Research Data Management Working Group Agenda

|  |  |
| --- | --- |
| Date | 21st June 2019 |
| Time | 09:00-10:00 |
| Attendees | Charles Ishak (CI), Emma Bell (EB), Helen Loo (HL), Roxana Shen (RS), Sajid Marhon (SM) |

# Meeting objectives

* Introduce the De Carvalho Lab Research Data Management (RDM) Working Group
* Review the current state of RDM within the De Carvalho lab
* Collectively define the scope, deliverables, and timeline of the Working Group

# Agenda

|  |  |
| --- | --- |
| Time | Item |
| 09:00-09:15 | 1. Introduction to the De Carvalho Lab Research Data Management (RDM) Working Group (EB) |
| 09:15-09:30 | 1. What is the current state of RDM within the De Carvalho lab? (RS) |
| 09:30-10:00 | 1. Discussion:    1. What is the scope of this working group?    2. What are the deliverables of this working group?    3. What should the timeline be for this working group? |

# Actions

* ALL: Mull over the contents of this meeting!
* ALL: Everyone who doesn’t have a github – make a github!
* ALL: Prepare a 5 minute walkthrough of how we each manage our data.
* EB: Send out a Doodle poll to agree the time of the next meeting.
* EB: Compile statistics on our usage of the clusters to present next meeting.
* HL: Prepare a brief informal presentation on how wet lab researchers manage their samples, experiments, and data

# Minutes

## Introduction to the De Carvalho Lab Research Data Management (RDM) Working Group (EB)

We collectively defined ground rules for these meetings:

* The purpose of this meeting is not to complain about individual lab members. Every member of this lab is responsible for the current state of our RDM.
* We will record the content of these meetings (EB will minute).
* We will meet regularly at a mutually agreed time.
* This working group will dissolve after 2 months (i.e. 21st August).

## What is the current state of RDM within the De Carvalho lab? (RS)

RS presented an overview of the current state of RDM within the lab.

A close up of text on a whiteboard

Description automatically generated

RS defined 5 stages of RDM within our lab. RS identified existing problems (👎) at several of these stages. The group collective identified good practices (👍), both pre-existing and suggestions for the future.

1. Wet lab
   1. 👍 Have a named bioinformatician on the project prior to sequencing submission.
   2. 👍 Discuss the experimental design with the named bioinformatician.
2. Submission to the sequencing service
   1. 👍 Forwarded the sequencing submission form to the bioinformatician.
   2. 👎 Bioinformaticians unaware of incoming data.
3. Data received on the clusters
   1. 👎 Bioinformaticians receive incoherent, inconsistent, or incomplete metadata.
   2. 👎 Wet lab researchers own the data and often don’t know how to make it available to others (i.e. how to open up permissions on a directory).
   3. 👎 Flow cell tracking lost once data is on cluster
4. Data analysis conducted
   1. 👍 Delete raw data at this point.
   2. 👎 Wet lab person doesn’t know or understand the extent of the analysis.
5. Submission to public repositories
   1. 👍 Once completed, owners archive the data associated with a project.
   2. 👍 Bioinformaticians submit all code to github.
   3. 👍 Project owners submit sequencing data to GEO/ENA.

### General discussion arising from this agenda point:

RS: There should be one sole person responsible for received sequencing data.

RS: Zhibin/Khun have to set permissions – concerns about individuals accidentally deleting data.

RS: Raw data stored for 10 years.

RS: Should we use Github for archiving code?

EB and CI: Yes.

CI: You can store sequencing data on GEO privately for up to 2 years.

CI: What does it mean to archive sequencing data?

RS: It is off the clusters (Mordor and H4H).

CI: Flow for wet-lab person: submit (include bioinformatician on e-mail), …

RS: It’s difficult to get the sequencing facilities to change their practices as we’ve been working with them for a long time.

EB: … And they have no incentive to change.

SM: Processed data takes up a lot of space (i.e. bams, sams). We need to remove these files once analysis is complete.

EB: Proposes collecting data on how much of our clusters are made up of intermediate analysis files.

RS: We need to shift to a lab github for code.

EB: Does everybody have a github?

CI: Maybe

HL: Maybe

EB: If we move to a lab github – we need github training

HL: What happens? We submit samples – they get sequenced – nobody knows what happens?

HL: There has to be some sort of mentality change around how we treat our bioinformaticians. Sometimes there’s no attempt to explain the experiment. We need to collectively agree on the analysis.

ALL: General agreement.

CI: References RS’s sequencing data sheet.

SM: The bioinformatician needs clear bioinformatics ids associated with samples.

RS: Are we OK with numeric IDs?

SM: We want the entire name in the sample. We’d prefer not to have numbers.

SM: When we upload to GEO do we use the original sample numbers?

RS: No.

CI: No, they will ask for more info.

EB: Do we need a named bioinformatician on a project before submission?

ALL: Yes.

RS: We need to lead by example with the recommendations we make.

EB: If we see people not following these recommendations…?

RS: We reinforce them.

EB: The bioinformaticians need to reinforce this by refusing to work with garbage data.

SM: What about people outside of the lab?

RS: There are people we collaborate with who use our cluster space. That takes up a lot of space.

EB: What space do we have?

* Mordor 25 TB (1 TB available)
* H4H 50 TB (6 TB available)

RS: There’s no more institutional cluster space. UHN may be acquiring more space, but not for H4H.

## Discussion

## What should the scope of this group be?

CI: Establishing best practices.

RS: Develop a pipeline to help communication between wet and dry lab researchers. The ultimate goal should be good communication. Everyone on the same page – no one being left out.

HL: Make things more efficient.

SM: To avoid duplication in our processes and in our raw data to solve our problems of data storage.

## What are the deliverables of this working group?

CI: Drafting a protocol to give step-by-step instructions for taking wet lab samples through to repository deposition.

RS: We will have a lab github!

HL: We cannot expect everyone to adhere to this all the time, but we aim to get 80% of sequencing projects to follow this.

RS: We need to know who’s responsible for what in handling sequencing experiments. We need to know who’s accountable for our data and its storage – making sure the metadata is complete.

RS: [Gives the example of Tiago’s data]. It’s on our cluster, no one is in charge of it.

RS: Do we set-up a spreadsheet on Google Drive and DropBox?

ALL: Google Drive.

RS: There’s a lot of data on the cluster that have no owner.

RS: More than half of our assigned space is stuff that we don’t even know what it is.

CI: We need to get rid of that.

HL: How long has it been since anyone touched that data?

RS: More than a year.

CI: Archive or delete it all.

RS: Concerned there’s no record keeping for this data.

HL: That’s not our responsibility.

EB: Agreed.

## What should the timeline be for this working group?

EB proposed a timeline for this working group. The group discussed, revised, and agreed to the timeline.

|  |  |
| --- | --- |
| Date | Meeting objective |
| Week 1 | Introduce the Research Data Management (RDM) Working Group   * Review the current state of RDM within the De Carvalho Lab |
| Week 2 | Review good RDM practices within and without the lab   * How do we individually manage our data? * How do wet lab researchers manage their experiments? * What are RDM best practices for sequencing data? |
| Week 3 | How do we apply good RDM practices to our lab? |
| Week 4 | Draft guidelines for RDM within the De Carvalho Lab |
| Week 5 | Redraft guidelines |
| Week 6 | Present to group |

## 