Biosciences computing resources and experience

High-performance computing, data storage, software development and research data lifecycle management is provided by ARC (research platforms listed below). Computational support for research and teaching in the Biosciences is provided by Dr James Gilbert, Senior Research Data Scientist, since autumn 2024.

COMPUTING

ARC develops, delivers and operates a wide range of research platforms.

Research Computing Platforms

We offer a range of advanced platforms for computationally intensive research which are free for UCL researchers to use on a fair share basis, enabling our users to perform all but the most demanding analyses and simulations.

All of our computing platforms use the same Linux-based software stack, which provides a consistent user interface and makes it easy for researchers to switch platforms according to their need.

Myriad

Myriad is designed for high I/O, high throughput jobs that will run within a single node rather than multi-node parallel jobs.

Kathleen

Kathleen is a compute cluster designed for extensively parallel, multi-node batch-processing jobs, having high-bandwidth connections between each individual node.

A cluster designed specifically for a high-speed Infiniband network linking together its compute nodes, allowing them to coordinate effort on a single task very effectively.

Data Safe Haven

UCL's Data Safe Haven (DSH) should be used for storing or processing sensitive data, including identifiable personal data and data that needs to be treated as special category for legal or contractual reasons. Myriad and Kathleen aren't suitable for use with this type of data.

Condenser

A private cloud platform, including virtualisation and containerisation, that underpins research activities at UCL.

RCNIC

The Research Computing and Networking Innovation Centre (RCNIC) aims to help with the adoption of new technologies across UCL's research computing facilities. It runs projects to evaluate, prove and incubate new technologies, which may become future IT services for researchers.

Materials & Molecular Modelling Hub

ARC hosts Young, the compute cluster for the UK National Tier-2 High Performance Computing Hub in Materials and Molecular Modelling.

MMM Hub

mmmhub.ac.uk

Teams in UGI primarily use the Myriad and CS clusters, with generally better experience reported for CS cluster usage. The advantages of the CS cluster are the fact that computing tasks can be run for long amounts of time (whereas they are limited to 48 hours in Myriad), additional storages space can be easily purchased and the support provided for by the cluster team is generally superior to that of Myriad. Both clusters have massive demand and there is a long waiting time for running jobs. Some users have reported both resources are inferior to the ones at the Crick, which is probably not surprising.

Detailed user experience (conveyed directly by members of the Balloux and Secrier labs):

Myriad cluster:

a. Long job queue times. This is a recurrent problem faced by many users.
b. Storage is limited. We are given 1Tb in scratch and I requested for more storage some time back but was told this is not possible. Unfortunately, this limits me to small scale jobs and precludes the use of automated pipelines.
c. Maximum job run time is 48hrs. This restricts us to short and small scale jobs.
d. There is no option to buy more computing resources last we checked (mid 2024).

CS cluster:

a. Long job queue times. This is a recurrent problem faced by many users.
b. Storage is functionally infinite (some teams have a 40Tb project space).
c. Job durations are functionally infinite (you can request for a job that runs for 480 days).
d. The login node is often quite laggy, potentially because the cluster system is quite outdated and inundated with students.
e. Disk I/O speed is the major bottleneck for the CS cluster. Large, parallelised bioinformatics pipelines that read and write a lot of data concurrently cannot be run without causing massive slowdowns to the cluster. In comparison, such jobs can be run concurrently on the Crick cluster without any slowdown.
f. Provides the option of buying a computing node managed by cluster team and used exclusively by the team who purchases the node. This is an amazing option and has largely solved the problem with job queue times. However, there is no standard protocol in place for this and it always seems like the cluster team is understaffed, so I suspect if many groups start asking the cluster team for exclusive nodes, this may become a problem.

DATA STORAGE

The data storage options at UCL are listed below. My team's experience is that data storage services are adequate, relatively cheap and sufficient for our needs (but others with more data may disagree).

Research Data Platforms

We offer a growing suite of research data platforms, including storage for current projects and a long-term repository, to support effective data management from planning to re-use.

Research Data Storage Service

This service is available for UCL research staff and their collaborators. It provides controlled access to research data storage, potentially at a very large scale. It is a managed service based on dedicated, resilient hardware which is backed up to tape on a daily basis. Contextual data about the project is captured to facilitate future curation and data management. The RDSS is available for use both by externally funded projects and internal 'unfunded' research.

UCL Research Data Repository

The Repository enables the long-term preservation of selected data beyond the lifetime of an active research project. It also acts as a data publication platform, enabling the citation and future re-use of datasets created by UCL staff.

Electronic Research Notebook

The UCL ERN service provides a means to edit and manage notes and data relating to your research. It is intended not only as an alternative to traditional paper lab notebooks, but as a wide-ranging solution for researchers who wish to gather their notes and related files in a single system where they can collaborate and selectively share their work with others in their team.

Medical Imaging Platforms

XNAT

This is an open source imaging informatics platform developed by the Neuroinformatics Research Group at Washington University. It helps UCL researchers store and share medical image data and associated files in compliance with GDPR and Information Governance requirements, allowing users to upload data either directly from hospital scanners and PACS systems, via ZIP files through the web interface or from a personal computer using a desktop client. UCL's XNAT servers are a central resource available to researchers at UCL, collaborating partners such as UCLH and the wider medical imaging research community.

Key gaps:

1. The high demand for HPC means that users need to frequently wait for several days until their jobs will run.
2. Myriad limits job runtime to 48 hours, so longer jobs can only be run on the CS cluster.
3. Resources are generally limited and more in demand compared to other research institutes.
4. Storing controlled-access datasets (such as human genetics data) in a way that complies with recent security requirements from NIH is not entirely straightforward.

Selected key contacts in Biosciences (and beyond):
James Heatherington, Director of ARC
Denise Gordon, Head of Faculty IT, Faculty of Life Sciences
Richard Poole, Associate Professor of Developmental Biology, Head of the Biosciences IT committee
Richard Pearson, Professor of Ecology and Associate Director of Research for Biosciences (overseeing computing-related aspects, among others)
James Gilbert, Senior Research Data Scientist (providing computational support to Biosciences teams)

Further information:
https://www.ucl.ac.uk/biosciences/biosciences-computational-biology
https://www.ucl.ac.uk/advanced-research-computing/