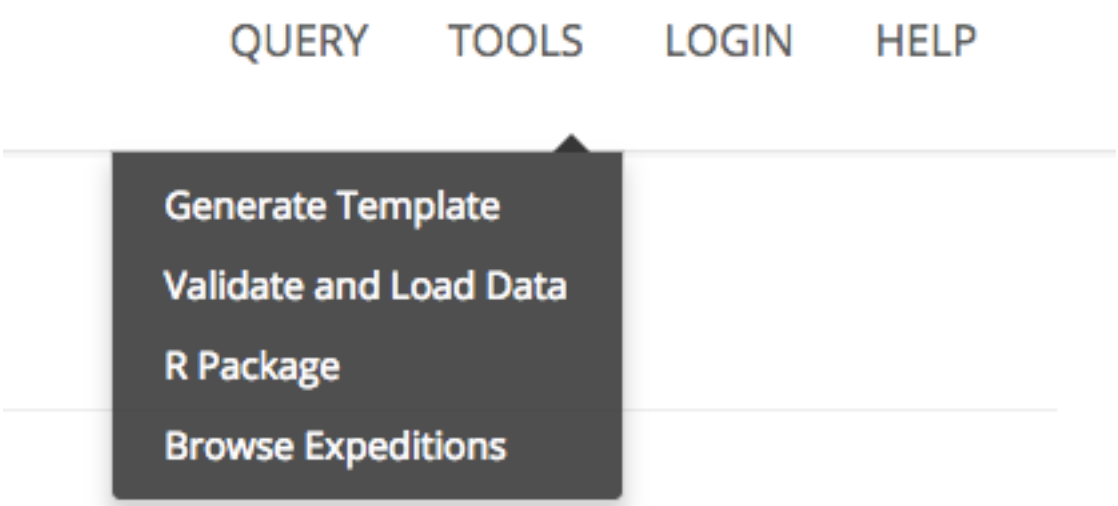
GeOMe Help Document

Introduction

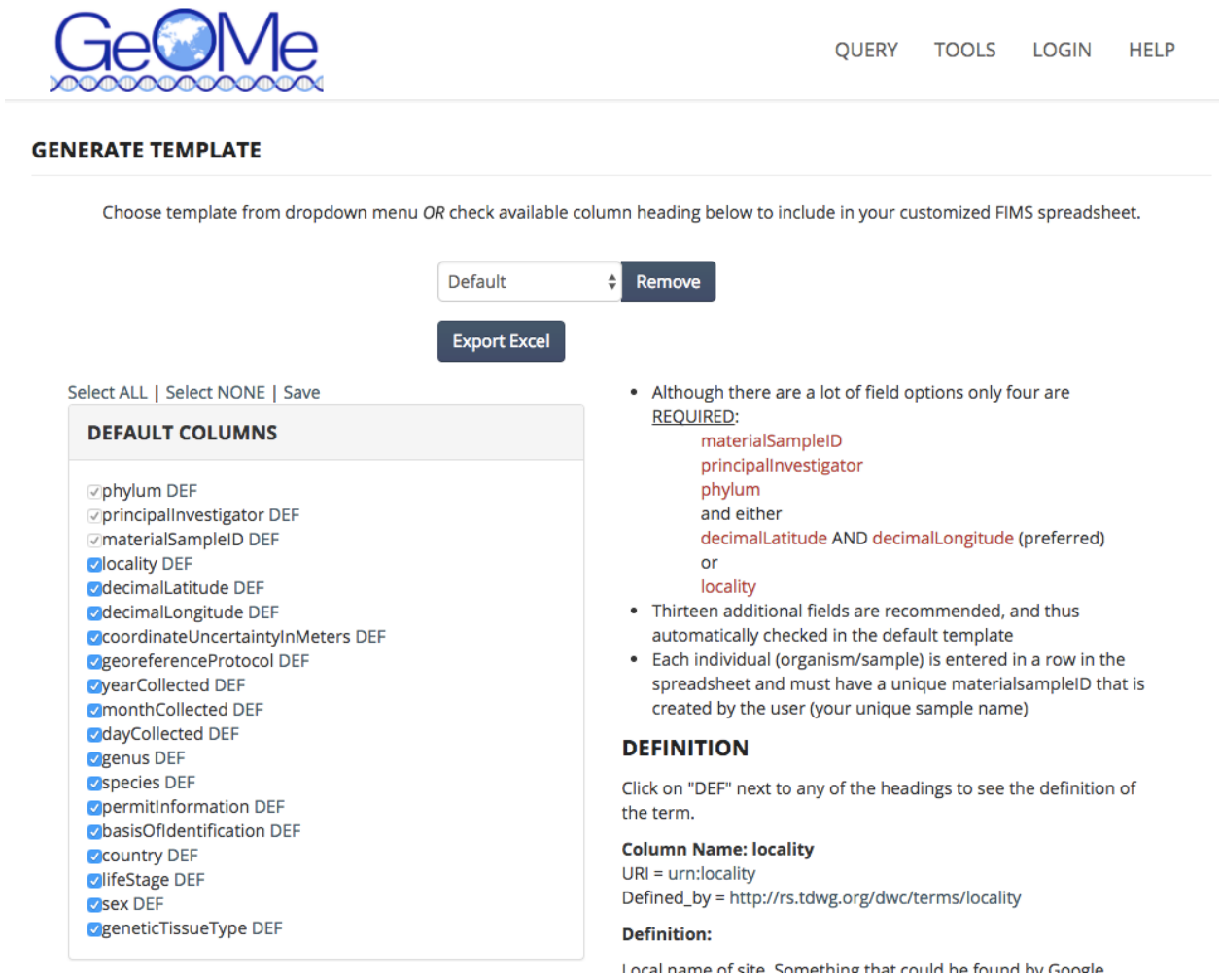
The Genomic Observatories Meta-Database (GeOMe) is a web-based database which captures metadata on biological samples, used for biodiversity inventories, population studies, and environmental metagenomics. GeOMe assigns persistent identifiers for all samples and sampling events and specifies the set of metadata attributes which satisfy the requirements of the genomic observatories 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.

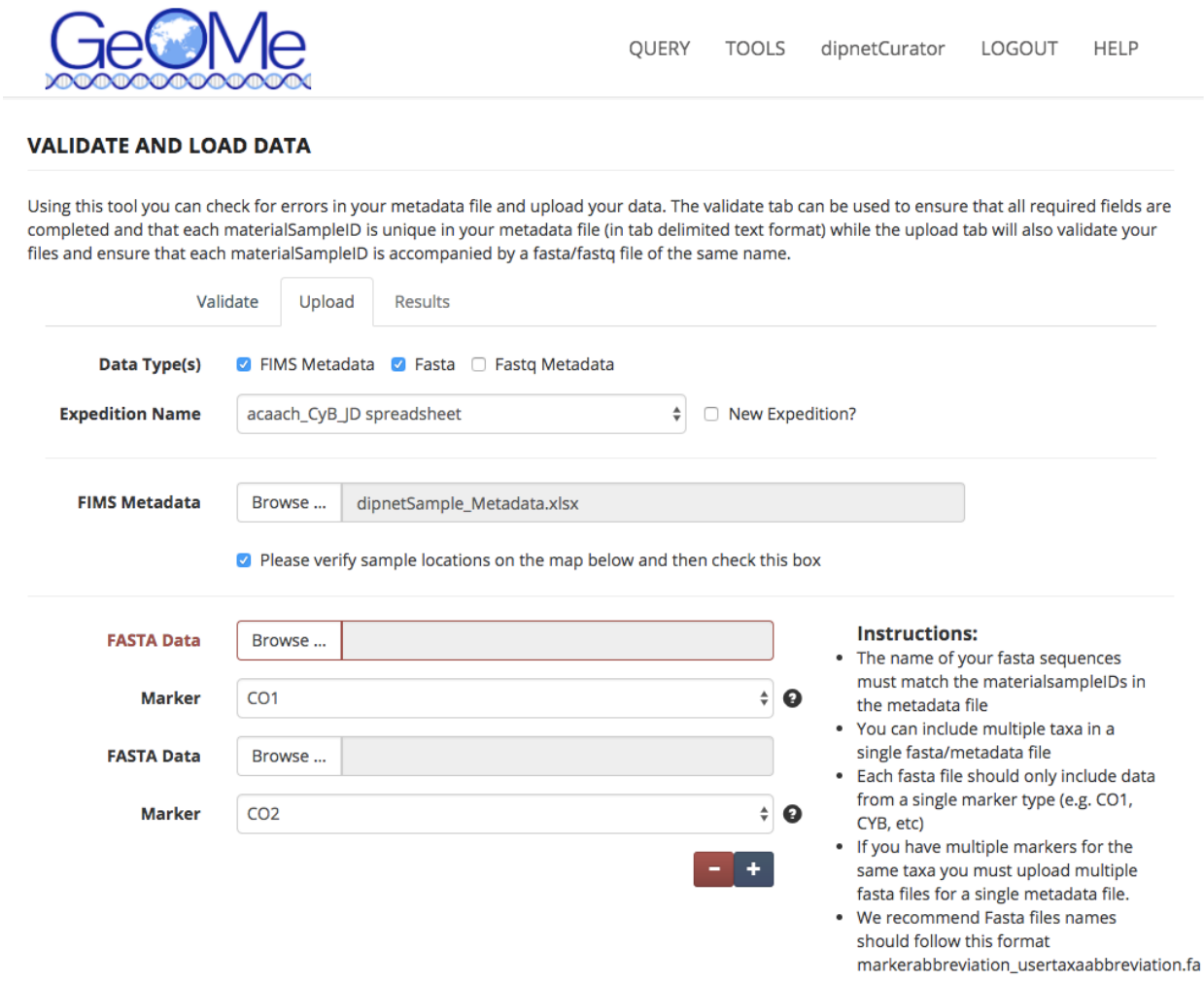


Validate and Load Data

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 file 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

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.



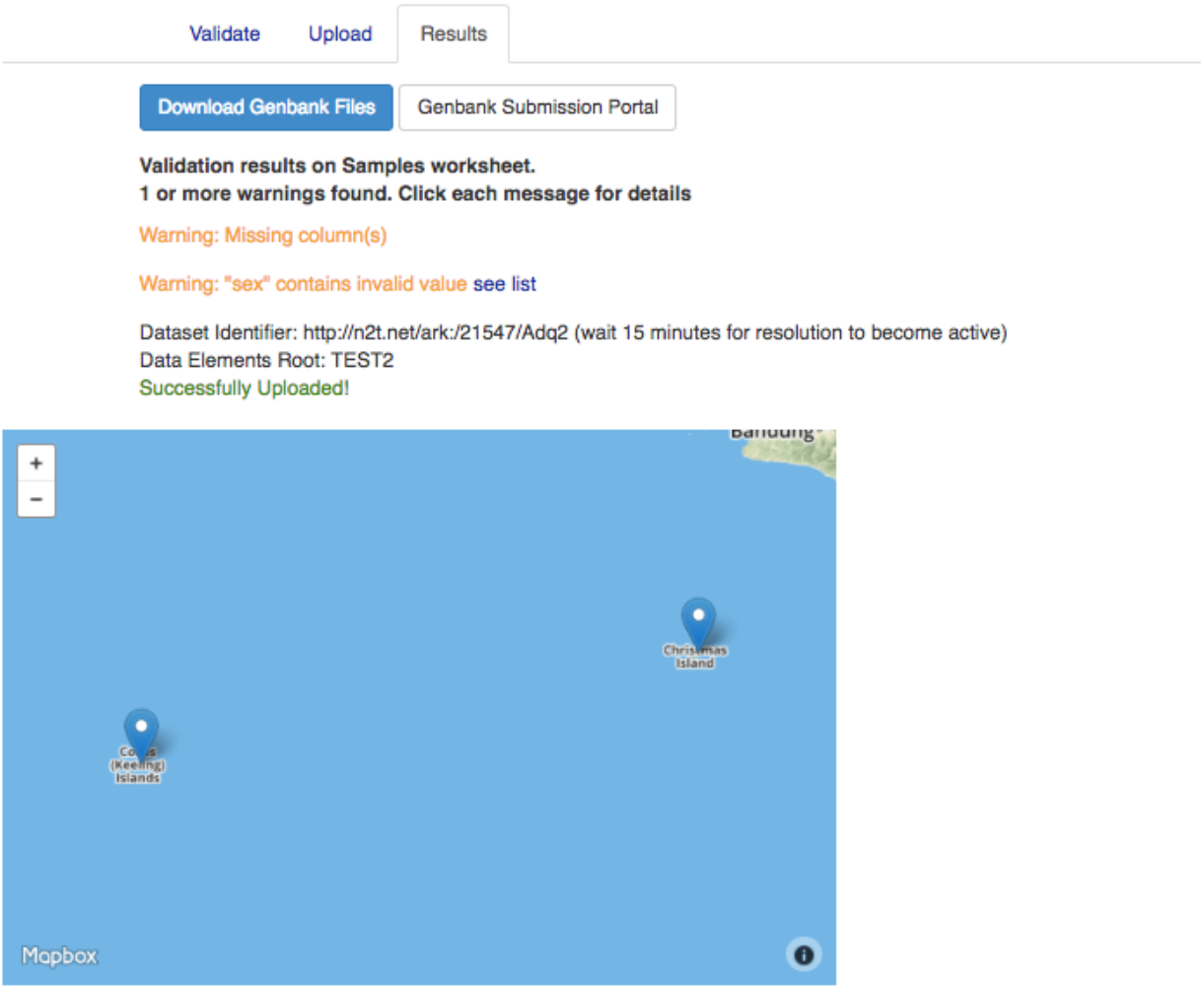
FASTQ Upload Example

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.



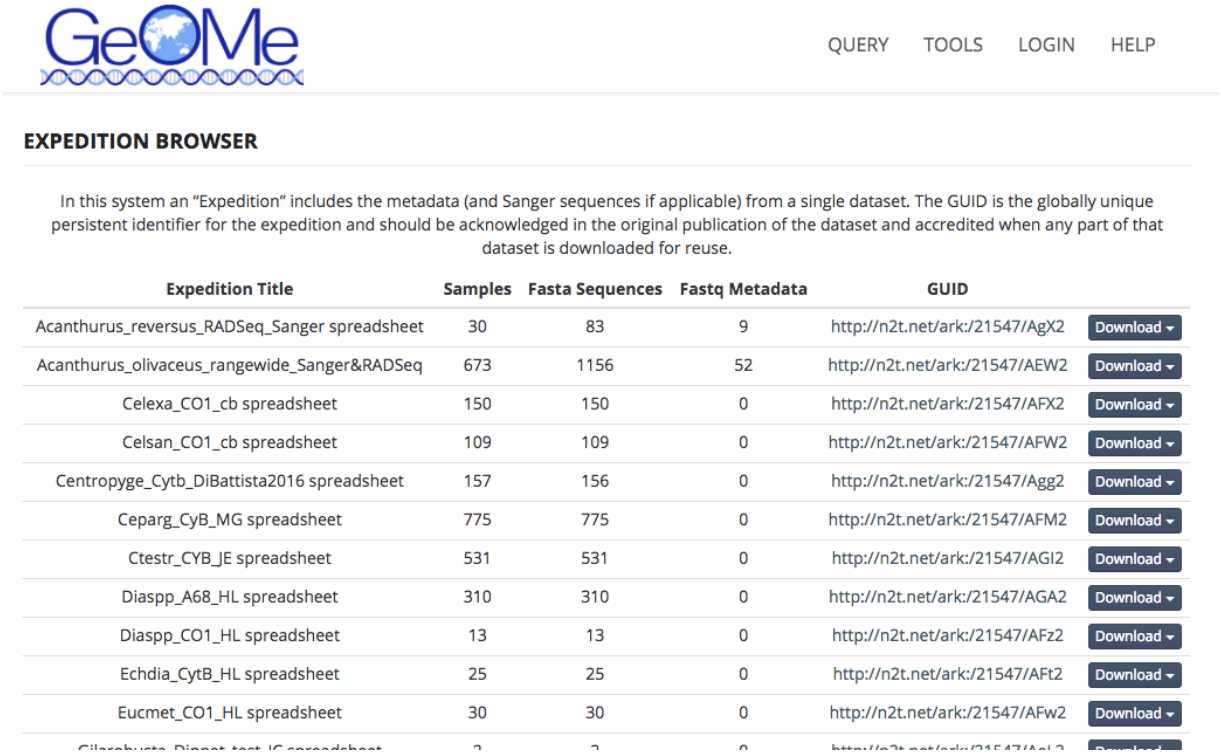
GeOMe R Package

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

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!

Citation Guide

How you cite GeOMe depends on the research you're doing and where you are submitting for publication. In light of this fact, we offer four different ways to cite information in GeOMe: citation of a dataset, citation of sample metadata, citation of NCBI accession identifiers, and citation of GeOMe itself. Following are ways of obtaining the correction citation information.

1) Citing a Dataset.
Expedition metadata and associated sequence files can be accessed under the Tools -> Browse Expeditions menu item. This will display all expeditions loaded for each project. Each expedition has an assigned GUID that you can use to cite the dataset itself. It appears in this form: http://n2t.net/ark:/21547/AgX2

2) Citing the Metadata for SRA Submissions
After loading your data into GeOMe and submitting to SRA, the GeOMe metadata and SRA metadata will be linked with 24 hours. You can go to the GeOMe query interface and search for your metadata. Click on Table view and find your sample of interest. Click the Row containing the metadata and it will bring you to a page that looks like, for example, https://www.geome-db.org/sample/ark:~2F21547~2FAut2Crup_J66 The top three lines contain an identifier for the sample metadata, the NCBI bioproject and the NCBI biosample.

3) Citing the Accession
Follow part 2 above and click the link to the biosample, which will take you to a biosample page that lists the accession.

4) Citing GeOMe Itself
Please cite the following paper:

Deck J., Gaither M.R., Ewing R., Bird C.E., Davies N., Meyer C., Riginos C., Toonen R.J., Crandall E.D. 2017 The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol* **15**(8), e2002925.