‹ Project Organization and Management for Genomics (../) (../01-tidiness/index.html) › (../03-ncbi-sra/in

Planning for NGS Projects

Overview

Teaching: 20 min
Exercises: 10 min
Questions

- How do I plan and organize a genome sequencing project?
- What information does a sequencing facility need?
- What are the guidelines for data storage?

Objectives

- Understand the data we send to and get back from a sequencing center.
- Make decisions about how (if) data will be stored, archived, shared, etc.

There are a variety of ways to work with a large sequencing dataset. You may be a novice who has not used bioinformatics tools beyond doing BLAST searches. You may have bioinformatics experience with other types of data and are working with high-throughput (NGS) sequence data for the first time. In the most important ways, the methods and approaches we need in bioinformatics are the same ones we need at the bench or in the field - *planning, documenting, and organizing* are the key to good reproducible science.

Discussion

Before we go any further, here are some important questions to consider. If you are learning at a workshop, please discuss these questions with your neighbor.

Working with sequence data

What challenges do you think you'll face (or have already faced) in working with a large sequence dataset?
What is your strategy for saving and sharing your sequence files?
How can you be sure that your raw data have not been unintentionally corrupted?
Where/how will you (did you) analyze your data - what software, what computer(s)?

Sending samples to the facility

The first step in sending your sample for sequencing will be to complete a form documenting the metadata for the facility. Take a look at the following example submission spreadsheet.

Sample submission sheet (../files/sample_submission.txt)

Download the file using right-click (PC)/command-click (Mac). This is a tab-delimited text file. Try opening it with Excel or another spreadsheet program.

Exercise

1. What are some errors you can spot in the data? Typos, missing data, inconsistencies?
2. What improvements could be made to the choices in naming?
3. What are some errors in the spreadsheet that would be difficult to spot? Is there any way you can test this?

Solution

Errors:

- Sequential order of well_position changes
- Format of client_sample_id changes and cannot have spaces, slashes, non-standard ASCII characters
- Capitalization of the replicate column changes
- Volume and concentration column headers have unusual (not allowed) characters
- Volume, concentration, and RIN column decimal accuracy changes
- The prep_date and ship_date formats are different, and prep_date has multiple formats
- Are there others not mentioned?

Improvements in naming

- Shorten client_sample_id names, and maybe just call them “names”
  - For example: “wt” for “wild-type”. Also, they are all “1hr”, so that is superfluous information
- The prep_date and ship_date might not be needed
- Use “microliters” for “Volume (µL)” etc.

Errors hard to spot:

- No space between “wild” and “type”, repeated barcode numbers, missing data, duplicate names
- Find by sorting, or counting

Retrieving sample sequencing data from the facility

When the data come back from the sequencing facility, you will receive some documentation (metadata) as well as the sequence files themselves. Download and examine the following example file - here provided as a text file and Excel file:

- Sequencing results - text (../files/sequencing_results_metadata.txt)
- Sequencing results - Excel (../files/sequencing_results_metadata.xls)

Exercise

1. How are these samples organized?
2. If you wanted to relate file names to the sample names submitted above (e.g. wild type…) could you do so?
3. What do the _R1/_R2 extensions mean in the file names?
4. What does the '.gz' extension on the filenames indicate?
5. What is the total file size - what challenges in downloading and sharing these data might exist?

Solution

1. Samples are organized by sample_id
2. To relate filenames use the sample_id, and do a VLOOKUP on submission sheet
3. The _R1/_R2 extensions mean "Read 1" and "Read 2" of each sample
4. The '.gz' extension means it is a compressed "gzip" type format to save disk space
5. The size of all the files combined is 1113.60 Gb (over a terabyte!). To transfer files this large you should validate the file size following transfer. Absolute file integrity checks following transfers and methods for faster file transfers are possible but beyond the scope of this lesson.

Storing data

The raw data you get back from the sequencing center is the foundation of your sequencing analysis. You need to keep this data, so that you can always come back to it if there are any questions or you need to re-run an analysis, or try a new analysis approach.

Guidelines for storing data

- Store the data in a place that is accessible by you and other members of your lab. At a minimum, you and the head of your lab should have access.
- Store the data in a place that is redundantly backed up. It should be backed up in two locations that are in different physical areas.
- Leave the raw data raw. You will be working with this data, but you don't want to modify this stored copy of the original data. If you modify the data, you'll never be able to access those original files. We will cover how to avoid accidentally changing files in a later lesson in this workshop (see File Permissions) (https://datacarpentry.org/shell-genomics/03-working-with-files/#file-permissions).

Some data storage solutions

If you have a local high performance computing center or data storage facility on your campus or with your organization, those are ideal locations. Get in touch with the people who support those facilities to ask for information.

If you don't have access to these resources, you can back up on hard drives. Have two backups, and keep the hard drives in different physical locations.

You can also use resources like Amazon S3 (https://aws.amazon.com/s3/), Microsoft Azure (https://azure.microsoft.com/en-us/pricing/details/storage/blobs/), Google Cloud (https://cloud.google.com/storage/) or others for cloud storage. The open science framework (https://osf.io) is a free option for storing files up to 5 GB. See more in the lesson "Introduction to Cloud Computing for Genomics" (http://www.datacarpentry.org/cloud-genomics/04-which-cloud/).

Summary

Before analysis of data has begun, there are already many potential areas for errors and omissions. Keeping organized and keeping a critical eye can help catch mistakes.

One of Data Carpentry's goals is to help you achieve *competency* in working with bioinformatics. This means that you can accomplish routine tasks, under normal conditions, in an acceptable amount of time. While an expert might be able to get to a solution on instinct alone - taking your time, using Google or another Internet search engine, and asking for help are all valid ways of solving your problems. As you complete the lessons you'll be able to use all of those methods more efficiently.

Where to go from here?

More reading about core competencies

L. Welch, F. Lewitter, R. Schwartz, C. Brooksbank, P. Radivojac, B. Gaeta and M. Schneider, 'Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3945096/)', PLoS Comput Biol, vol. 10, no. 3, p. e1003496, 2014.

Key Points

- Data being sent to a sequencing center also needs to be structured so you can use it.
- Raw sequencing data should be kept raw somewhere, so you can always go back to the original files.

‹ (../01-tidiness/index.html)

› (../03-ncbi-sra/in

Licensed under CC-BY 4.0 () 2018–2019 by The Carpentries (https://carpentries.org/)
Licensed under CC-BY 4.0 () 2016–2018 by Data Carpentry (http://datacarpentry.org)