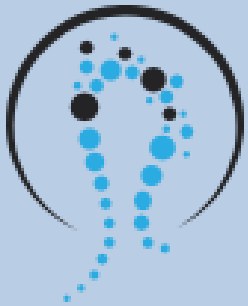


exRNA Communication Consortium
Data Management And Resource Repository

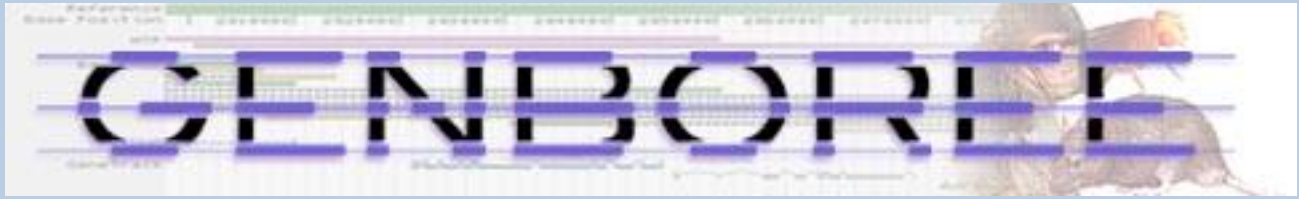
exRNA Data & Metadata Submission Infrastructure at the DMRR

Sai Lakshmi Subramanian, William Thistlethwaite, Aaron Baker
Aleks Milosavljevic
Data Coordination Center – Baylor College of Medicine
16th October 2014

Baylor College of Medicine

Outline

- Genboree Services
- exRNA Data Submission using FTP
- exRNA Metadata Data Model
- GenboreeKB exRNA Metadata Tracking System
- exRNA Data Analysis Tools



DCC - Genboree Services for the ERCC

exRNA Data Analysis Tools - Genboree Workbench
http://genboree.org/

exRNA Document Sharing & Discussion Forums – Genboree Commons
http://genboree.org/theCommons

exRNA Metadata Tracking System – GenboreeKB
http://genboree.org/genboreeKB

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

exRNA Data - FTP Submission

FTP Server	• ftp.genboree.org
If you have data	• Email Sai (sailakss@bcm.edu) for an account in the FTP server.
Login	• Your Genboree user name
Password	• Your Genboree password
Upload directory	• A dedicated unique and private directory for your lab/group, shared only by your lab members. • Will be provided when your FTP account is created

Data upload to FTP server

To upload data to the FTP server, use:

FTP client like Filezilla

Command line FTP data transfer in Unix/Linux/Mac/Windows

Instructions to upload files to the FTP server:

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

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

FTP Submission – Files to submit

Manifest file

REQUIRED - .manifest.json

- JSON Format
- See template and edit fields

A single data archive file

REQUIRED - .tar.gz or .zip

- Data file – in FASTQ format (can be compressed) - **REQUIRED - .fastq or .fastq.gz or .fastq.zip**
- Metadata file – in TAB format – **REQUIRED - .metadata.tsv**
- Config file – in JSON format – **OPTIONAL** - .config.json

Data Submission - Manifest file – JSON format


```
{
  "studyName": "STUDYNAME",
  “userLogin”: “GENBOREE LOGIN NAME”,
  “group”: “GENBOREE GROUP”,
  “db”: “GENBOREE DATABASE”,
  "md5CheckSum": "MD5 CHECKSUM OF DATA ARCHIVE",
  "runMetadataFileName": "RUN_METADATAFILE.metadata.tsv",
  "submissionMetadataFileName": "SUBMISSION_METADATAFILE.metadata.tsv",
  "analysisMetadataFileName": "ANALYSIS_METADATAFILE.metadata.tsv",
  "studyMetadataFileName": "STUDY_METADATAFILE.metadata.tsv",
  "experimentMetadataFileName": "EXPERIMENT_METADATAFILE.metadata.tsv",
  "manifest":
  [
    {
      "sampleName": "SAMPLE1",
      "dataFileName": "DATAFILE1.fastq.gz",
      "metadataFileName": "METADATAFILE1.metadata.tsv",
      "configFileName": "DATAFILE1.config.json"
    },
    {
      "sampleName": "SAMPLE2",
      "dataFileName": "DATAFILE2.fastq.gz",
      "metadataFileName": "METADATAFILE1.metadata.tsv",
      "configFileName": "DATAFILE2.config.json"
    }
]
}
```

Manifest file

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

Templates are available in the Tutorial

File should have extension **.manifest.json**

Data submitter should provide a valid Genboree user login

Genboree Group name and Genboree database name are required for uploading pipeline outputs

Compute MD5 checksum of the data archive that you are submitting, 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

Tool Config file (for smallRNA Pipeline) - JSON format – OPTIONAL FILE

Templates are provided

Provide a sensible "analysisName" for your sample

TIP for analysisName:

Your Sample name with data/time string as shown

TIP: Do not provide this file if you want to use the default settings.

```
{
    "settings":{
        "tRNAmapping": "on",
        "clippedInput": "on",
        "adapterSequence": "",
        "piRNAmapping": "on",
        "mapPlantsViruses": "on",
        "mapRfam": "on",
        "snoRNAmapping": "on",
        "analysisName": "SAMPLENAME-
            2014-10-16-10:05:07"
        }
}
```

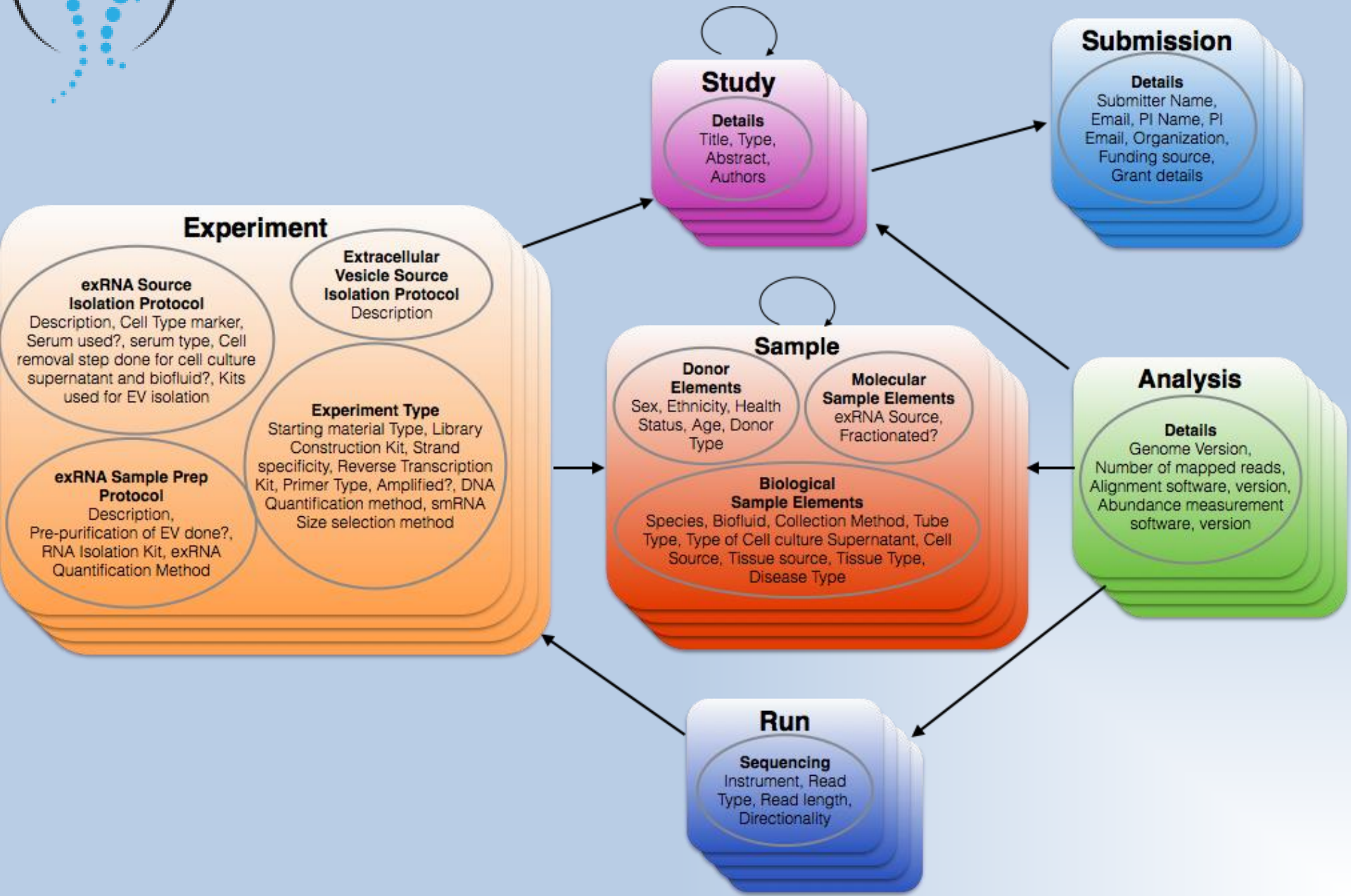
exRNA Metadata Standards

Submitting metadata is extremely IMPORTANT

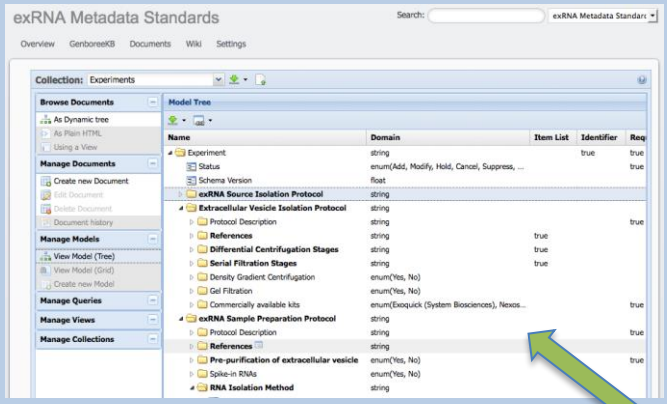
A couple of reasons:

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

exRNA Metadata Standards



GenboreeKB exRNA Metadata Tracking System



Editing, Browsing

GenboreeKB REST-APIs

MongoDB – Collections of Documents

Query Tools, Other applications

JSON (native)

XML (archiving at NCBI)

RDF (Linked Data)

Metadata Documents – JSON format {"attribute": "value"}

Upload and Validation

exRNA Metadata Data Models

Metadata Data models for GenboreeKB

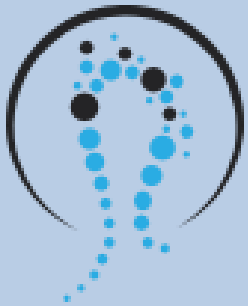
Available for download at:
http://genboree.org/theCommons/projects/exrna-mads/wiki/exRNA%20Metadata%20Standards
Look for **Schema** column in the table

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



Genboree KnowledgeBase (GenboreeKB)

GenboreeKB → Mongo Database (Backend), Redmine plugin (UI)

- ✓ 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 Document Accession

EXR-([A-Z0-9]{8})-[ST|SU|BS|AN|EX|RU|FL]

Prefix **EXR**,
followed by "-",
alphanumeric string with exactly 8 characters with uppercase alphabets and numbers,
followed by "-",
a suffix with two-letter code for each collection type.

Examples:

Biosample: EXR-2P3LNW78-BS
Experiment: EXR-2P3LNW78-EX
Analysis: EXR-2P3LNW78-AN
Submission: EXR-2P3LNW78-SU
Run: EXR-2P3LNW78-RU
Study: EXR-2P3LNW78-ST
File : EXR-2P3LNW78-FL

Metadata file – Tabbed format

Download templates from tutorial and edit

- http://genboree.org/theCommons/projects/exrna-mads/wiki/exRNA%20Metadata%20Standards
- Download the templates from the last column **Doc Template** in the table given in tutorial 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**

Generate a template using the GenboreeKB UI

- Just add those required properties and other optional properties for which you have metadata
- All ontology terms are dynamically retrieved from Bioportal
- Download this template document, and then use it for creating multiple metadata files.

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 the rows. Remember that you should maintain the level of nesting (i.e. the number of leading dashes and stars).

Metadata file – Nested document model

A	B	
#property	value	domain
Study		autoID(EXR, uniqAlphaNum, ST)
- Status		enum(Add, Modify, Hold, Cancel, Suppress, Release, Protect, Validate)
- 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

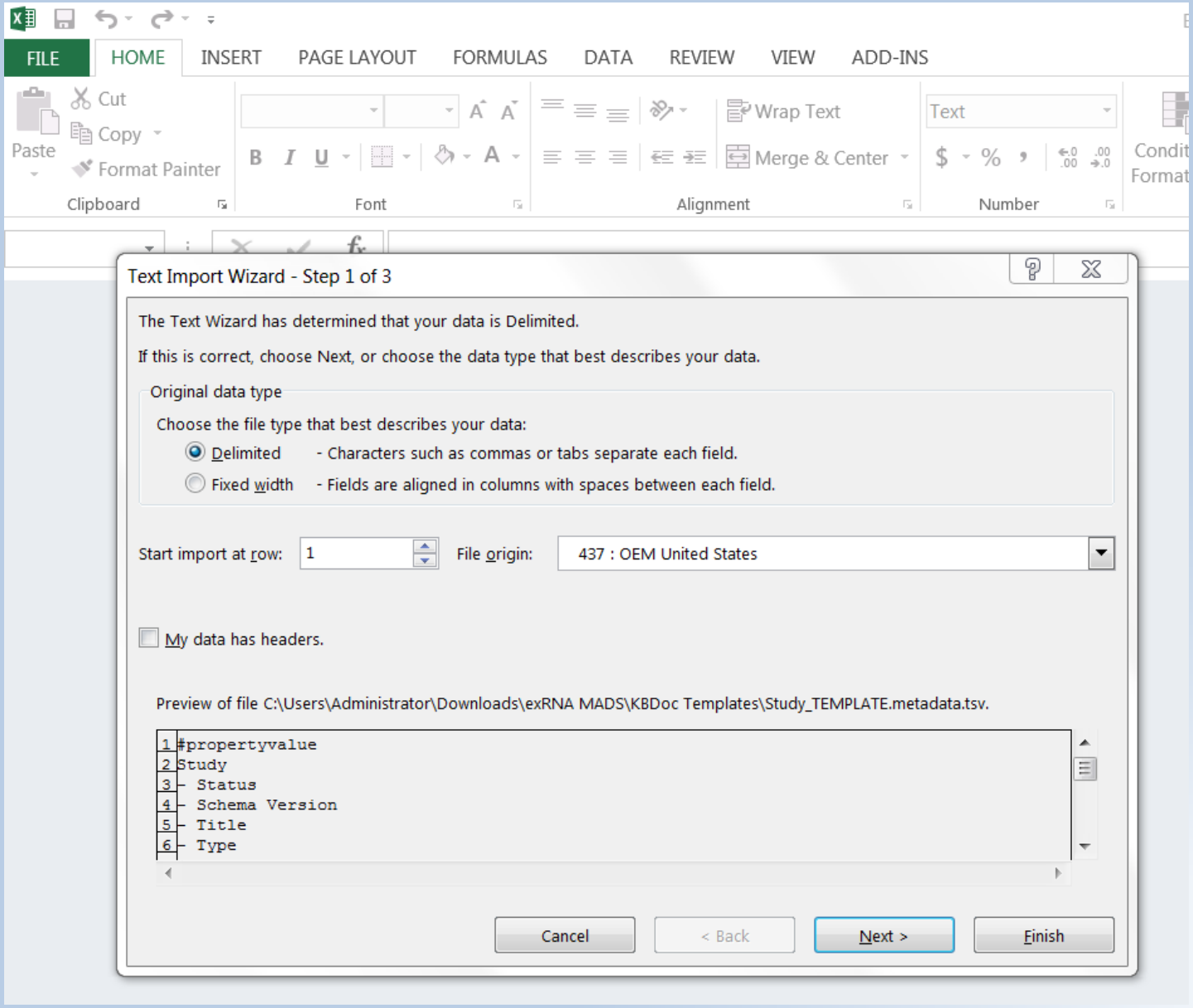
Tabbed Document Template

Nested model in the UI

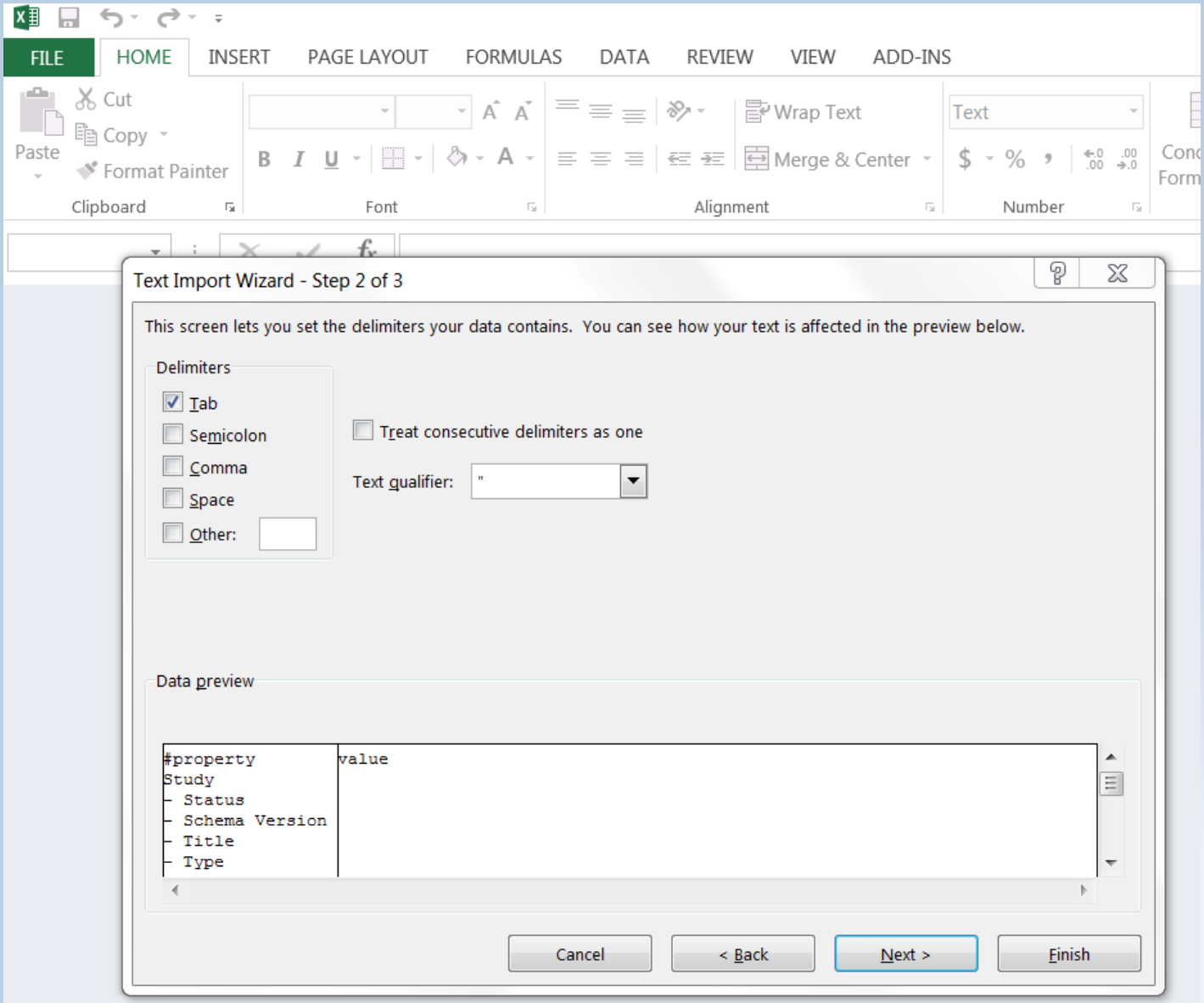
Model Tree

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	

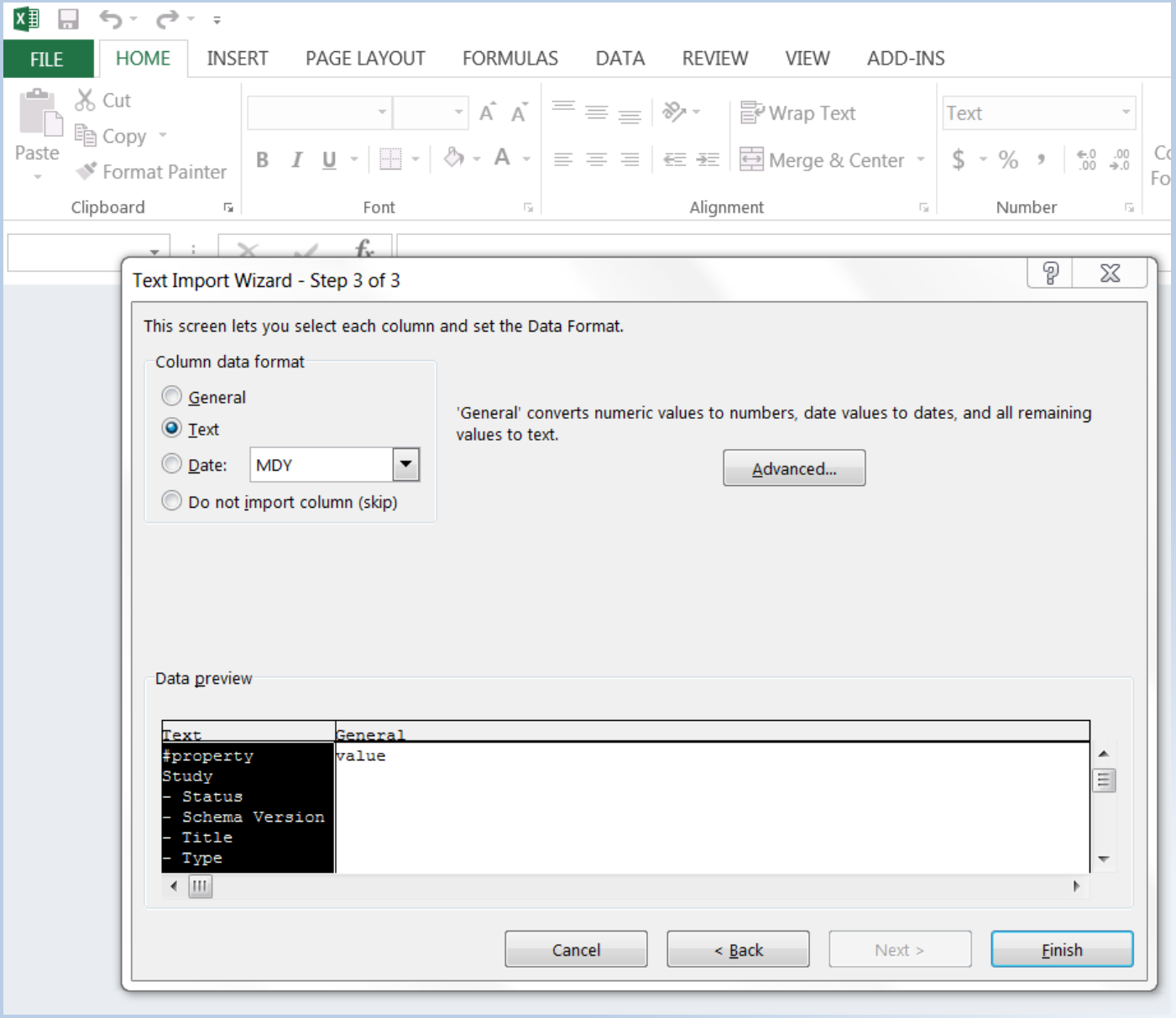
Opening nested doc template in Microsoft Excel – preserve the tree structure



Opening nested doc template in Microsoft Excel – preserve the tree structure



Opening nested doc template in Microsoft Excel – preserve the tree structure



Metadata file – 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 with all valid choices.

Measurement Domain

- Numeric values followed by the ***unit*** specified in the domain or other acceptable conversions of the defined unit.
- **EXAMPLE**: The **Age** property in **Biosample** document has the domain ***measurement(years)***.
- You can enter 10 years or 10 days or 10 months – All units related to years are acceptable.

Valueless Domain

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

Metadata file – Ontology Validation

bioportalTerm domain	Require values that match the correct ontology term
EXAMPLE	For **Disease Type** in the **Biosamples** Collection, schema specifies this domain: bioportalTerm(http://data.bioontology.org/search?ontology=NCIT&subtree_root=http%3A%2F%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C7057)
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

Genboree Workbench

Introduction to Genboree Workbench - Watch Video Tutorial

- https://docs.google.com/file/d/0Bz3_YiJBA_j3Tk1uOFlIazdMbkk/

Create a new account in Genboree

- http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=493

Using the Genboree Workbench

- Find Help within the Genboree Workbench http://genboree.org/theCommons/ezfaq/show/public-commons?faq_id=497
  - Help » Getting Started
  - Help » Tool Map
  - Help » Tool Help Resources

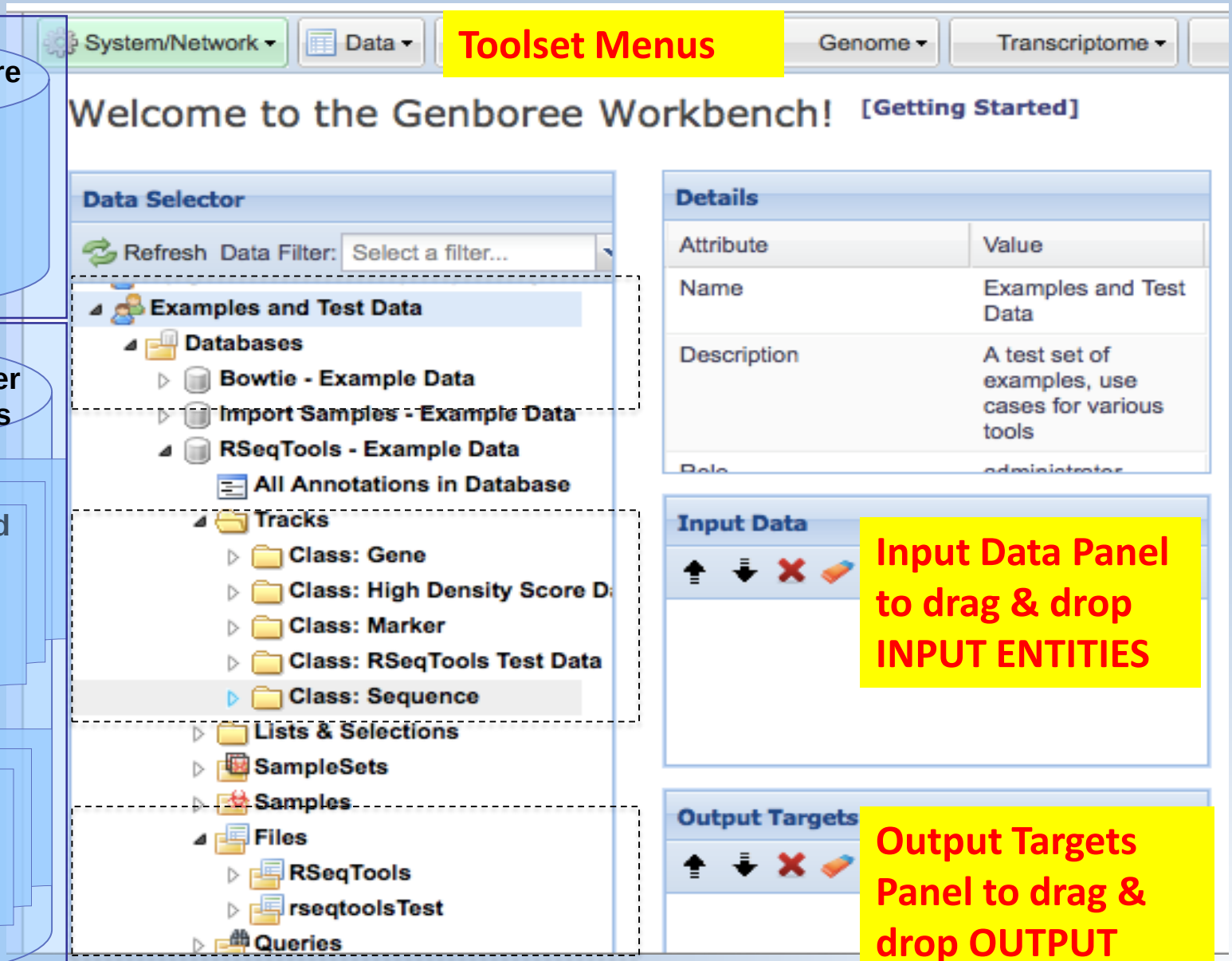
Genboree Workbench

Genboree Core Services
- Authorization & access
- User data repository management

Genboree User Data Services

Genome-Based Tables
- Genomic annotations
- High-density score data

Generic Entity Tables
- Loosely typed metadata
- AVP based



exRNA Data Analysis Tools in Genboree Workbench

Long RNA-seq using RSeqTools

Transcriptome » Analyze RNA-Seq Data » Analyze RNA-Seq data by RSEQtools

Small RNA-seq Pipeline

Transcriptome » Analyze Small RNA-Seq Data » smallRNA-seq Pipeline

Tutorials

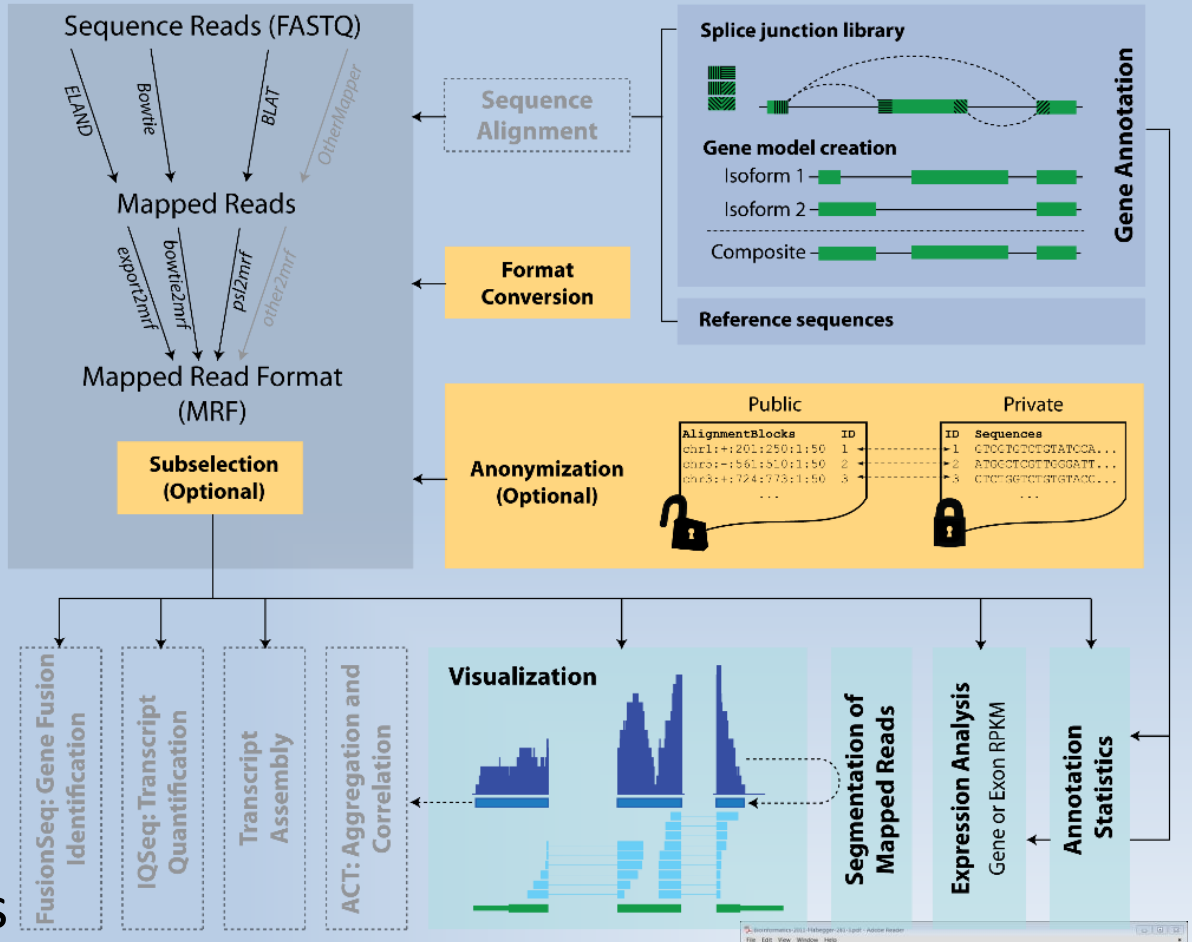
http://genboree.org/theCommons/projects/exrna-tools-may2014/wiki

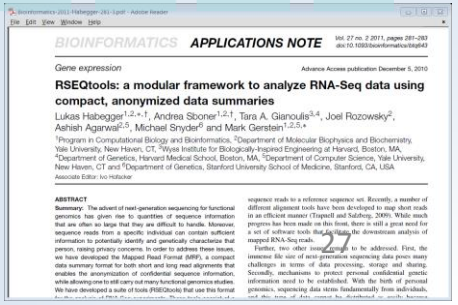
Long RNA-seq Pipeline - RSEQtools

A modular and flexible framework to perform common long RNA-Seq analysis tasks

Current implemented pipeline

Generates genome-wide signal maps (hg18/19) as well as gene level expression quantifications (RPKM) for UCSC known genes.

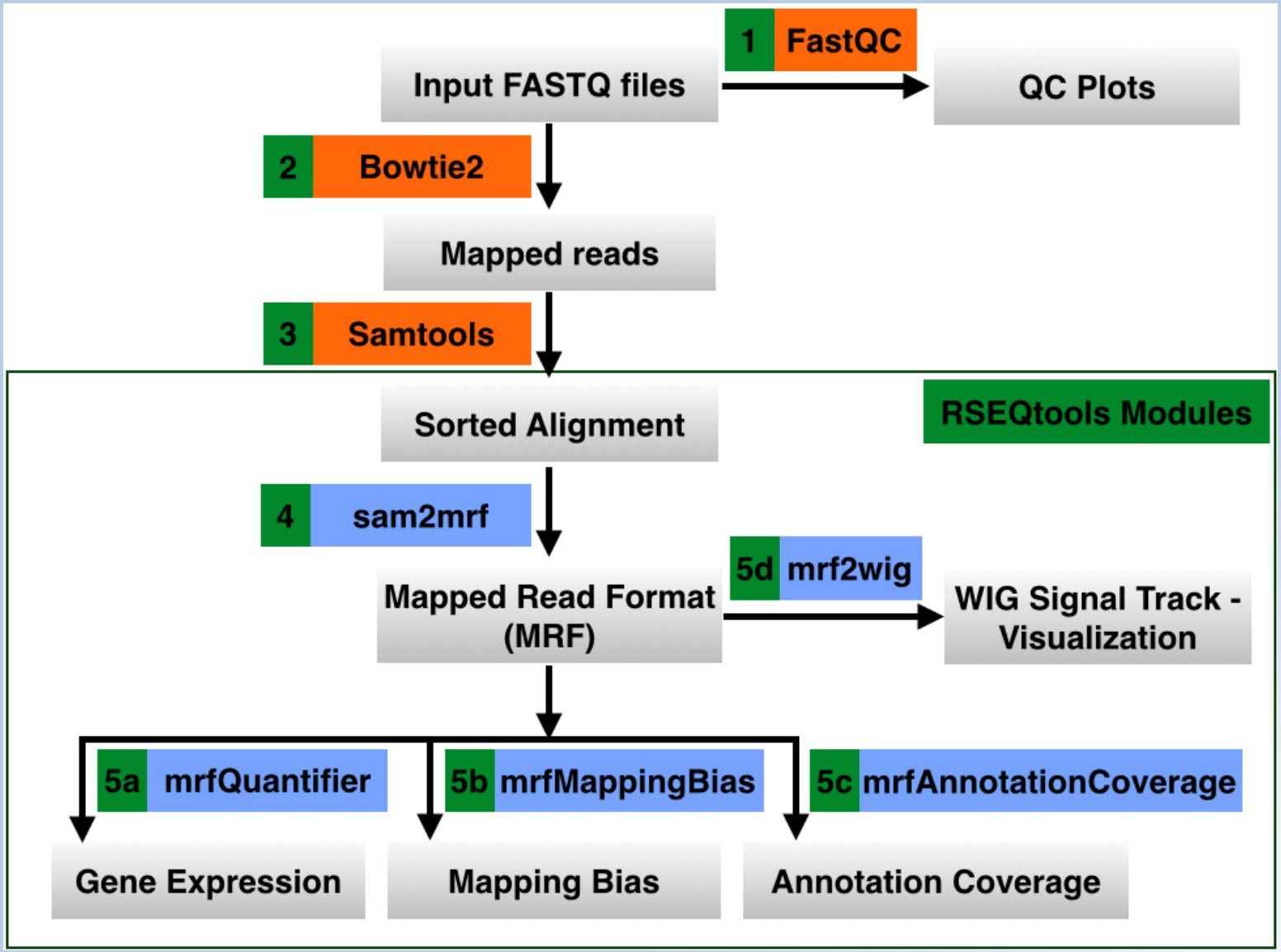



BIOINFORMATICS APPLICATIONS NOTE

RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries

Developed by Gerstein Lab – DIAC, exRNA consortium

Long RNA-Seq Analysis pipeline in the Genboree Workbench

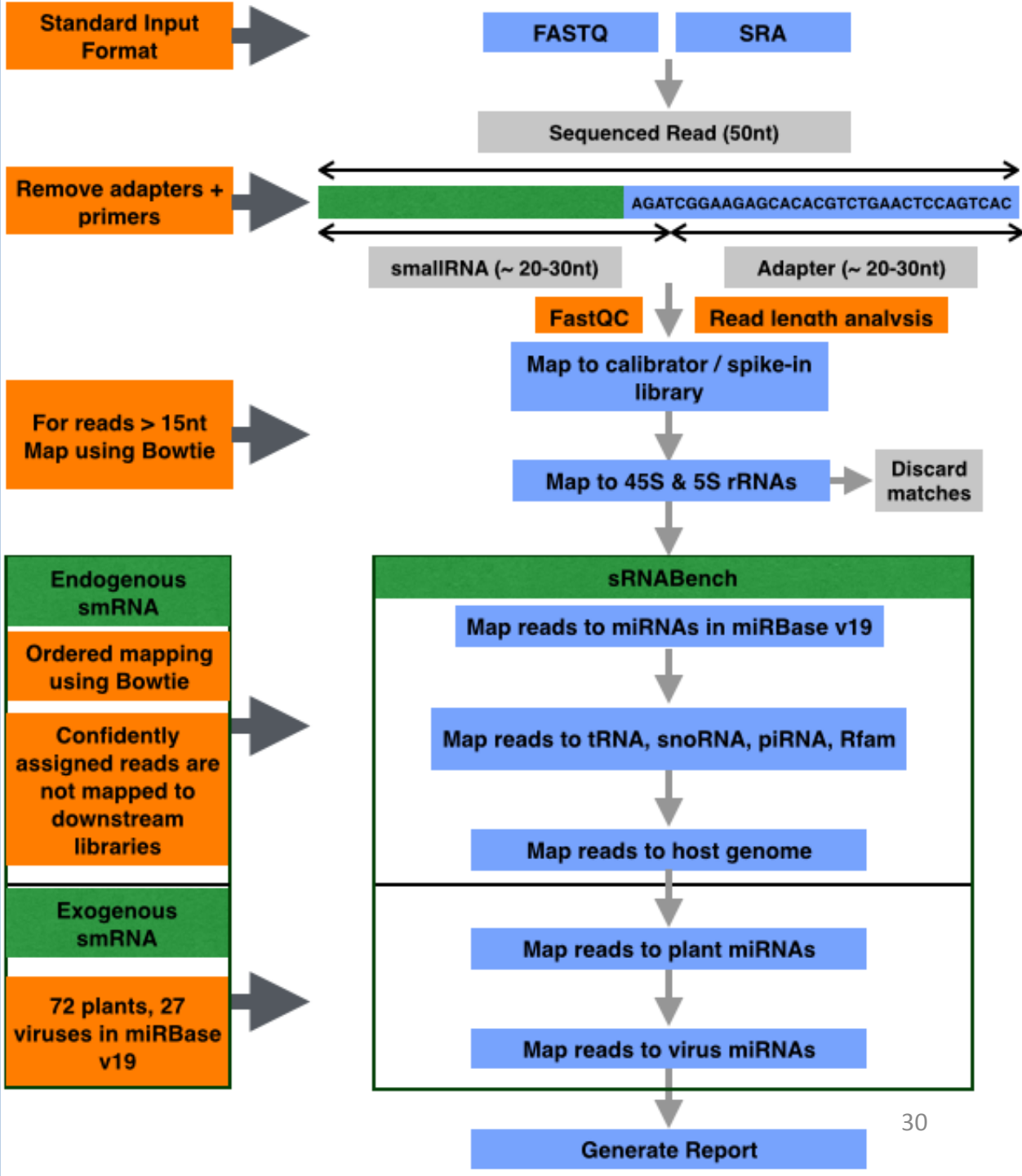


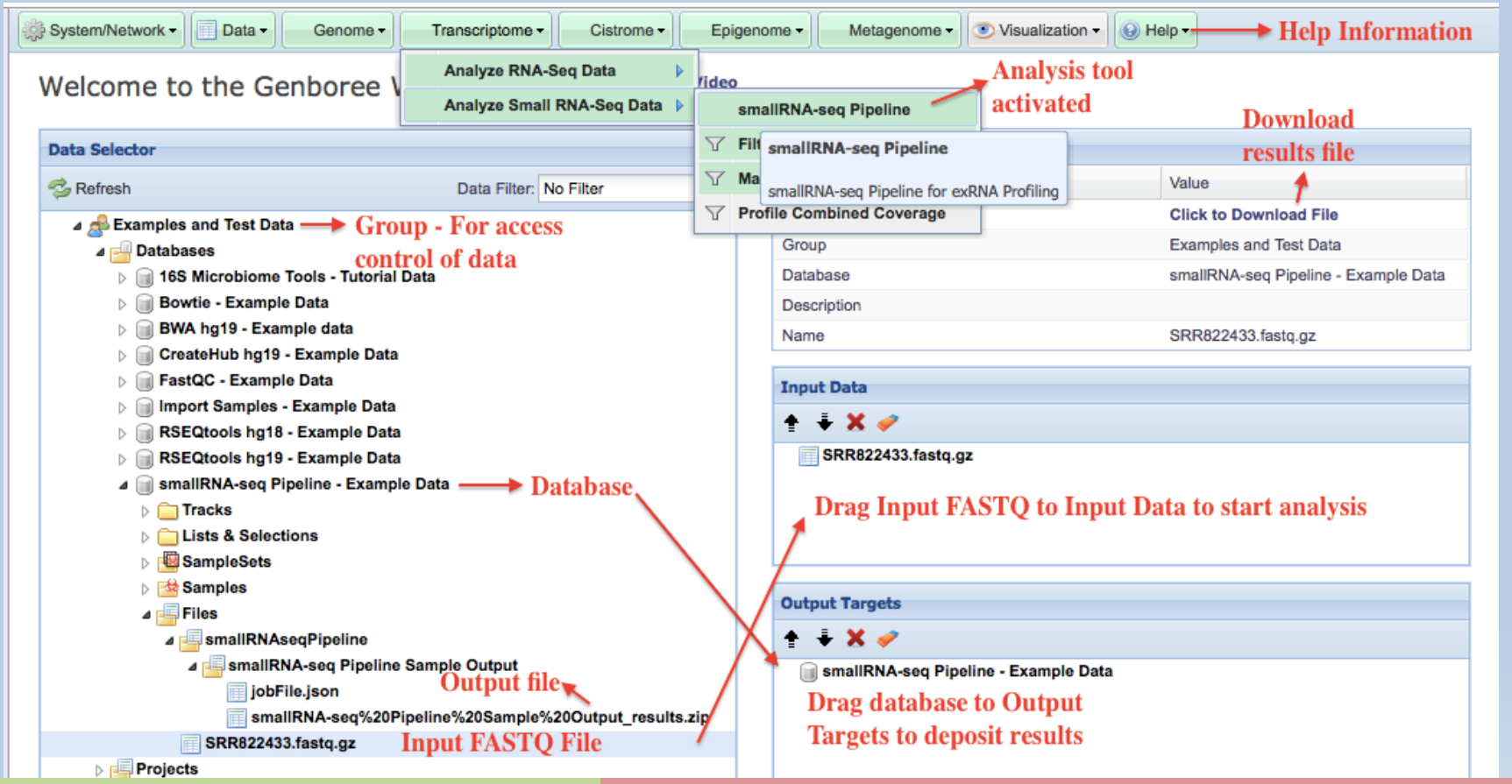
Small exRNA Data analysis

smallRNA-seq Pipeline

- pre-processing: remove adapter & primer sequences
- contaminants: no poly-A purification leaves rRNAs, etc
- mapping: small reads tend to multi-map to large genomes
- diversity: many species of small RNA (not just miRNAs!)
- quantification: different normalization

smallRNA-seq pipeline implemented in the Genboree Workbench





To submit a smallRNA pipeline job in Genboree workbench

INPUT: Single-end FASTQ sequence file (can be compressed)

Supported Genome build: hg19

OUTPUT: All result files compressed in *AnalysisName*_results.zip

OUTPUT FILES:

- Read length distribution
- Endogenous Mapping results – rRNA, miRNA, tRNA, snoRNA, piRNA, Rfam
- Exogenous Mapping results – Plant and virus miRNAs in miRBase
- Files containing sequence reads following Adapter removal, rRNA removal, Endogenous mapping & Exogenous mapping

Questions?

Appendix