(../)

Project Organization and Management for Genomics (../)

(../02-project-plannin

Data Tidiness

Overview

Teaching: 20 min
Exercises: 10 min
Questions

- What metadata should I collect?
- How should I structure my sequencing data and metadata?

Objectives

- Think about and understand the types of metadata a sequencing experiment will generate.
- Understand the importance of metadata and potential metadata standards.
- Explore common formatting challenges in spreadsheet data.

Introduction

When we think about the data for a sequencing project, we often start by thinking about the sequencing data that we get back from the sequencing center, but just as important, if not more so, is the data you've generated about the sequences before it ever goes to the sequencing center. This is the data about the data, often called the metadata. Without the information about what you sequenced, the sequence data itself is useless.

Discussion

With the person next to you, discuss:

What kinds of data and information have you generated before you sent your DNA/RNA off for sequencing?

Solution

Types of files and information you have generated:

- Spreadsheet or tabular data with the data from your experiment and whatever you were measuring for your study.
- Lab notebook notes about how you conducted those experiments.
- Spreadsheet or tabular data about the samples you sent off for sequencing. Sequencing centers often have a particular format they need with the name of the sample, DNA concentration and other information.
- Lab notebook notes about how you prepared the DNA/RNA for sequencing and what type of sequencing you're doing, e.g. paired end Illumina HiSeq. There likely will be other ideas here too. Was this more information and data than you were expecting?

All of the data and information just discussed can be considered metadata, i.e. data about the data. We want to follow a few guidelines for metadata.

Notes

Notes about your experiment, including how you prepared your samples for sequencing, should be in your lab notebook, whether that's a physical lab notebook or electronic lab notebook. For guidelines on good lab notebooks, see the Howard Hughes Medical Institute "Making the Right Moves: A Practical Guide to Scientific Management for Postdocs and New Faculty" section on Data Management and Laboratory Notebooks (http://www.hhmi.org/sites/default/files/Educational%20Materials/Lab%20Management/Making%20the%20Right%20Moves/moves2_ch8.p

Including dates on your lab notebook pages, the samples themselves and in any records about those samples helps you associate everything with each other later. Using dates also helps create unique identifiers, because even if you process the same sample twice, you don't usually do it on the same day, or if you do, you're aware of it and give them names like A and B.

📌 Unique identifiers

Unique identifiers are a unique name for a sample or set of sequencing data. They are names for that data that only exist for that data. Having these unique names makes them much easier to track later.

Data about the experiment

Data about the experiment is usually collected in spreadsheets, like Excel.

What type of data to collect depends on your experiment and there are often guidelines from metadata standards.

📌 Metadata standards

Many fields have particular ways that they structure their metadata so it's consistent and can be used across the field.

The Digital Curation Center maintains a list of metadata standards (http://www.dcc.ac.uk/resources/metadata-standards/list) and some that are particularly relevant for genomics data are available from the Genomics Standards Consortium (http://gensc.org/projects/).

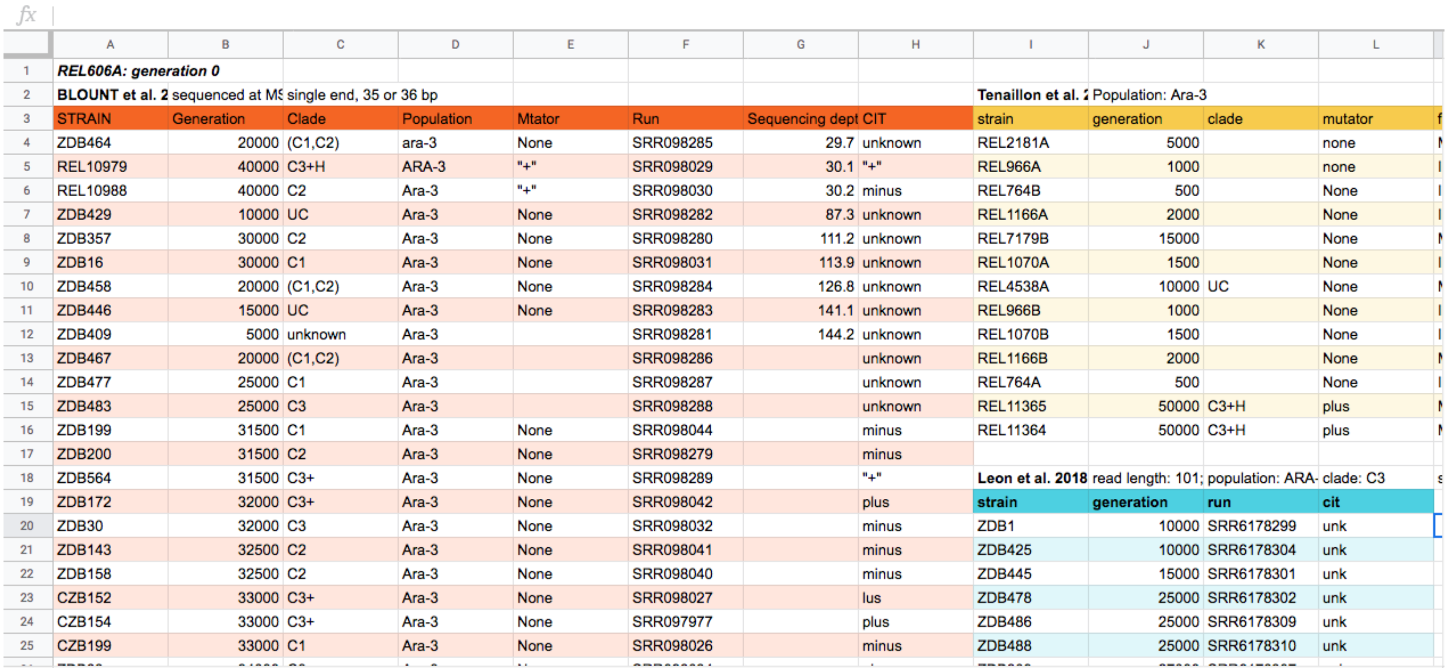
If there aren't metadata standards already, you can think about what the minimum amount of information is that someone would need to know about your data to be able to work with it, without talking to you.

Structuring data in spreadsheets

Independent of the type of data you're collecting, there are standard ways to enter that data into the spreadsheet, to make it easier to analyze later. We often enter data that makes it easy for us as humans to read and work with it, because we're human! Computers need data structured in a way that they can use it. So to use this data in a computational workflow, we need to think like computers when we use spreadsheets.

The cardinal rules of using spreadsheet programs for data:

- Leave the raw data raw - don't change it!
- Put each observation or sample in its own row.
- Put all your variables in columns - the thing that vary between samples, like 'strain' or 'DNA-concentration'.
- Have column names be explanatory, but without spaces. Use '-', '_' or camel case (https://en.wikipedia.org/wiki/Camel_case) instead of a space. For instance 'library-prep-method' or 'LibraryPrep'is better than 'library preparation method' or 'prep', because computers interpret spaces in particular ways.
- Don't combine multiple pieces of information in one cell. Sometimes it just seems like one thing, but think if that's the only way you'll want to be able to use or sort that data. For example, instead of having a column with species and strain name (e.g. *E. coli* K12) you would have one column with the species name (*E. coli*) and another with the strain name (K12). Depending on the type of analysis you want to do, you may even separate the genus and species names into distinct columns.
- Export the cleaned data to a text-based format like CSV (comma-separated values) format. This ensures that anyone can use the data, and is required by most data repositories.



	A	B	C	D	E	F	G	H	I	J	K	L
1	***REL606A: generation 0***											
2	**BLOUNT et al. 2**	sequenced at MS	single end, 35 or 36 bp						**Tenaillon et al. 2**	Population: Ara-3		
3	STRAIN	Generation	Clade	Population	Mtator	Run	Sequencing dept	CIT	strain	generation	clade	mutator
4	ZDB464	20000	(C1,C2)	ara-3	None	SRR098285	29.7	unknown	REL2181A	5000		none
5	REL10979	40000	C3+H	ARA-3	"+"	SRR098029	30.1	"+"	REL966A	1000		none
6	REL10988	40000	C2	Ara-3	"+"	SRR098030	30.2	minus	REL764B	500		None
7	ZDB429	10000	UC	Ara-3	None	SRR098282	87.3	unknown	REL1166A	2000		None
8	ZDB357	30000	C2	Ara-3	None	SRR098280	111.2	unknown	REL7179B	15000		None
9	ZDB16	30000	C1	Ara-3	None	SRR098031	113.9	unknown	REL1070A	1500		None
10	ZDB458	20000	(C1,C2)	Ara-3	None	SRR098284	126.8	unknown	REL4538A	10000	UC	None
11	ZDB446	15000	UC	Ara-3	None	SRR098283	141.1	unknown	REL966B	1000		None
12	ZDB409	5000	unknown	Ara-3		SRR098281	144.2	unknown	REL1070B	1500		None
13	ZDB467	20000	(C1,C2)	Ara-3		SRR098286		unknown	REL1166B	2000		None
14	ZDB477	25000	C1	Ara-3		SRR098287		unknown	REL764A	500		None
15	ZDB483	25000	C3	Ara-3		SRR098288		unknown	REL11365	50000	C3+H	plus
16	ZDB199	31500	C1	Ara-3	None	SRR098044		minus	REL11364	50000	C3+H	plus
17	ZDB200	31500	C2	Ara-3	None	SRR098279		minus				
18	ZDB564	31500	C3+	Ara-3	None	SRR098289		"+"	**Leon et al. 2018**	read length: 101;	population: ARA-	clade: C3
19	ZDB172	32000	C3+	Ara-3	None	SRR098042		plus	**strain**	**generation**	**run**	**cit**
20	ZDB30	32000	C3	Ara-3	None	SRR098032		minus	ZDB1	10000	SRR6178299	unk
21	ZDB143	32500	C2	Ara-3	None	SRR098041		minus	ZDB425	10000	SRR6178304	unk
22	ZDB158	32500	C2	Ara-3	None	SRR098040		minus	ZDB445	15000	SRR6178301	unk
23	CZB152	33000	C3+	Ara-3	None	SRR098027		lus	ZDB478	25000	SRR6178302	unk
24	CZB154	33000	C3+	Ara-3	None	SRR097977		plus	ZDB486	25000	SRR6178309	unk
25	CZB199	33000	C1	Ara-3	None	SRR098026		minus	ZDB488	25000	SRR6178310	unk

(https://github.com/datacarpentry/organization-genomics/raw/gh-pages/files/Ecoli_metadata_composite_messy.xlsx)

Discussion

This is some potential spreadsheet data generated about a sequencing experiment. With the person next to you, for about 2 minutes, discuss some of the problems with the spreadsheet data shown above. You can look at the image, or download the file to your computer via this link (https://github.com/datacarpentry/organization-genomics/raw/gh-pages/files/Ecoli_metadata_composite_messy.xlsx) and open it in a spreadsheet reader like Excel.

Solution

A full set of types of issues with spreadsheet data is at the Data Carpentry Ecology spreadsheet lesson (http://www.datacarpentry.org/spreadsheet-ecology-lesson/02-common-mistakes/). Not all are present in this example. Discuss with the group what they found. Some problems include not all data sets having the same columns, datasets split into their own tables, color to encode information, different column names, spaces in some columns names. Here is a "clean" version of the same spreadsheet:

Cleaned spreadsheet (https://raw.githubusercontent.com/datacarpentry/wrangling-genomics/gh-pages/files/Ecoli_metadata_composite.tsv)
Download the file using right-click (PC)/command-click (Mac).

Further notes on data tidiness

Data organization at this point of your experiment will help facilitate your analysis later, as well as prepare your data and notes for data deposition now often required by journals and funding agencies. If this is a collaborative project, as most projects are now, it's also information that collaborators will need to interpret your data and results and is very useful for communication and efficiency.

Fear not! If you have already started your project, and it's not set up this way, there are still opportunities to make updates. One of the biggest challenges is tabular data that isn't formatted so computers can use it, or has inconsistencies that make it hard to analyze.

More practice on how to structure data is outlined in our Data Carpentry Ecology spreadsheet lesson (http://www.datacarpentry.org/spreadsheet-ecology-lesson/02-common-mistakes/)

Tools like OpenRefine (http://www.datacarpentry.org/OpenRefine-ecology-lesson/) can help you clean your data.

Key Points

- Metadata is key for you and others to be able to work with your data.
- Tabular data needs to be structured to be able to work with it effectively.

(../)

(../02-project-plannin

Edit on GitHub (https://github.com/datacarpentry/organization-genomics/edit/gh-pages/_episodes/01-tidiness.md) / Contributing (https://github.com/datacarpentry/organization-genomics/blob/gh-pages/CONTRIBUTING.md) / Source (https://github.com/datacarpentry/organization-genomics/) / Cite (https://github.com/datacarpentry/organization-genomics/blob/gh-pages/CITATION) / Contact (mailto:team@carpentries.org)

Using The Carpentries theme (https://github.com/carpentries/carpentries-theme/) — Site last built on: 2019-09-26 00:11:03 +0000.