Research Data Management Working Group Agenda

|  |  |
| --- | --- |
| Date | 28th June 2019 |
| Time | 09:30-10:30 |
| Attendees | Charles Ishak (CI), Emma Bell (EB), Helen Loo (HL), Roxana Shen (RS), Sajid Marhon (SM) |

# Aims and objectives

To review good RDM practices within the lab by answering the following questions:

* How do we individually manage our data?
* How do wet-lab researchers manage their experiments?

# Agenda

|  |  |
| --- | --- |
| Time | Item |
| 09:00-09:05 | 1. Review the minutes from last week’s meeting. (EB) |
| 09:05-09:15 | 1. A statistical analysis of how we use our clusters (SM, EB) |
| 09:15-09:30 | 1. How do we individually manage our data? (ALL) |
| 09:30-09:40 | 1. How do wet-lab researchers manage their experiments? (HL) |
| 09:40-09:45 | 1. Next week’s meeting. (ALL) |

# Actions

* EB: Assign each group member an RDM guideline document to review for next meeting.
* EB: Visualise cluster usage stats.
* RS: Get a breakdown of sequencing spending by year from Julie.
* ALL: Prepare a brief overview of the RDM guidelines

# Minutes

## Review the minutes from last week’s meeting (EB)

### Actions

* ~~ALL: Mull over the contents of this meeting!~~
* ~~ALL: Everyone who doesn’t have a github – make a github!~~
* ~~ALL: Prepare a 3 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.~~

## A statistical analysis of how we use our clusters (SM, EB)

RS: There’s a Tobedeleted directory that I don’t have permissions to delete.

SM: If there’s bad data on the cluster, should it stay on the cluster?

CI: Yes, this should be incorporated into best practices.

SM: Important that the finished project go else.

HL: It’s interesting how many projects are unfinished.

SM: If after 5 years an unfinished project doesn’t go anywhere, we’ll need 200 TB.

RS: What happens when the bioinformatician has their own data? E.g. public repositories. Do we aliquot space for that? Do we set a quota for that?

RS: Do we want to organise the data on the cluster by “internally generated” and “externally generated”?

EB: HPC Core have data sets stored. We could retrieve that data when needed.

SM: There’s high security at UHN preventing connection to other clusters.

SM: Do the wet-lab people have any large data other than fastqs?

CI and HL: No, it’s just fastqs.

SM: Fear of duplication.

EB: Need a financial breakdown of how much sequencing experiments and cluster storage costs.

RS: Data hosting is not a major cost relative to sequencing experiment cost.

SM: It’s important we ensure data is compressed (e.g. bams, .fastq.gz)

SM: This meeting needs follow up – this group should not end.

EB: Meet weekly for this 6 weeks. Meet once a quarter.

## How do we individually manage our data? (ALL)

### Helen

HL: I keep a folder for each donor. Instead of a lab notebook I keep a detailed calendar.

EB: Important point about date-stamping everything.

HL: This system is not good for other people looking at my stuff.

CI: Maybe an electronic lab book would help?

HL: All in the cloud.

### Charles

CI: [Outlines typical start-to-finish from wet- to dry-lab]

CI: If it’s going to be sequenced I assign a unique numerical identifier. This is something the lab was already doing. It’s the way to go for sure.

CI: I complete a submission form for the genomics centre.

CI: I have this personal spreadsheet that details project, stake-holders, description of experiment, genotypes, technique, index used, size selection, cluster location, date submitted for sequencing.

CI: When sequencing returned, I add the flow-cell ID to the spreadsheet. [Stresses the importance of including flow cell IDs.]

CI: Problem: sample IDs not detailed when fastq files returned. This is a breakdown in our pipeline. How do we find out what they are without having to ask Julie?

RS: I have to go in and look at the files.

…

SM: How should we name files? After the ID or after the description? [Expresses strong preference for the latter.]

EB: [Expresses preference for the former.] Easier to keep track of files.

RS: [Also prefers this system.]

…

CI: Bottleneck – I have to open up bioinformaticians.

CI: Get everything onto github before submission and make it private for 2 years.

CI: A lot of what I do already existed in the lab.

### Roxana

RS: I assign aliases to my samples.

RS: I keep track of exactly what happens to my samples in one spreadsheet.

RS: Partly because I am ultimately held accountable to the stake-holder or company.

RS: Most of the protocols I work with, if they’re very repetitive, I set up a checklist.

RS: When I started, we didn’t have a submission form for the sequencing core.

RS: I submit for BioA. If it looks good I submit for sequencing.

RS: All this information is shared with the bioinformatician.

RS: I have a master sequencing list. All my stuff has patient IDs, so I keep it local. It contains the sequencing ID, any unique identifiers, when it was submitted for sequencing, … .

RS: Within my computer I have my collaboration folder.

EB: Are we backing up our stuff?

RS: I do hard back-ups.

### Emma

RS: We don’t keep track of who’s contributed to a project.

CI: Maybe we need a spreadsheet that track project contribution

### Sajid

SM: In the home directory the main folder is the project folder. This contains code only.

SM: If there’s a resource that’s not available on the cluster I’ll generate it and keep it in my data folder.

SM: All files generated from the analysis stays in the same folder – easier to organise. E.g. everything pertaining to replicate 2 is in the replicate 2 folder.

CI: Is this easy to convert into a github organisation?

SM: This would be hard to do that. What should go to the github is code and analysis?

## How do wet-lab researchers manage their experiments? (HL)

HL: There’s insufficient communication between those responsible for data collection and sample sequencing, and bioinformatic analysis.

HL: Current issues: Lack of/poor: communication between wet- and dry-lab; metadata.

HL: [Gives example of orphan box – dump for samples not willing to throw away.]

CI: There’s no date on the tubes?

HL: No…

HL: Most of the time this doesn’t happen. Now we have printers and label makers. The labels survive freeze-thaw.

SM: These are IDed and linked to a speadsheet?

HL: Yes, so if I leave it’s easy to work out what the samples are.

EB: Where do these IDs come from?

HL: Predefined because they’re clinical samples.

RS: Clinical samples are much better looked after.

HL: [Gives examples of RS’s beautiful data and metadata plan.]

## Next week’s meeting (ALL)

Aim: To review RDM guidance from other research groups, institutions, and policy-makers.

|  |  |
| --- | --- |
| 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? |
| Week 3 | How do other research groups and institutions manage their data? |
| Week 4 | Draft guidelines for RDM within the De Carvalho Lab |
| Week 5 | Redraft guidelines |
| Week 6 | Present to group |

## 