data transfer in Unix/Linux/Mac

TIP: Typing the FTP server address on the web browser will not allow you to upload files.

Jobs will be queued in the Genboree Cluster and processed in the order of submission.

Larger submissions (100+) will take a few days to complete.

On completion (success or failure), you will receive an email – **PLEASE READ THE EMAIL CAREFULLY.**

Failed samples should be resubmitted separately, after being fixed.

Order of Processing: Manifest -> Metadata -> Data

Error in any stage => Fix only that file (details will be available in the email) and re-upload to your ***FTP inbox***

Viewing your Results

http://genboree.org/theCommons/projects/exrna-mads/wiki/Viewing_Your_Results

FTP Server: *exrna-picode/finished/* directory

Genboree Workbench: Genboree Group/Database specified in the manifest

Metadata Files: GenboreeKB UI

Accession numbers assigned for your metadata will be provided in the email.

Useful Links

- exRNA Portal Software Resources
  http://exrna.org/resources/software
- Genboree Workbench
  http://genboree.org/java-bin/workbench.jsp
- Data Coordination Center Wiki
  http://genboree.org/theCommons/projects/exrna-mads/wiki
- exRNA Data Analysis Tools Wiki
  http://genboree.org/theCommons/projects/exrna-tools-may2014/wiki