Enumeration and Gene-Sequencing of Bacteriophages: A Genomics Data Management Teaching Case

by Watinee Sae-Lim

Summary of Teaching Points

Overview

* Educating new lab members about RDM practices in the lab
* Using multiple storage locations (network, laptop, external drive)
* Challenges of storing and accessing “big data”
* Using various operating systems
* Labeling and linking images to their analyzed files
* Software license management

Types, Formats, and Stages of Data (see page 7)

* Outcomes of the project include various types of data (e.g., numerical data, images, graphs, genome sequences, and sequence alignments)
* Types of data are in proprietary file formats require export to open formats for storage and preservation purposes
* Unable to open generated files using proprietary genomics software because of expired licensing
* Issues converting a file format to compatible formats for each operating system

Contextual Details

* Control in file naming conventions is required for every format of data so that images can be linked to their analyzed files
* No data description, especially the raw data which will be reused

Storage, Backup, and Security

* Data stored in various places; there is neither a plan for storage nor documenting which data are stored where
* No security or backup plan

Data Ownership

* Lack of defined ownership

Policies for Reuse

* The raw data are stored for reuse purpose, but there is no plan or policy for reuse

Data Access and Sharing

* There is neither a clear plan nor policy for duration of remote access
* There is no formal process for sharing data

Archiving and Preservation

* Data are stored in proprietary file formats needing requiring conversion to open file formats for preservation.

A doctoral student has newly joined a microbiology and immunology lab. The lab investigates and categorizes genomes of bacteriophages isolated from shellfish samples to (1) identify host specificity (2) generate morphological[[1]](#footnote-1) data to describe a genomic structure of phages and (3) analyze genomic sequencing. A bacteriophage (phage) is a virus that infects and replicates within bacteria. The primary goal is to discover a new phage. Since genomic sequencing is novel and immensely difficult, the lab’s Principal Investigator (PI) assigned a postdoctoral student in the lab to be her research mentor. Mostly, the mentor will impart his expertise and share his experience with the postdoctoral student. The student attended several trainings such as bio-safely training and environment safety, but none related to instruction on how to properly utilize any of the lab’s tools, equipment, or data management (i.e., documenting and storing her raw data, etc.).

“When I need assistance, I ask around. I ask everyone in the lab. Our tools and equipment share basic features and functionalities, so everyone at least knows enough to get me started and then I figure it out from there.”

The student’s study is an individual project, but there is joint authorship with the lab. The student believes that the data she collects will be jointly owned by her and the lab. Her mentor monitors her research protocol, and gives suggestions and advice regarding how to perform genomic sequencing, but he is not involved in managing, sharing or using her research data, which is solely generated by her. The student records everything in her lab notebook, which is divided into two parts: the first part is what she plans to do, and the second part is what she actually does and the results. She later types all the notes into her own Mac™ laptop using Microsoft Word:macTM. Every day she collects data from a variety of assays, which are procedures to quantitatively measure the concentrations of phages in sample. Then she documents everything in her lab notebook. Next, she compares a number of isolated bacteriophages, identifies negative and positive samples, and calculates the percentage of phages using proprietary software to generate graphs. These images she saves on her laptop. To capture a phage’s genome structure, she conducts a morphology scan, using a Transmission electron microscope (TEM) technique[[2]](#footnote-2). The outcomes are recorded in a special file format, but she normally exports the images to JPEG or TIFF formats. She stores these files on one of the lab’s computers, using Windows OSTM. There are individual folders on its hard drive and lab members can store their data there, labeled under their names.

Genome sequencing, using the Next-generation sequencing[[3]](#footnote-3), which uses high-through put instrumentation, produces large amounts of data in file formats such as .qseq, .fasta, and .fastq. These large data sets are stored on the medical school’s digital space storage, under the lab and her name, respectively. To analyze the sequence data, she opens a file from the digital space and works with the file on the lab’s computer (Windows OSTM) because CLC Genomic WorkbenchTM software is stalled there. The analyzing process produces genomic data in various file formats such as sequence alignment (.phy, .msf, .aln) and CLC format package[[4]](#footnote-4) genome annotations (.clc: sequences, trees, alignments, and reports, etc.). She works on data analysis using two different platforms: Mac OSTM and Windows OSTM. She has to convert the files to compatible versions, which sometimes results in formatting errors or worse, the loss of some data. She currently cannot open CLC Genomics WorkbenchTM because the lab’s license has expired, so she cannot open some original files she created by using the software.

She tries to remember to save all of her images in the same folder as their related text files, but some of the images she has saved in the image folder she can no longer identify because she used default file names (i.e., t40001.gif, IMG4431.jpg) that were generated by an application or digital camera; therefore, they cannot be linked to the analyzed content. Since she had no explicit research data management training or training on using the lab’s tools, including special software like CLC Genomics Workbench™, she sometimes incidentally saved a file in a wrong format. For example, if she exported the results of an alignment to a CSV annotation format, it will export just the annotation information. Conversely, if she exported to the FASTA alignment format, she would get the aligned sequences in the FASTA format, but no annotations. The CLC format package will hold every file, but they can only be opened by the software itself.

For this project, she has two big folders: one is for her study’s literature review, and the other is for the results. Each folder contains sub-folders named topically (i.e., ecology, morphology, bio-control, therapy, images, raw data, and analyzed data, etc.). She labels each file with information that is meaningful to her like dates, abbreviations of an experiment, and phage number. If she revised content in a file, she would add dates to the file name. If she continued to work on that file or added some new data, she would not change the name of the file.

She does not really have a backup plan for her data. In addition, her laptop has limited space. Thus, she uses a one-terabyte external hard drive to store more data. She does not export and save all of the raw data stored in the digital space because the files are too enormous. The raw data stays in the digital space for future reuse. After the project ends, she will access the data via a web-based platform through the lab’s website.

|  |  |  |
| --- | --- | --- |
| Methods | Outcomes | Storages |

An external hard drive

The lab’s main computer

The student’s laptop

An external hard drive

An interviewee’s laptop or

Lab’s computers (Windows OSTM, and OS XTM)

Read lengths (bps) in qseq.text file

The medical school’s digital space storage

Genome annotation, sequences (.fasta), sequence alignment (.phy, .msf, .aln): CLC format package (.clc: sequences, trees, alignments, reports, etc.)

Data analysis by CLC Genomics Workbench and ABySS: download raw data from the space

Genome Sequencing by Next-generation sequencing,

The lab’s main computer (Windows OSTM): in each individual folder

Images (Jpeg or TIFF)

Morphology: transmission electron microscope (TEM)

Textual and numerical data

Plaque assay to study host-range identification

The lab notebook

Textual and numerical data

Comparing a number of bacteriophages and identifying negative & positive samples

Plaque assay to enumeration of bacteriophages

A lab notebook

Textual and numerical data

**Discussion Questions**

1. Types of data

a. What types of data are being collected for this study?

* 1. How did the lab ensure all research staff used the same data sources and data definitions?
  2. What needs to be in the data management plan related to the data capture for the various data sets?
  3. What analytical methods and mechanisms will be applied to the data either prior to or post integration?
  4. What type of outcome data will be generated?

1. File Formats and Contextual details
   1. What file formats and naming conventions will be used for the separate data sources and for the integrated file used for analysis?
   2. What impact would the naming conventions and the use of proprietary software have on later data access?
   3. What other contextual details would she specifically need to document to make the data meaningful to others?
   4. In what form could they capture these details?
2. Data Storage, Backup, Security
3. Where and on what media will the data from each data source be stored?
4. How, how often and where will the data from each source be backed up?
5. How will the lab manage data security across research staff on the study for each data source?
6. How long following the completion of the study will the data be stored?
7. Data protection/privacy
8. Who will own any copyright or intellectual property rights to the data from each source?
9. How will the dataset be licensed if rights exist?
10. Policies for reuse of data
    1. Will the data be restricted to be re-used only for certain purposes or by specific researchers?
    2. Are there any reasons not to share or re-use data?
11. Policies for access and sharing
    1. Will some kind of contribution or fee be charged for subsequent access to this data?
    2. What process should be followed to gain future access to the study data?
12. Archiving and preservation
13. What is the long-term strategy for maintaining, curating, and archiving the data?
14. What data will be included in an archive?
15. Where and how will it be archived?
16. What other contextual data or other related data will be included in the archive?
17. How long will the data be kept beyond the life of the project?

Creative Commons License This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States License](http://creativecommons.org/licenses/by-nc-sa/3.0/us/).You are free to re-use part or all of this work elsewhere, with or without modification. In order to comply with the attribution requirements of the Creative Commons license (CC-BY), we request that you cite:

* **Editor**: Lamar Soutter Library, University of Massachusetts Medical School
* **the title of the work**: New England Collaborative Data Management Curriculum
* **the URL where the original work can be found**: http://library.umassmed.edu/necdmc

Revised June 12, 2015

1. **Morphology** is a branch of biology dealing with the study of the form and structure of organisms and their specific structural features. Source: <http://en.wikipedia.org/wiki/Morphology_(biology)> [↑](#footnote-ref-1)
2. Transmission electron microscopy (TEM) is a microscopy technique in which a beam of electrons is transmitted through an ultra-thin specimen, interacting with the specimen as it passes through. Source: <http://en.wikipedia.org/wiki/Transmission_electron_microscopy> [↑](#footnote-ref-2)
3. <http://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods> [↑](#footnote-ref-3)
4. http://www.clcbio.com/support/file-formats/ [↑](#footnote-ref-4)