# GeOMe Help Document

GeOMe Help Document 1

Introduction 1

Generate Template 1

Validate and Load Data 3

FASTA Upload Example 3

FASTQ Upload Example 4

GeOMe R Package 5

Browse Expeditions 5

Query 6

Accession Numbers and Sample Identifiers 6

Physical Sample Identifiers 6

Dataset/Expedition Identifiers 7

Sequence Identifiers 7

Cheat sheet for uploading your DIPnet data to the NCBI Short Read Archive (SRA) 7

## Introduction

The Genomic Observatories 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 collecting events, samples, and tissues and specifies the set of metadata attributes which satisfy the requirements of the [genomic observatories model](https://gigascience.biomedcentral.com/articles/10.1186/2047-217X-3-2), 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](http://software.mooreabiocode.org/). GeOMe also packages submissions for easy delivery to the Sequence Read Archive (SRA) and Genbank's Nucleotide database.

## Generate Template

Generate a template (see Workbench -> "Generate Template") to create spreadsheet templates for your metadata. Fields in the template generator that are checked and greyed out are mandatory fields. Fields that are checked and in blue are recommended fields. Check all fields that you wish to include in your sheets that you download. Spreadsheets can be downloaded as a workbook (containing all available sheets), or as individual sheets.

The following options are available for downloading metadata template sheets:

* **Workbook**: Selecting this option will create a single workbook containing all available sheets (described below) for project. See descriptions of the following sheets for more information about each sheet.
* **Events**: eventID, principalInvestigator, and yearCollected are required fields. Also, either decimalLatitude and decimalLongitude or Locality is required.
* **Samples**: specimenID eventID are required fields. In many cases, multiple samples are collected at the same location. You will create an event in the Events sheet with a unique identifier (e.g. "Event1") and then insert the eventID into the Samples sheet.
* **Tissues**: tissueID and specimenID are required fields. If you do not have distinct tissues for each sample, you can just name the tissueID field the same as the specimenID field. If there is more than one tissue per sample, you will need to name each tissue with a unique identifier. For example, specimenID = "sample1" with two tissues could have tissueIDs of "sample1.1" and "sample1.2".
* **sample\_photos\***: This is a worksheet that relates samples to photos of samples
* **event\_photos\***: This is a worksheet that relates events to photos of events

\* sample\_photos and event\_photos do not appear in all projects.

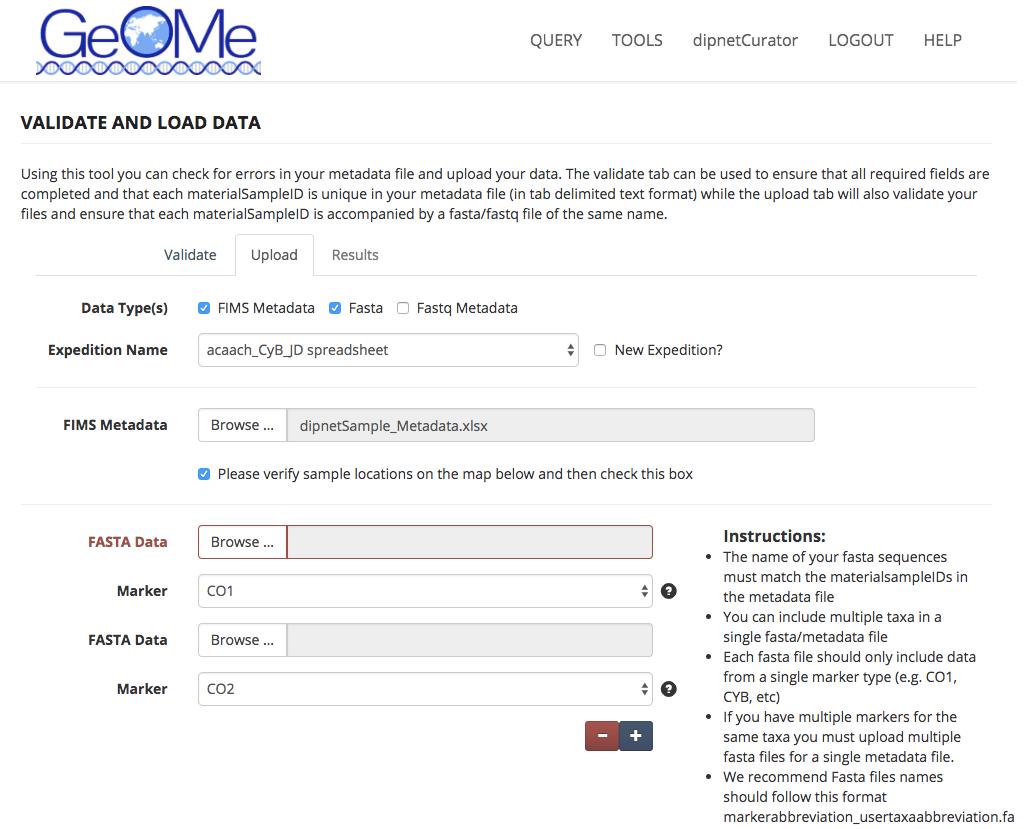
## Validate and Load Data

You can load data into the system using the Loading interface (see Workbench -> "Load Data") . If you do not have an account and a project is public, you may validate your data to make sure it passes all required checks. If you have an account you will be able to load your data for each project, assuming that it passes all validation checks. For information about acquring an account, please send an email with you name, title and affiliation to [geome.database@gmail.com](mailto:geome.database@gmail.com)

In addition to loading the metadata sheets described in the "Create Template" section above, you may also upload FASTA sequences and FASTQ metadata for projects that are configured to accept sequence data (Note: we are currently not accepting microsatellite datasets).

### FASTA Upload

You must create, or select a pre-existing expedition name for your dataset before continuing. Select your FIMS Metadata file, along with a FASTA filename and a Marker name. 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.



### FASTQ Upload

The FASTQ Upload example follows the same protocols as the FASTA upload example. The following points should be followed when uploading FASTQ data:

* FIMS 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 FIMS 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 PE 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 FASTQ file, a screen is presented that shows you 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.

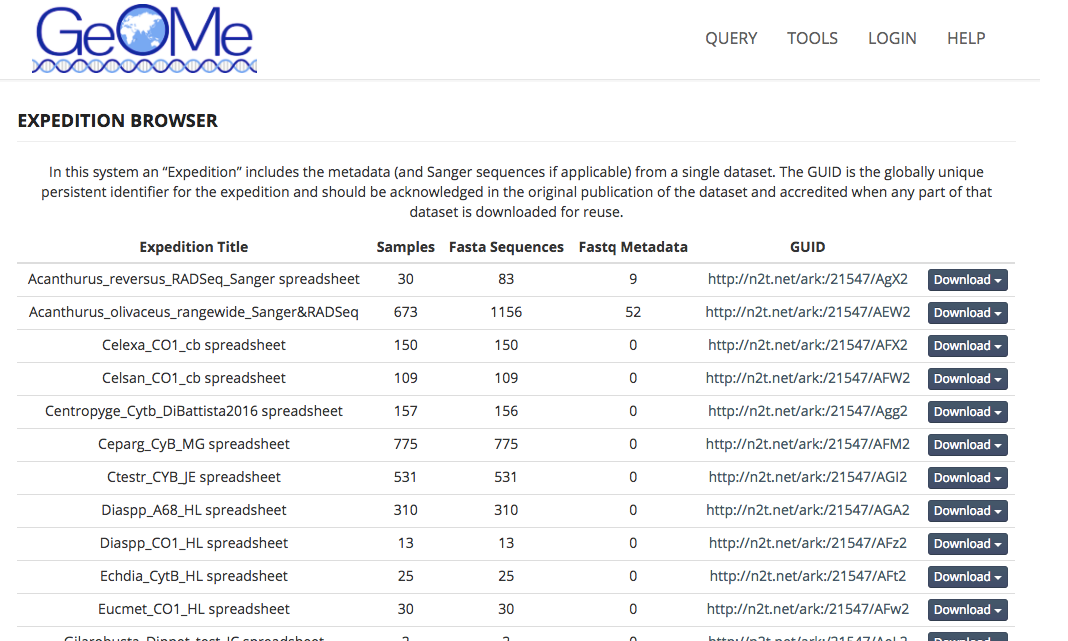
GeOMe accepts FASTQ submissions because we want to archive unmanipulated sequence data that are free of filtering biases. By doing this we are reducing ascertainment bias (a site that is not a SNP in your dataset may be a SNP when combined with data from a different population). We are also avoiding the subjective choices that go into calling SNPs, such as thresholds for trimming, filtering, coverage and likelihood. Our objective is for future users of your data to be able to make these choices in the context of their own question and dataset.

## GeOMe R Package

You can query datasets that have been loaded to the GeOMe database using our [R package](https://github.com/DIPnet/fimsR-access) (<https://github.com/DIPnet/fimsR-access>). Please note, the R package was released for version 1 of the GeOMe site and has not yet been fully tested with the current release. If you encounter issues, please contact the development team.

## 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

The GeOMe query interface enables users to filter on geographic information, 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 metadata download of the queried results.

## Accession Numbers and Sample Identifiers

When you submit your work for publication you may be asked for Genbank accession numbers, dataset identifiers, or even sample identifiers. GeOMe creates identifiers for physical samples and datasets, as well as automatically syncing sequence read archive SRA numbers. The following information describes how to handle these identifiers.

### Physical Sample Identifiers

As you may have seen, you can obtain a globally unique form of the materialSampleID in the "bcid" column at then end of the row of metadata when you download a CSV file and it looks like:

|  |
| --- |
| ark:/21547/Apj2Acaoli\_262 |

The above follows the same principal of reporting a doi, where you just put doi: and some string following it. if you want it resolvable, then you can report it with the resolver in front of it, like:

<https://n2t.net/ark:/21547/Apj2Acaoli_262>

### Dataset/Expedition Identifiers

You can find dataset identifiers by going to "Tools -> Browse Expeditions" and  you'll see a column called "GUID" that if you click on will bring you to information about your expedition. E.g. <https://n2t.net/ark:/21547/Apc2>

### Sequence Identifiers

For nextgen sequences that have followed the GeOMe path described in this document you can enter the resolvable GUID for the materialSample and find links to the BioProject and BioSample identifier, e.g. check out the following record:

<http://n2t.net/ark:/21547/le2Acaoli_CAS44>

GeOMe currently doesn’t link Genbank Accession identifiers for FASTA data submissions, so these will need to be researched independently.

# Cheat sheet for uploading your DIPnet data to the NCBI Short Read Archive (SRA)

After submitting your metadata to DIPnet two files will be produced the bioSample-attributes.tsv and the sra- metadata.tsv files and you will be directed to SRA to upload your data. There are several steps but the creation of those two files will streamline the process significantly!

If you don’t already have a NCBI account you will need to create one. If you do have an account then sign in using the tab at the top right corner of page.

After you sign in start a new submission

**Step 1**: Submitter

Enter your personal information

**Step 2**: General Info

You will be asked two important questions here:

1. Did you already register a BioProject for this data set?
2. Did you already register BioSamples for this data set?

In the majority of cases the answer to both questions will be NO

The following instructions are based on the user answering “NO” to both of the above questions.

**Step 3**: Project Info

Fill in project information. For example:

Project Title: Acanthurus\_reversus\_RADSeq\_data

Project Description: RADSeq data for the reef fish Acanthurus reversus Relevance: Evolution

Is your project part of a larger initiative that is already registered with NCBI?

Most likely No

External links: Add if relevant

Select your grants: If relevant

**Step 4**: Biosample type

Here you choose your sample type. Most DIPnet members will check either “Invertebrates” OR “Model organism or animal sample” for vertebrates.

**Step 5**: Biosample attributes

Upload the bioSample-attributes file (.tsv) produced by GeOMe. You may see additional warnings or error messages produced by the SRA validator. You must fix error messages. In some cases, you may safely ignore warnings. For example, we have seen cases for users working in marine system where locality is often based on nearby terrestrial locations, and the SRA responds with a warning that the locality is invalid since it is located in the warning. This particular message may be ignored for marine users where this is intentional.

**Step 6**: SRA metadata

Check the Upload a file option and ppload the sra-metadata file (.tsv) produced by GeOMe

**Step 7**: Files

Follow the directions on SRA and upload your files. You will be asked to download the latest version of Aspera Connect. This will speed upload tremendously. Once Aspera is installed go directly to the Choose Files option, choose your zipped folder, and Aspera will automatically open.

**Step 8**: Overview Submit!