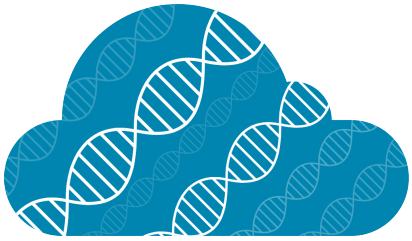
Cloud-SPAN Handbook

Evelyn Greeves

Table of contents

1 Introduction

Welcome!



Cloud-SPAN

Cloud-based High Performance Computing
SPecialised **AN**alyses for 'omics

Welcome to the Cloud-SPAN handbook! It's great to have you here!

Introduction

About Cloud-SPAN

Cloud-SPAN deploys high quality learning resources that will train researchers to effectively generate and analyse a range of 'omics data using Cloud computing resources. Read more about our courses.

Cloud-SPAN is a collaboration between the Department of Biology at the University of York and the Software Sustainability Institute, and funded by the UKRI innovation scholars award under project reference MR/V038680/1.

About this handbook

This handbook is intended as a reference for both the core Cloud-SPAN team (see below) and for our wider community of learners. It's where you'll find our Code of Conduct, contributing guidelines and other practical information which will help you make the most of our resources in a friendly, understanding environment.

Project team

Name	Role	Institution
Emma Rand	Project oversight	Uni of York
Jorge Buenabad-Chavez	Cloud deliverer	Uni of York
Evelyn Greeves	FAIR Training Lead	Uni of York
Pasky Miranda	Research Training Lead	Uni of York
Sarah Dowsland	Project Manager	Uni of York

2 Code of Conduct

Preamble

The Cloud-SPAN team are dedicated to providing a welcoming and supportive environment for all people, regardless of background or identity. As such, we do not tolerate behaviour that is disrespectful to our community members or that excludes, intimidates, or causes discomfort to others. We do not tolerate discrimination or harassment based on characteristics that include, but are not limited to: gender identity and expression, sexual orientation, disability, physical appearance, body size, citizenship, nationality, ethnic or social origin, pregnancy, familial status, veteran status, genetic information, religion or belief (or lack thereof), membership of a national minority, property, age, education, socio-economic status, technical choices, and experience level.

Everyone who participates in Cloud-SPAN project activities is required to conform to this Code of Conduct. This Code of Conduct applies to all spaces managed by the Cloud-SPAN project including, but not limited to, in person focus groups and workshops, and communications online via GitHub. By participating, contributors indicate their acceptance of the procedures by which the project core development team resolves any Code of Conduct incidents, which may include storage and processing of their personal information.

Our Code of Conduct

We are confident that our community members will together build a supportive and collaborative atmosphere at our events and during online communications. The following bullet points set out explicitly what we hope you will consider to be appropriate community guidelines:

- **Be respectful of different viewpoints and experiences.** Do not engage in homophobic, racist, transphobic, ageist, ableist, sexist, or otherwise exclusionary behaviour.
- **Use welcoming and inclusive language.** Exclusionary comments or jokes, threats or violent language are not acceptable. Do not address others in an angry, intimidating, or demeaning manner. Be considerate of the ways the words you choose may impact others. Be patient and respectful of the fact that English is a second (or third or fourth!) language for some participants.

- **Do not harass people.** Harassment includes unwanted physical contact, sexual attention, or repeated social contact (see below for an extended list of behaviours we consider to be harassment). Know that consent is explicit, conscious and continuous—not implied. If you are unsure whether your behaviour towards another person is welcome, ask them. If someone tells you to stop, do so.
- **Respect the privacy and safety of others.** Do not take photographs of others without their permission. Do not share other participant's personal experiences without their express permission. Note that posting (or threatening to post) personally identifying information of others without their consent ("doxing") is a form of harassment.
- **Be considerate of others' participation.** Everyone should have an opportunity to be heard. In update sessions, please keep comments succinct so as to allow maximum engagement by all participants. Do not interrupt others on the basis of disagreement; hold such comments until they have finished speaking.
- **Don't be a bystander.** If you see something inappropriate happening, speak up. If you don't feel comfortable intervening but feel someone should, please feel free to ask a member of the Code of Conduct response team for support.

Unacceptable Behaviour

Examples of unacceptable behaviour by community members at any project event or platform include:

- written or verbal comments which have the effect of excluding people on the basis of membership of any specific group
- causing someone to fear for their safety, such as through stalking, following, or intimidation
- violent threats or language directed against another person
- the display of sexual or violent images
- unwelcome sexual attention
- nonconsensual or unwelcome physical contact
- sustained disruption of talks, events or communications
- insults or put downs
- sexist, racist, homophobic, transphobic, ableist, or exclusionary jokes
- excessive swearing
- incitement to violence, suicide, or self-harm
- continuing to initiate interaction (including photography or recording) with someone after being asked to stop
- publication of private communication without consent

CloudSPAN prioritises marginalised people's safety over privileged people's comfort. We will not act on complaints regarding:

- 'Reverse' -isms, including 'reverse racism,' 'reverse sexism,' and 'cisphobia'.
- Reasonable communication of boundaries, such as "leave me alone," "go away," or "I'm not discussing this with you."
- Communicating in a 'tone' you don't find congenial.
- Criticism of racist, sexist, cissexist, or otherwise oppressive behavior or assumptions.

Incident Reporting Guidelines

Contact points

If you feel able to, please contact Emma Rand by email at emma.rand@york.ac.uk

Alternate contact points

If you do not feel comfortable contacting Emma Rand, please report an incident to Evelyn Greeves by email at evelyn.greeves@york.ac.uk

Acknowledgements

This Code was adapted from the Turing Way Code of Conduct, which itself draws from the Carpentries and Alan Turing Institute Data Study Group codes of conduct. Both are licensed for reuse under a CC BY 4.0 CA license.

Material was additionally drawn from the R Community Diversity, Equity, and Inclusion Working Group, also licensed under CC BY 4.0 CA.

3 Our Courses

We offer courses on topics such as using the command line, genomic analysis and whole metagenome sequencing (WMS) workflows. All of our courses are offered **free of charge**.

Here's why we think our courses are special:

- They're designed to be as accessible as possible to beginners, with lots of guidance and reassurance along the way.
- We use a technique called live coding to keep learners engaged - less lecturing, more doing.
- Most of our courses can either be studied as part of a taught online workshop or as self-study modules. Read more about how to self-study our courses here.
- If you participate in a taught course, you'll get free access to an AWS instance containing all the software and data you'll need for the duration of the course.

> **Note**
>
> The only course which requires significant prior experience in programming is Automated Management of AWS Instances which is aimed at computing professionals and research software engineers.

Prenomics

Prenomics is an interactive online course on understanding file systems and using the command line which takes place over 2 half days (roughly six hours of content). We developed this course after finding that people taking the Genomics course vary in their experiencing of navigating file systems and the command line (shell).

Topics covered include:

- file directory structure
- logging onto a cloud instance
- basic shell commands
- using the shell to manipulate and search files

NO prior experience is necessary.

Genomics

Genomics is a practical, tutor-led course taking place over 4 half days or 2 full days (roughly twelve hours of content). It teaches data management and analytical skills using cloud resources for genomic research.

Topics covered include:

- project management for cloud genomics
- writing shell scripts
- assessing read quality
- trimming/filtering reads
- finding sequence variants.

A basic knowledge of the shell is required, which can be obtained by attending/self-studying the Prenomics course.

Metagenomics

Metagenomics teaches data analysis for metagenomics projects. It runs asynchronously with a mixture of taught sessions, offline self-study and drop-in sessions.

Topics covered include:

- an introduction to metagenomics and what it can be used for
- performing quality control on long and short reads
- improving your assembly with polishing
- binning into species/metagenome-assembled genomes (MAGs)
- taxonomic assignment and functional annotation

No prior knowledge of the shell is required as this course includes two sessions of Prenomics material.

Create Your Own AWS Instance

This course shows you how to create and manage an Amazon Web Service instance like the ones used in all our courses, in order to:

- study course materials further
- perform more complex analyses on your own data.

Core R

This short 2 hour course provides a whistle-stop tour of the R programming language and RStudio environment for total beginners. You will learn how to navigate RStudio, use the manual, import data and summarise/visualise it with a plot.

Automated Management of AWS Instances

Automated Management of AWS Instances is aimed at experienced command line users who have an interest in deploying and managing cloud resources for training purposes.

No prior knowledge of AWS concepts or tools is needed but learners should have significant experience with the Unix terminal and programming with Bash.

4 Self-studying

5 The Cloud-SPAN commmunity

Under construction

Our aim is to build a friendly and involved community of people who have used our resources, are interested in our resources, or who have expertise in the areas we cover.

That means that whether you are…

an expert in 'omics analyses
a complete newbie at 'omics
not entirely sure what an "omic" is
an experienced Cloud user
a little bewildered by the Cloud

…then the Cloud-SPAN community is for you!

There are lots of ways to contribute and we welcome all of them! Contributing, and joining our community, doesn't have to mean writing technical code.

Here are some ideas of ways you can contribute:

Ways to contribute

Learn

- Attend or work through one of our courses.

Connect

- Join our *Community of Practice*.
- Come along to one of our code retreats to get help applying your skills and meet other researchers working in similar fields.

Help

- Tell us about bugs or problems you encounter in the course.

Expand

- Suggest new/different software tools for analysis.
- Contribute new examples.
- Attend one of our 'Train the Trainer" courses so you can take parts of the course back to teach at your home institution.

Using GitHub to contribute

We use GitHub as a tool for managing version control (AKA keeping a record of the project's development). This helps us stay accountable and transparent. It's also one of the ways we are making steps towards adhering to the [FAIR Principles](#fair-principles).

If you want to contribute any content such as an update to the course or a new example then via GitHub is the best place to get in contact. For lots more guidance about how to contribute via GitHub, read our GitHub Contribution Guide.

Git (the programming language underlying Github) and Github can be a little intimidating at first but don't worry, the team are here to hold your hand!

6 FAIR Principles

What is FAIR data?

FAIR data is **Findable**, **Accessible**, **Interoperable** and **Reusable**.

These principles are designed to help both humans and machines find and reuse data as easily as possible. They are aspirational but tangible steps can be made towards realising them.

You can read about the ethical values underlying the FAIR principles via the FAIR Cookbook here.

Findable

Findable is all about making sure data/resources are as easy to find as possible.

How we're making resources **findable** at Cloud-SPAN:

- We have added rich metadata to our teaching resources using the Bioschemas protocol for training materials.
- We have assigned DOIs to our training materials by depositing them in Zenodo.
- We will be registering our training materials with TeSS, a repository for life sciences training resources.
- We will be assigning persistent identifiers to our teaching materials to prevent "link rot", or broken links.

Accessible

Accessible means it is easy to find out how to access the data/resources.

How we're making resources **accessible** at Cloud-SPAN:

- Our training materials will be openly available, with no caveats, for use by those who cannot attend our workshops or who prefer self-led study.
- We state this in the metadata of resources and on the webpages hosting the courses.

Interoperable

Interoperable means data/resources can be easily integrated with other data/resources, and be viewable in different programs, applications or workflows.

How we're making resources **interoperable** at Cloud-SPAN:

- We provide data for analysis in de facto standard file formats, such as the FASTQ format for sequencing data.
- We write our training resources in Markdown, a widely used and platform-independent text formatting language which renders in all browsers.
- We use Bioschemas markup to add metadata to our resources, which is part of an initiative to standardise how search engines read webpages containing data.
- Within our metadata we use the EDAM topic ontology to describe the topics our courses cover.

Reusable

Reusable is about making sure that data/resources are suitable for re-use in different settings by including "richly described metadata" and applying a suitable licence.

How we're making resources **reusable** at Cloud-SPAN:

- We have tagged our resources with metadata properties which conform to the Bioschemas suggested list of properties for biosciences training materials.
- We have applied Creative Commons Attribution 4.0 International (CC-BY) licences to our training materials - this is stated in the metadata, in the GitHub repository and on the webpages hosting the courses.
- We welcome (and encourage!) outside contributions of explanations and examples - see the Ways to contribute for more information.