

GeOMe Help Document

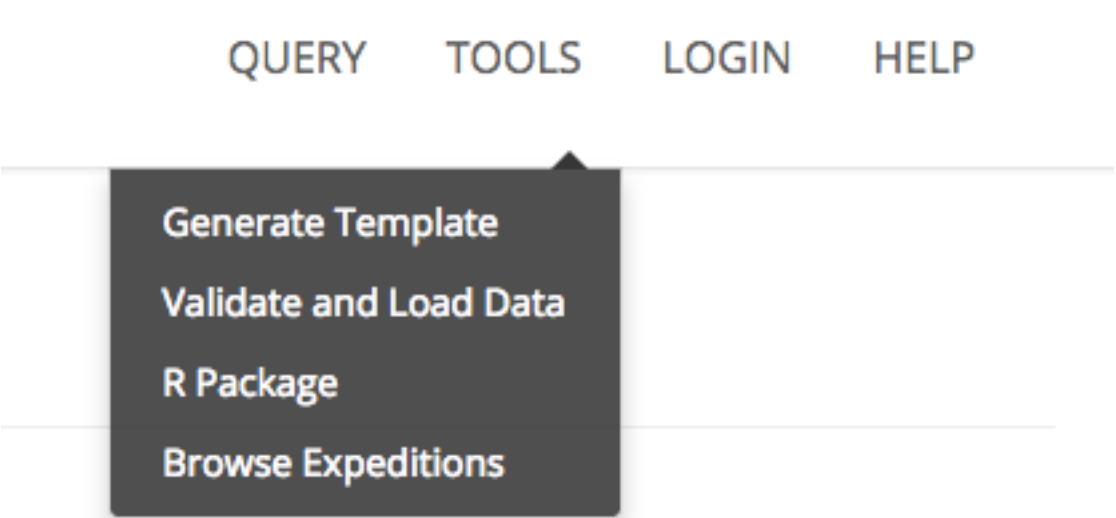
GeOMe Help Document	1
Introduction	1
Generate Template	1
Validate and Load Data	2
FASTA Upload Example	3
FASTQ Upload Example	4
GeOMe R Package	5
Browse Expeditions	6
Query	6

Introduction


The Genomic Observatory Meta-Database (GeOMe) is a web-based database which captures metadata on biological samples, used for biodiversity inventories, population studies, and environmental metagenomics. GeOMe assigns persistent identifiers for all samples and sampling events and specifies the set of metadata attributes which satisfy the requirements of the [genomic observatory model](#), including capturing the who, what, where, and when associated with all samples. GeOMe provides instant feedback to users on the quality of their data and packages data for further analysis for use in a laboratory information system (LIMS) using the [Biocode LIMS plugin](#). GeOMe also packages submissions for easy delivery to the Sequence Read Archive (SRA) and Genbank's Nucleotide database.

Generate Template

Sample metadata is recorded on an Excel Spreadsheet and you can create and customize your own templates under “Tools -> Generate Template”



On the Generate Template page, you can select columns that you want to include on your spreadsheet. Click on the “DEF” link beside each column name to view the definition of the column name. Columns that are pre-checked and shown in grey, indicate that they are mandatory fields and not able to be un-checked. Columns that are pre-checked and shown in blue indicate they are suggested and can be un-checked. Once you have checked the columns you wish to include in your spreadsheet, press the “Export Excel” button to download an Excel Spreadsheet which you can then use to fill in Sample Metadata.



[QUERY](#)
[TOOLS](#)
[LOGIN](#)
[HELP](#)

GENERATE TEMPLATE

Choose template from dropdown menu *OR* check available column heading below to include in your customized FIMS spreadsheet.

Default

Remove

Export Excel

Select ALL | Select NONE | Save

DEFAULT COLUMNS

- ☒ phylum DEF
- ☒ principalInvestigator DEF
- ☒ materialSampleID DEF
- ☒ locality DEF
- ☒ decimalLatitude DEF
- ☒ decimalLongitude DEF
- ☒ coordinateUncertaintyInMeters DEF
- ☒ georeferenceProtocol DEF
- ☒ yearCollected DEF
- ☒ monthCollected DEF
- ☒ dayCollected DEF
- ☒ genus DEF
- ☒ species DEF
- ☒ permitInformation DEF
- ☒ basisOfIdentification DEF
- ☒ country DEF
- ☒ lifeStage DEF
- ☒ sex DEF
- ☒ geneticTissueType DEF

- Although there are a lot of field options only four are **REQUIRED**:
materialSampleID
principalInvestigator
phylum
and either
decimalLatitude AND decimalLongitude (preferred)
or
locality
- Thirteen additional fields are recommended, and thus automatically checked in the default template
- Each individual (organism/sample) is entered in a row in the spreadsheet and must have a unique materialsampleID that is created by the user (your unique sample name)

DEFINITION

Click on "DEF" next to any of the headings to see the definition of the term.

Column Name: locality
URI = urn:locality
Defined_by = <http://rs.tdwg.org/dwc/terms/locality>

Definition:
Local name of site. Something that could be found by Google

Validate and Load Data

The Validate and Load Data option can be found under “Tools -> Validate and Load Data”. The first step is validating your sample metadata. Use the Browse button to browse for your filled out spreadsheet file that was generated as a blank template above and select the “Validate” button. After data validation, you can Upload your dataset and include just the metadata or include FASTA or FASTQ metadata.

FASTA Upload Example

If you have organelle or single-copy nuclear DNA sequence data from Sanger sequencing, you can upload a FASTA file of your sequences directly to GeOMe. Nuclear data should be uploaded as a single unphased sequence with ambiguity codes intact.

You must create or select a pre-existing expedition name for your dataset before continuing. An expedition is any set of sample metadata with internally unique identifiers (materialSampleID) that has not previously been uploaded to GeOMe. Expeditions may contain data from multiple species or sample types so long as they meet the above criteria. Examples of expeditions include all of the sample metadata from a sampling expedition, or all of the sample metadata from a particular publication.

Select your FIMS Metadata file for this expedition, along with a FASTA filename and a Marker name. If your particular locus is not available in the dropdown list, select “New Marker”. After selecting the FIMS Metadata file, you must check a box stating that you have visually verified the sample locations on the map at the bottom of the page. The name of your FASTA sequences must match the sample identifiers in the metadata file. Each FASTA file should only include data from a single marker type. If you have multiple markers for the same taxa you must upload multiple FASTA files for a single metadata file, which can be added by clicking on the “+” button.

VALIDATE AND LOAD DATA

Using this tool you can check for errors in your metadata file and upload your data. The validate tab can be used to ensure that all required fields are completed and that each materialSampleID is unique in your metadata file (in tab delimited text format) while the upload tab will also validate your files and ensure that each materialSampleID is accompanied by a fasta/fastq file of the same name.

Validate
Upload
Results

Data Type(s) ☒ FIMS Metadata ☒ Fasta ☐ Fastq Metadata

Expedition Name ☐ New Expedition?

FIMS Metadata

☒ Please verify sample locations on the map below and then check this box

FASTA Data

Marker

FASTA Data

Marker

Instructions:

- The name of your fasta sequences must match the materialsampleIDs in the metadata file
- You can include multiple taxa in a single fasta/metadata file
- Each fasta file should only include data from a single marker type (e.g. CO1, CYB, etc)
- If you have multiple markers for the same taxa you must upload multiple fasta files for a single metadata file.
- We recommend Fasta file names should follow this format
markerabbreviation_usertaxaabbreviation.fa

FASTQ Upload Example

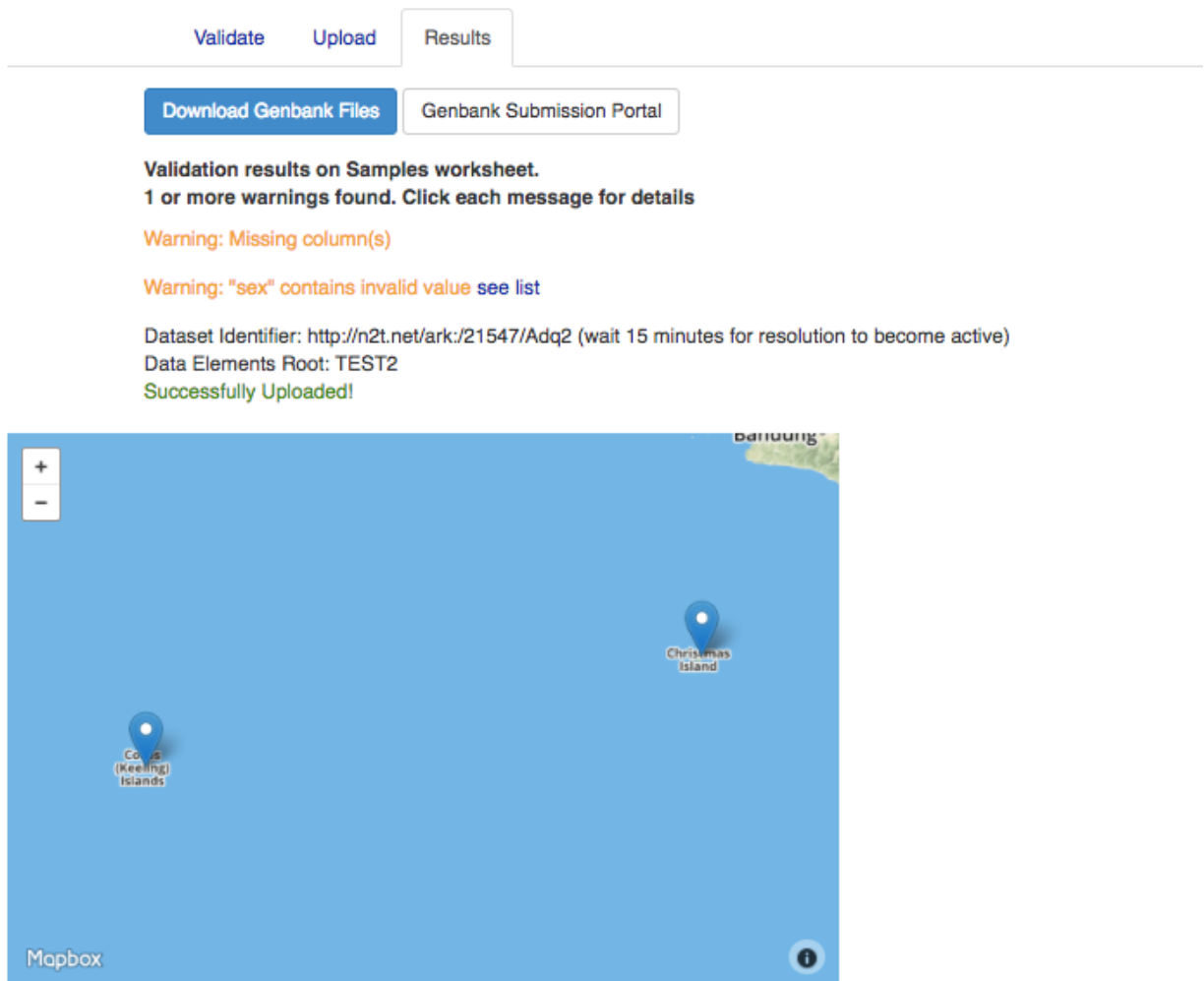
If you have any type of sequence data generated by massively parallel “next-generation” methods, it is preferable to store your raw FASTQ data at the NCBI sequence read archive, where they can also be discovered via BLAST. GeOME accepts your expedition metadata, and packages your submission for easy upload to the SRA, bypassing the need to deal with SRA’s submission portal wizard. GeOME will automatically harvest the relevant accession numbers from the SRA, creating a solid link between the genetic data and its metadata. A FASTQ upload generally follows the same protocol as a FASTA upload, with the following additional points:

- GeOME will accept single and paired end read data
- Each FASTQ file should contain reads from a single individual
- Names of fastq files must match the materialsampleIDs in the metadata file up to the file extension (e.g., R1.fq.gz, .1.fq, etc)
- The actual fastq sequence files will not be uploaded here and stored on the GeOME system. Instead the metadata file will be uploaded and stored here.
- For validation purposes a text file of the FASTQ file names (one name per line and including the file extension) will be uploaded here. If you are uploading paired end data

there should be two file names per sample. This process ensures that required fields are complete, that each materialsampleID is unique, and that the materialsampleIDs match the fastq file names.

- Once uploading is complete the FIMS system will produce two files (SRA metadata and BioSample attributes files) that will ease the upload process to NCBI's Short Read Archive (SRA). When these files are downloaded a set of simple instructions are included that will speed your SRA submission.

Once you have validated and uploaded your file of FASTQ filenames, you will see a screen with two buttons and your validation results. One button enables you to download pre-generated Genbank submission files. The second button is available which opens a browser window taking you to Genbank's SRA Portal.



GeOMe R Package

A link is available under the tools menu which takes you to the GeOMe R package github page, located at <https://github.com/DIPnet/fimsR-access>. More instructions are available at that link.

Browse Expeditions

The “Browse Expeditions” option shows all available uploaded expeditions that are part of GeOMe. This pages shows you the number of samples, FASTA sequences, and FASTQ metadata provided for each sample. Here you have the option of downloading CSV, FASTA, or FASTQ formatted metadata.



[QUERY](#) [TOOLS](#) [LOGIN](#) [HELP](#)

EXPEDITION BROWSER

In this system an “Expedition” includes the metadata (and Sanger sequences if applicable) from a single dataset. The GUID is the globally unique persistent identifier for the expedition and should be acknowledged in the original publication of the dataset and accredited when any part of that dataset is downloaded for reuse.

Expedition Title	Samples	Fasta Sequences	Fastq Metadata	GUID	
Acanthurus_reversus_RADSeq_Sanger spreadsheet	30	83	9	http://n2t.net/ark:/21547/AgX2	Download ▾
Acanthurus_olivaceus_rangewide_Sanger&RADSeq	673	1156	52	http://n2t.net/ark:/21547/AEW2	Download ▾
Celexa_CO1_cb spreadsheet	150	150	0	http://n2t.net/ark:/21547/AFX2	Download ▾
Celsan_CO1_cb spreadsheet	109	109	0	http://n2t.net/ark:/21547/AFW2	Download ▾
Centropyge_Cytb_DiBattista2016 spreadsheet	157	156	0	http://n2t.net/ark:/21547/Agg2	Download ▾
Ceparg_CyB_MG spreadsheet	775	775	0	http://n2t.net/ark:/21547/AFM2	Download ▾
Ctestr_CYB_JE spreadsheet	531	531	0	http://n2t.net/ark:/21547/AGI2	Download ▾
Diaspp_A68_HL spreadsheet	310	310	0	http://n2t.net/ark:/21547/AGA2	Download ▾
Diaspp_CO1_HL spreadsheet	13	13	0	http://n2t.net/ark:/21547/AFz2	Download ▾
Echdia_CytB_HL spreadsheet	25	25	0	http://n2t.net/ark:/21547/AFt2	Download ▾
Eucmet_CO1_HL spreadsheet	30	30	0	http://n2t.net/ark:/21547/AFw2	Download ▾
Glycybute_Direct test 16 spreadsheet	2	2	0	http://n2t.net/ark:/21547/AFs2	Download ▾

Query

The GeOMe query interface enables users to filter on geographic information (via a bounding box), any word string as part of the metadata (e.g. “Moorea”), Darwin core terms, expedition names, or any other column that is part of the GeOMe specification. The Query interface returns results either in map form or table form, selectable by clicking on the “Map” or “Table” buttons on the upper right corner of the interface. The “Download” link enables download of the queried results, with the user able to specify whether they want just the metadata in Excel or KML format, or whether they also want associated Sanger sequence data or SRA accession numbers.