The Rosalind Franklin Institute

Accelerating Structural Biology: From PC to HPC

Dr Dimitrios Bellos

Rosalind Franklin Institute, Baskerville HPC, Collaborative Computational Project for Electron cryo-Microscopy (CCP-EM)

Rosalind Franklin Institute
Artificial Intelligence and Informatics:
Dr Laura Shemilt
Gabryel Mason-Williams
Dr Joss Whittle

Rosalind Franklin Institute
Structural Biology:
William Bowles

Baskerville HPC:
Dr Gavin Yearwood
Dr James Allsopp
Dr Jenny Wong
Dr Simon Hartley

CCP-EM:
Dr Colin Palmer
Dr Tom Burnley

The Rosalind Franklin Institute

Software used for Structural Biology

Structural Biology Software :

- It is tested to run on a single machine - PC.
  Computational bottleneck
- Have Graphical User Interface (GUI) that it is preferred
- It is being used to process multi terabytes of data
  Even though, not large local storage or slow connection with external
- A single user can use the PC at a time

Can we bring and use Structural Biology Software on HPCs ?

High Performance Compute cluster Software:

- Ideally should run on multiple machines
  High compute resources
- They are operated via terminal commands
- High speed connection with large storage
- Multiple users can schedule jobs and use it simultaneously

REgularised LIkelihood OptimisatioN (RELION)

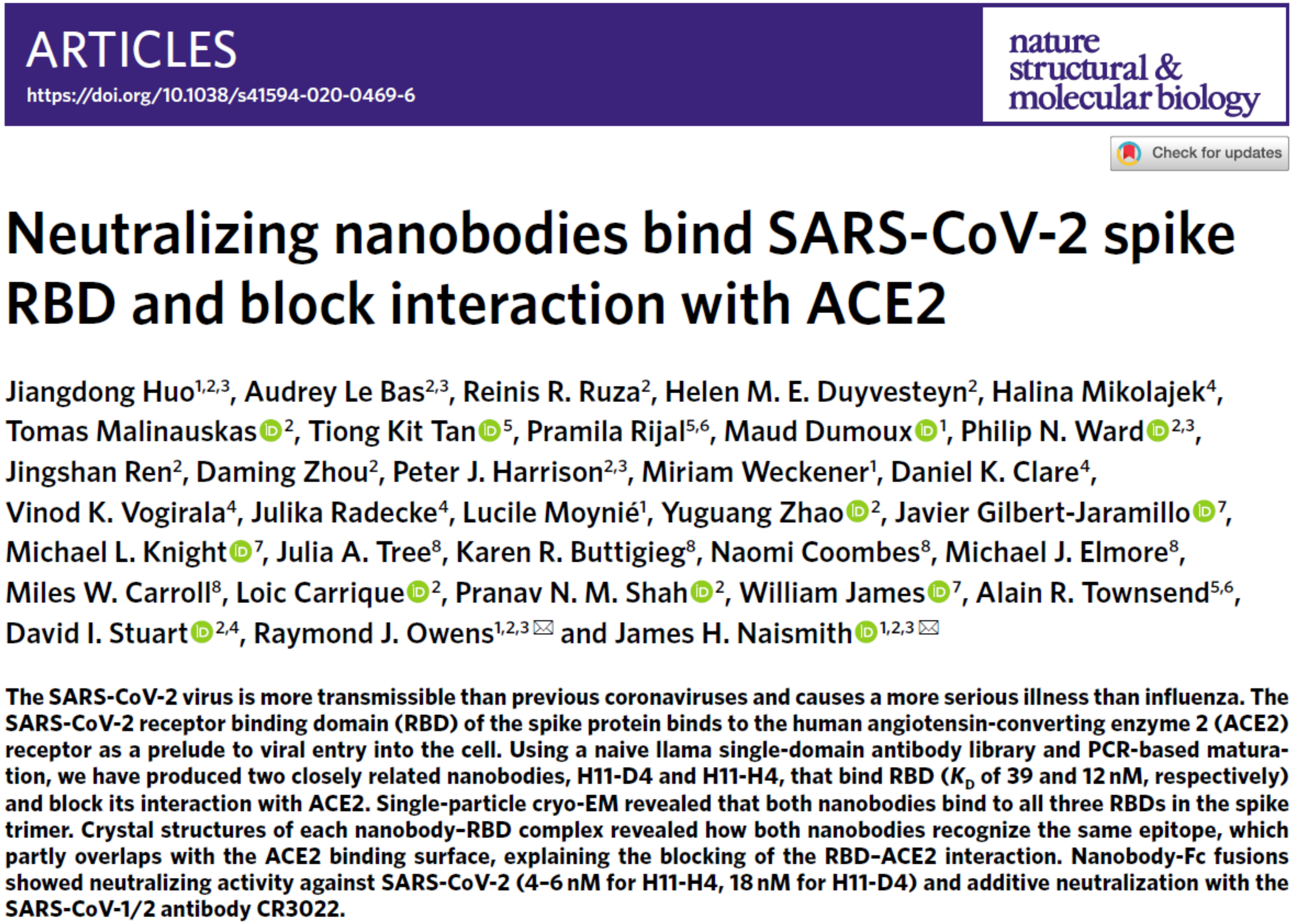
One of the most commonly used software for Structural Biology is RELION.
It employs empirical Bayesian approaches for electron cryo-microscopy (cryo-EM) structure determination.

- Bringing RELION on an HPC can accelerate science tremendously and increase the number of publications.
- Operations that take 2 weeks in a single machine may now be done in a few days.
- The advantage of RELION is that it has the capacity to be run HPC, especially a GPU cluster.

The Rosalind Franklin Institute

REgularised LIkelihood OptimisatioN (RELION)

An example of highly impactfully publication thanks to RELION



ARTICLES

https://doi.org/10.1038/s41594-020-0469-6

nature structural & molecular biology

Check for updates

Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2

Jiangdong Huo[1,2,3], Audrey Le Bas[2,3], Reinis R. Ruza[2], Helen M. E. Duyvesteyn[2], Halina Mikolajek[4], Tomas Malinauskas[2], Tiong Kit Tan[5], Pramila Rijal[5,6], Maud Dumoux[1], Philip N. Ward[2,3], Jingshan Ren[2], Daming Zhou[2], Peter J. Harrison[2,3], Miriam Weckener[1], Daniel K. Clare[4], Vinod K. Vogirala[4], Julika Radecke[4], Lucile Moynié[1], Yuguang Zhao[2], Javier Gilbert-Jaramillo[7], Michael L. Knight[7], Julia A. Tree[8], Karen R. Buttigieg[8], Naomi Coombes[8], Michael J. Elmore[8], Miles W. Carroll[8], Loic Carrique[2], Pranav N. M. Shah[2], William James[7], Alain R. Townsend[5,6], David I. Stuart[2,4], Raymond J. Owens[1,2,3] ✉ and James H. Naismith[1,2,3] ✉

The SARS-CoV-2 virus is more transmissible than previous coronaviruses and causes a more serious illness than influenza. The SARS-CoV-2 receptor binding domain (RBD) of the spike protein binds to the human angiotensin-converting enzyme 2 (ACE2) receptor as a prelude to viral entry into the cell. Using a naive llama single-domain antibody library and PCR-based maturation, we have produced two closely related nanobodies, H11-D4 and H11-H4, that bind RBD (K_D of 39 and 12 nM, respectively) and block its interaction with ACE2. Single-particle cryo-EM revealed that both nanobodies bind to all three RBDs in the spike trimer. Crystal structures of each nanobody–RBD complex revealed how both nanobodies recognize the same epitope, which partly overlaps with the ACE2 binding surface, explaining the blocking of the RBD–ACE2 interaction. Nanobody–Fc fusions showed neutralizing activity against SARS-CoV-2 (4–6 nM for H11-H4, 18 nM for H11-D4) and additive neutralization with the SARS-CoV-1/2 antibody CR3022.

REgularised LIkelihood OptimisatioN (RELION)

However, to bring RELION in an HPC (Baskerville) there were some requirements to make it highly accessible to its users

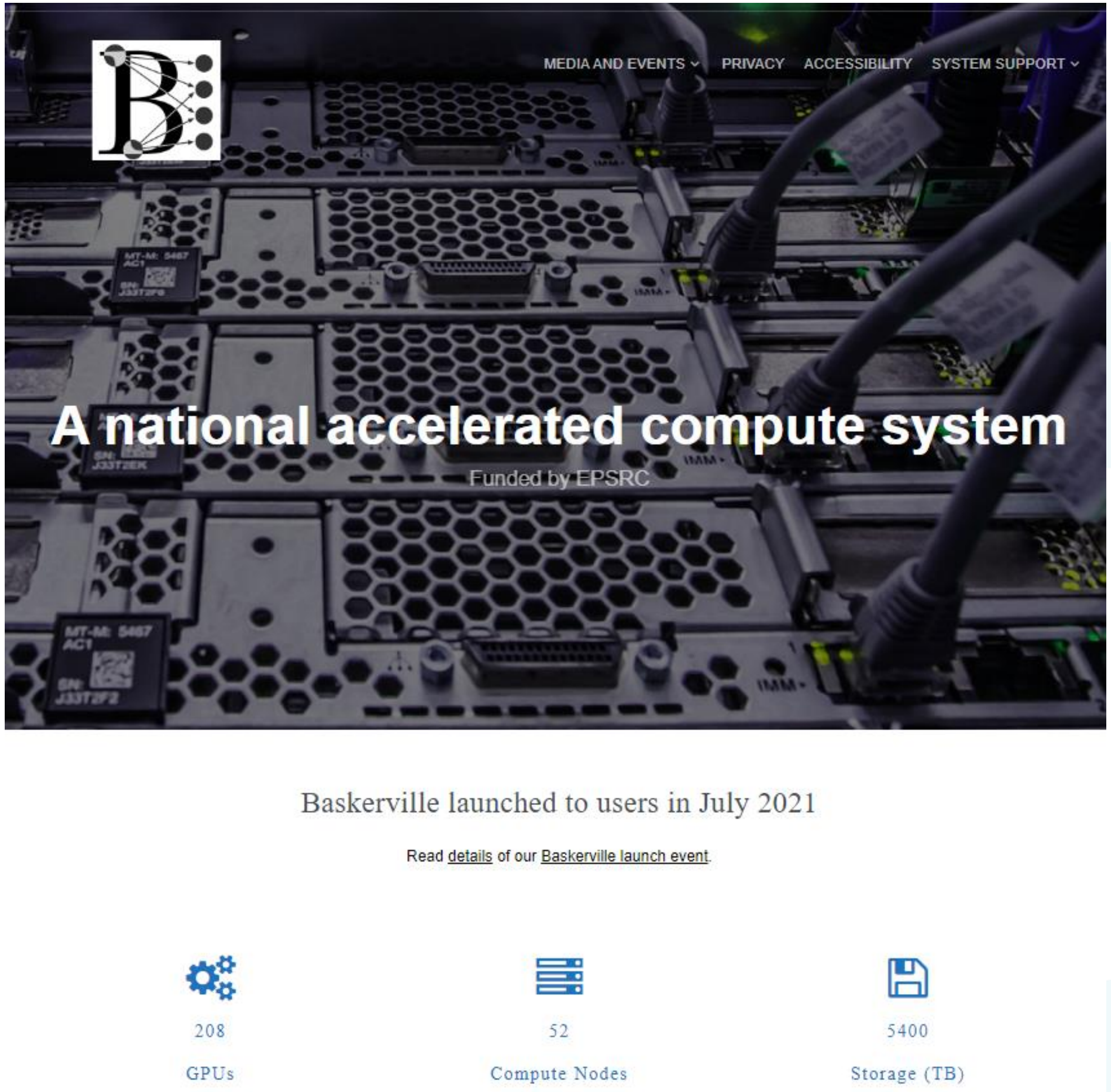
1. Offer a way to use RELION GUI approach

2. Allow fast data transfers between the cluster and the data storage

3. Resolve potential technical issues, perform testing and profiling when this requires multiple groups to co-ordinated (Franklin biologist and AI core Team, Baskerville HPC, CCP-EM)

4. Educate the users how to use RELION on an HPC instead of PCs and single machines.

Baskerville HPC cluster



A national accelerated compute system

Funded by EPSRC

Baskerville launched to users in July 2021

Read details of our Baskerville launch event.

208

GPUs

52

Compute Nodes

5400

Storage (TB)

Baskerville HPC offers provides a high compute resources and in particular there are 52 nodes and a total sum of 208 high performance Nvidia A100s GPUs.

Offers a large data storage with fast connection

Optimal for running software that can run in a multi-node multi-GPU setting

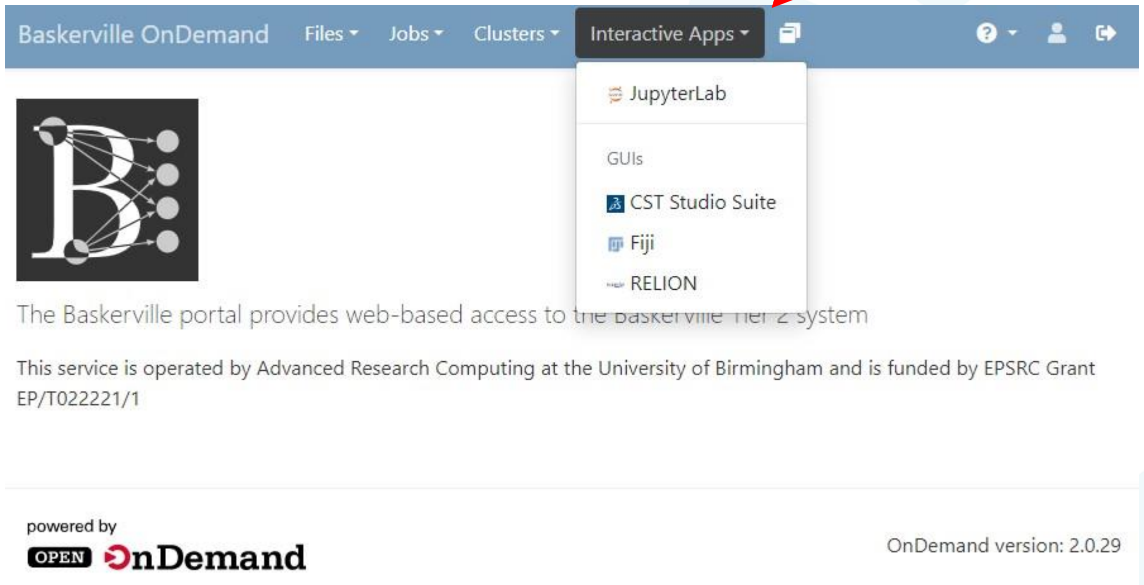
The Rosalind Franklin Institute

Interactive RELION App on Baskerville HPC

As it is widely known, learning to use a software via a GUI is is easier than with terminal commands

An interactive app has been created on Baskerville Portal that allows the users to launch RELION on Baskerville and to display its GUI

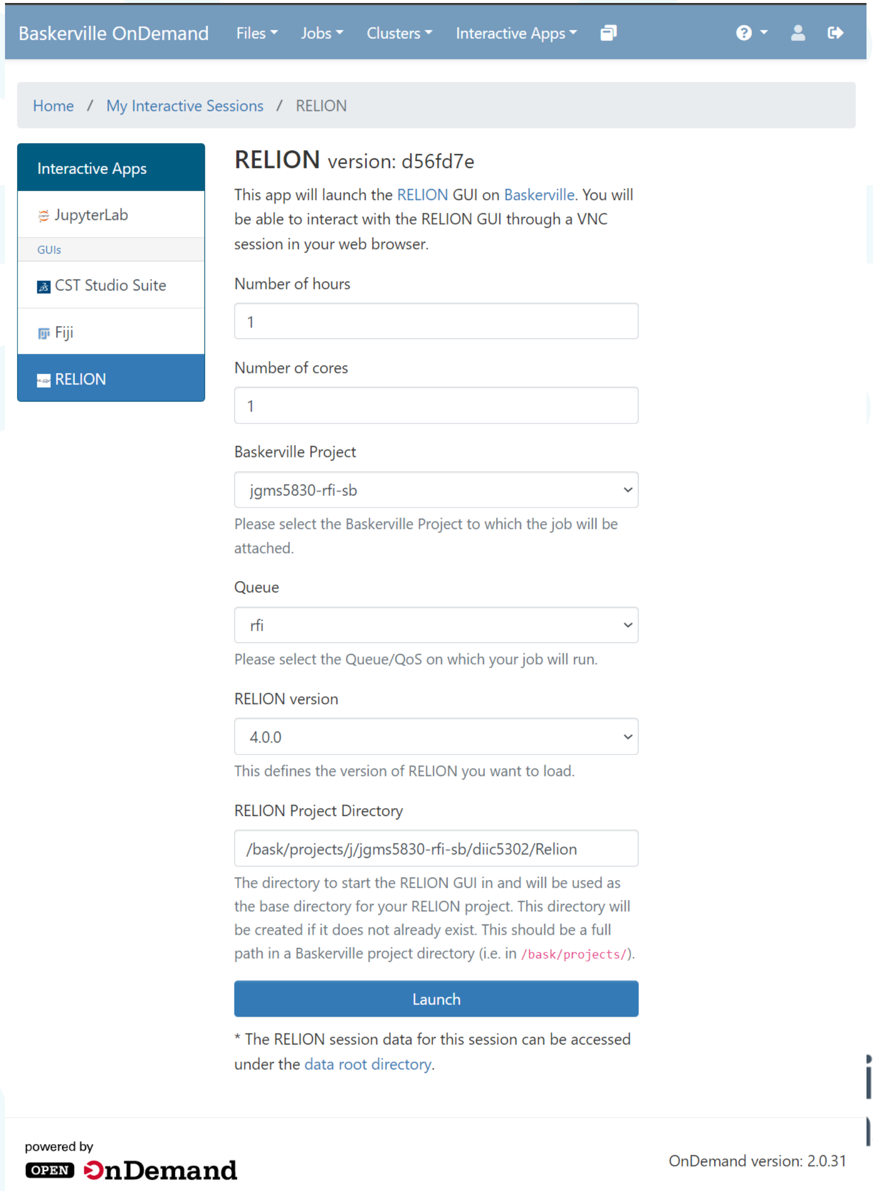
To login to Baskerville Portal 2FA is being used



The Baskerville portal provides web-based access to the Baskerville Tier 2 system

This service is operated by Advanced Research Computing at the University of Birmingham and is funded by EPSRC Grant EP/T022221/1

OnDemand version: 2.0.29



RELION version: d56fd7e

This app will launch the RELION GUI on Baskerville. You will be able to interact with the RELION GUI through a VNC session in your web browser.

Number of hours: 1

Number of cores: 1

Baskerville Project: jgms5830-rfi-sb

Please select the Baskerville Project to which the job will be attached.

Queue: rfi

Please select the Queue/QoS on which your job will run.

RELION version: 4.0.0

This defines the version of RELION you want to load.

RELION Project Directory: /bask/projects/j/jgms5830-rfi-sb/diic5302/Relion

The directory to start the RELION GUI in and will be used as the base directory for your RELION project. This directory will be created if it does not already exist. This should be a full path in a Baskerville project directory (i.e. in `/bask/projects/`).

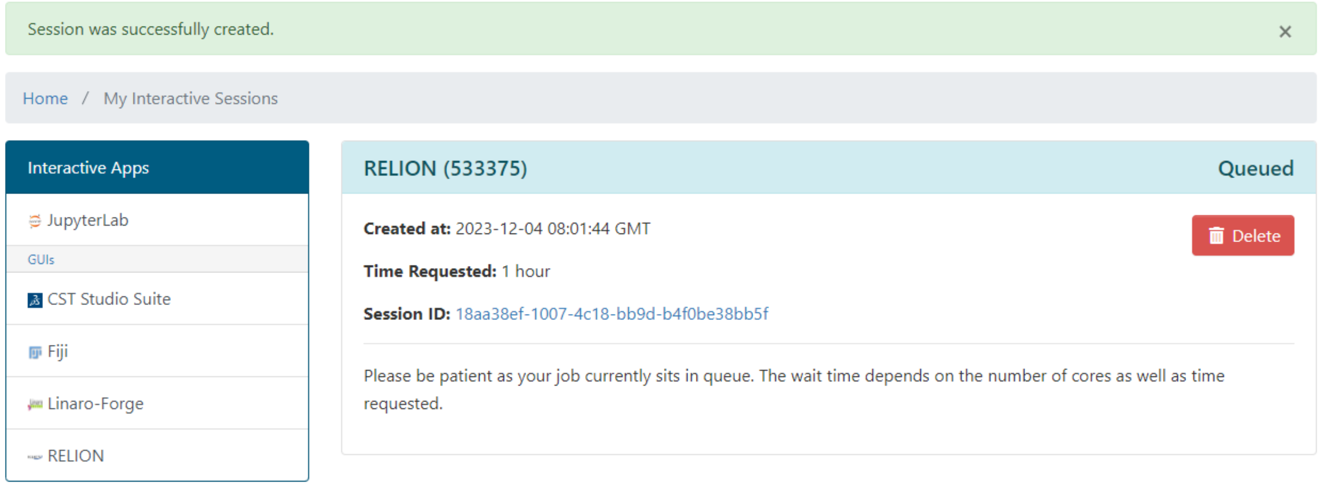
* The RELION session data for this session can be accessed under the data root directory.

OnDemand version: 2.0.31

Interactive RELION App on Baskerville HPC

When a new interactive app is being requested a single GPU on a single node has to be reserved

This is needed for the graphics part of the Graphical User Interface



Session was successfully created.

RELION (533375) Queued

Created at: 2023-12-04 08:01:44 GMT

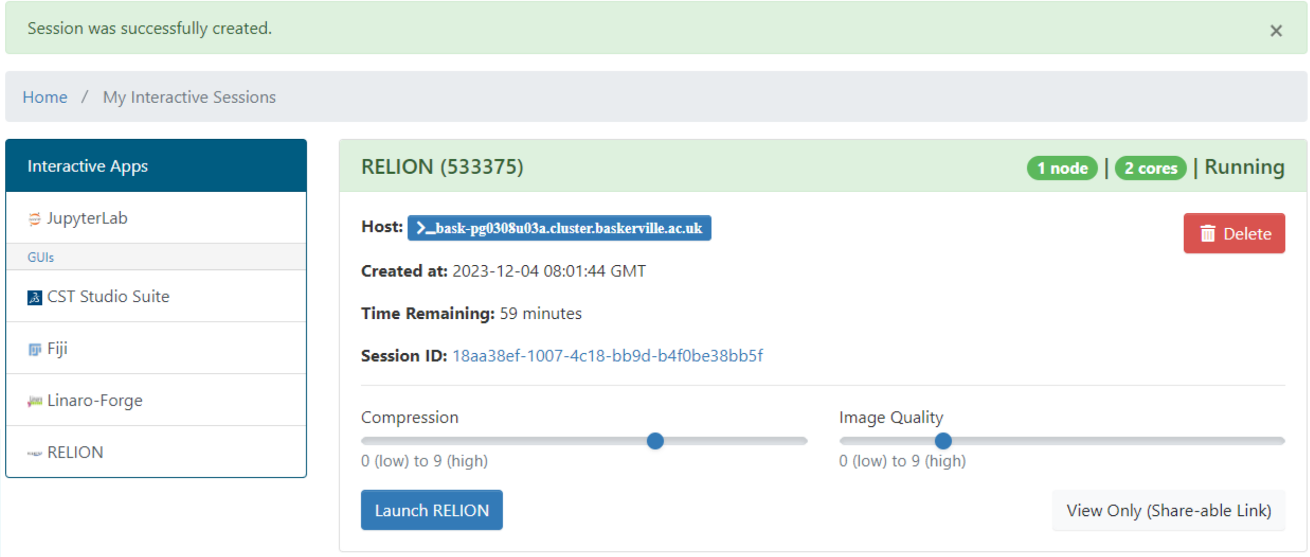
Time Requested: 1 hour

Session ID: 18aa38ef-1007-4c18-bb9d-b4f0be38bb5f

Please be patient as your job currently sits in queue. The wait time depends on the number of cores as well as time requested.

After the user clicks the button ‘Launch RELION’ a new browser tab opens with RELION GUI.

When the GUI is no longer need the user is advised to delete the session, thus releasing the allocated GPU



Session was successfully created.

RELION (533375) 1 node | 2 cores | Running

Host: >_bask-pg0308u03a.cluster.baskerville.ac.uk

Created at: 2023-12-04 08:01:44 GMT

Time Remaining: 59 minutes

Session ID: 18aa38ef-1007-4c18-bb9d-b4f0be38bb5f

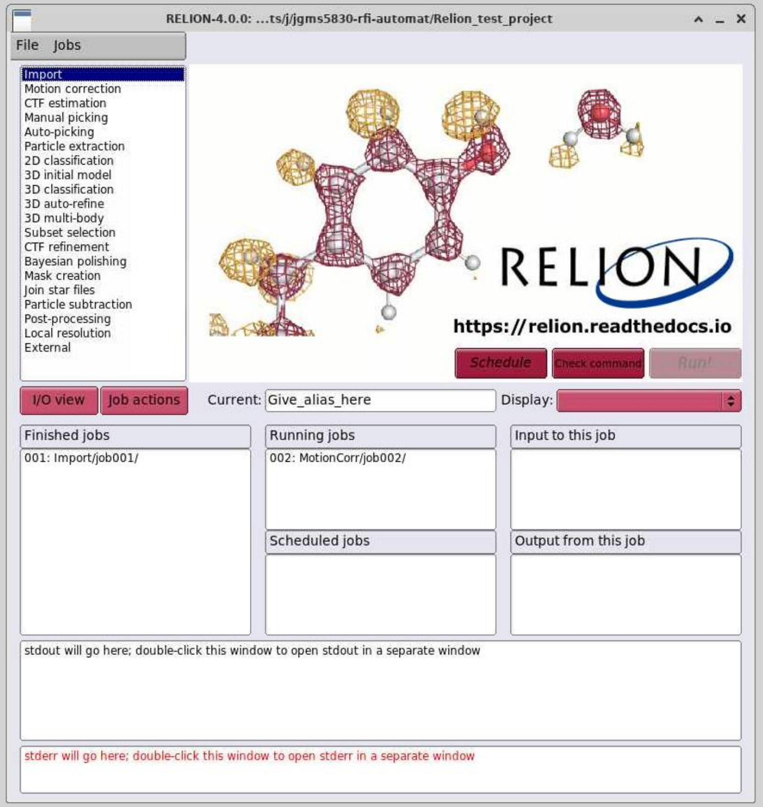
Compression
0 (low) to 9 (high)

Image Quality
0 (low) to 9 (high)

Interactive RELION App on Baskerville HPCs

For almost every process on the list there is a GPU acceleration option. In the final 'Running' tab the user can specify slurm scheduler options.

After clicking 'Run!' a new independent slurm job will be submitted that will not be killed if the Interactive session is closed



RELION-4.0.0: ...ts/j/jgms5830-rfi-automat/Relion_test_project

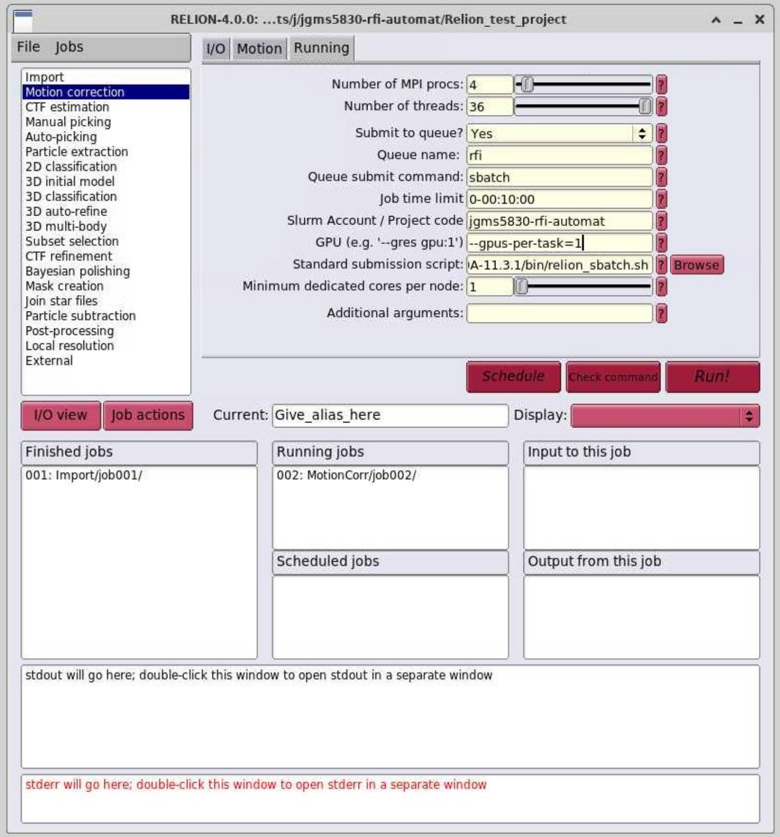
Current: Give_alias_here

Finished jobs: 001: Import/job001/

Running jobs: 002: MotionCorr/job002/

stdout will go here; double-click this window to open stdout in a separate window

stderr will go here; double-click this window to open stderr in a separate window



RELION-4.0.0: ...ts/j/jgms5830-rfi-automat/Relion_test_project

Number of MPI procs: 4

Number of threads: 36

Submit to queue? Yes

Queue name: rfi

Queue submit command: sbatch

Job time limit 0-00:10:00

Slurm Account / Project code jgms5830-rfi-automat

GPU (e.g. '--gres gpu:1') --gpus-per-task=1

Standard submission script: A-11.3.1/bin/relion_sbatch.sh

Minimum dedicated cores per node: 1

Additional arguments:

Current: Give_alias_here

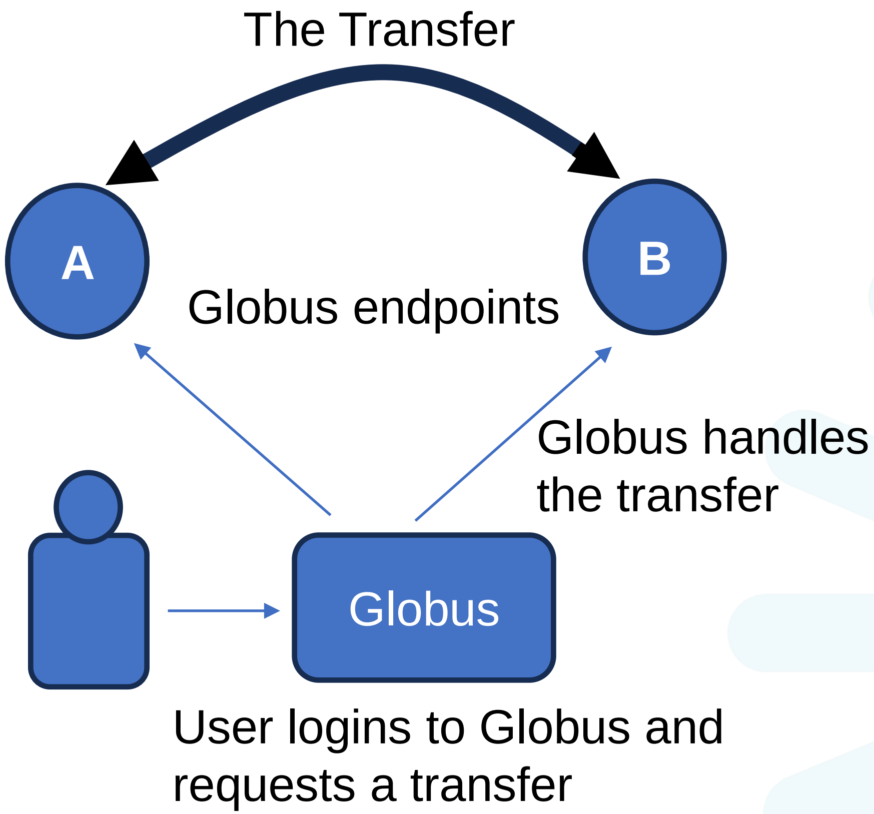
Finished jobs: 001: Import/job001/

Running jobs: 002: MotionCorr/job002/

stdout will go here; double-click this window to open stdout in a separate window

stderr will go here; double-click this window to open stderr in a separate window

Fast data transfers to Baskerville using Globus


The Transfer
A
B
Globus endpoints
Globus handles the transfer
Globus
User logins to Globus and requests a transfer

Globus is a service that allows fast data transfers between machines where a Globus endpoint has been setup

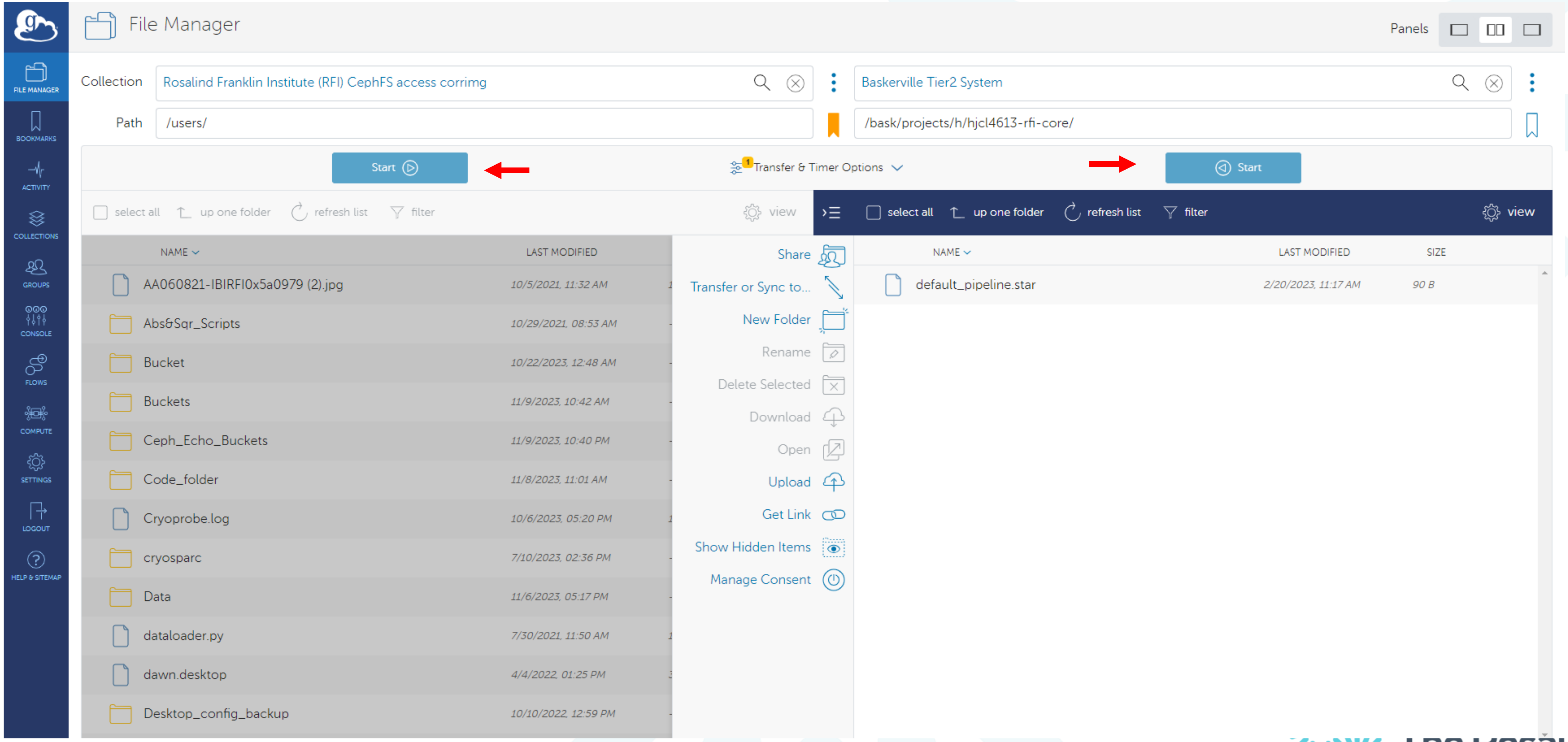
Globus allows encryption and different levels of the visibility for its endpoints and permissions

Transfers initiated via Globus website

Very intuitive website design

The Rosalind Franklin Institute

Fast data transfers to Baskerville using Globus



Collection: Rosalind Franklin Institute (RFI) CephFS access corrimg

Path: /users/

Collection: Baskerville Tier2 System

Path: /bask/projects/h/hjcl4613-rfi-core/

NAME	LAST MODIFIED
AA060821-IBIRFI0x5a0979 (2).jpg	10/5/2021, 11:32 AM
Abs&Sqr_Scripts	10/29/2021, 08:53 AM
Bucket	10/22/2023, 12:48 AM
Buckets	11/9/2023, 10:42 AM
Ceph_Echo_Buckets	11/9/2023, 10:40 PM
Code_folder	11/8/2023, 11:01 AM
Cryoprobe.log	10/6/2023, 05:20 PM
cryosparc	7/10/2023, 02:36 PM
Data	11/6/2023, 05:17 PM
dataloader.py	7/30/2021, 11:50 AM
dawn.desktop	4/4/2022, 01:25 PM
Desktop_config_backup	10/10/2022, 12:59 PM

NAME	LAST MODIFIED	SIZE
default_pipeline.star	2/20/2023, 11:17 AM	90 B

The Rosalind Franklin Institute

Fast data transfers to Baskerville using Globus

To put things in perspective:
From 29/09 - 25/10 were able to transfer 724TB to Baskerville

Logins to Baskerville’s Globus endpoint it requires to authenticate with 2FA

The authentication last only for 30 days

Transfers to and from Baskerville are forced to be encrypted and this option cannot be disabled by the users

Baskerville has a large storage and every Baskerville project had different storage quota set by the project investigator (PI). It is the users’ responsibility to do periodic clean ups.

Engaging all related groups

We organised meetings with different groups separately (CCP-EM, Baskerville, Franklin biologists), but also joined meetings

During these meetings we arranged to reserve Baskerville resources to be able to solve issues with live testing

Using these meeting we were able to resolve:

- Allow submitted RELION to not be killed after the interactive app session is deleted
- Find additional dependencies that were missing
- Find argument limits that we are working on removing
- Provide some initial compute argument recommendations to the users

Training materials and documentation for users

- A week ago, on 27th and 28th of November the annual Baskerville training took place in Harwell campus.

- We have a website to our users where they can read recommended RELION compute settings for a single machine

- We are working on a guide for recommended compute settings on Baskerville. To do so we are working with the users helping them test and profile their Baskerville jobs.

- There is a training and documentation material offered to our users on how to use Baskerville

- Baskerville has a very comprehensive documentation page (https://docs.baskerville.ac.uk/)

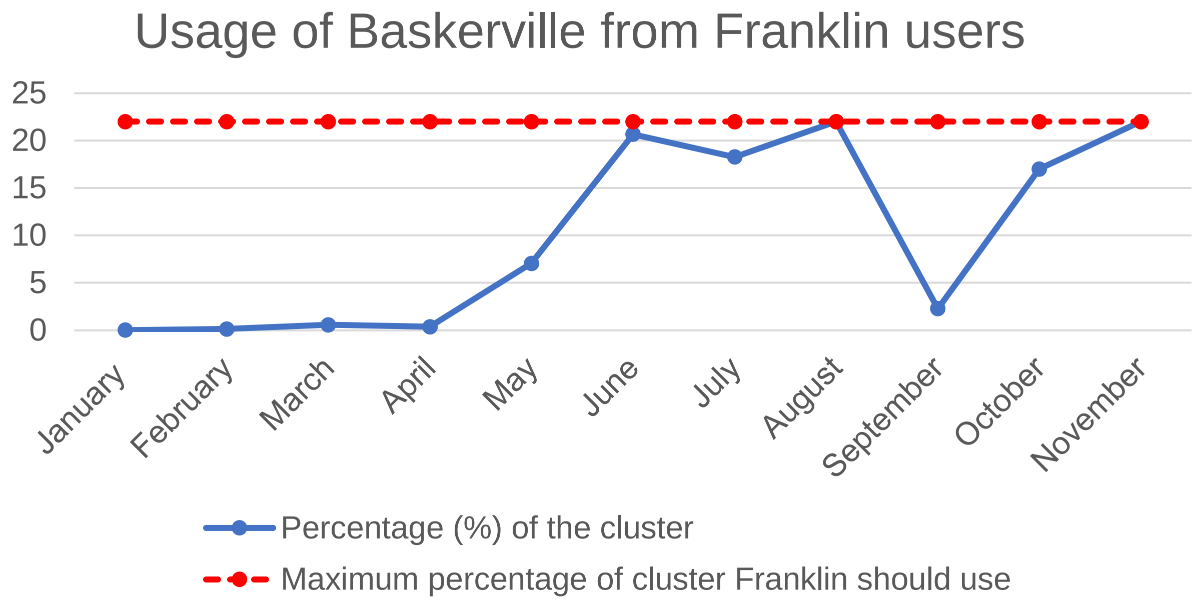
The Rosalind Franklin Institute

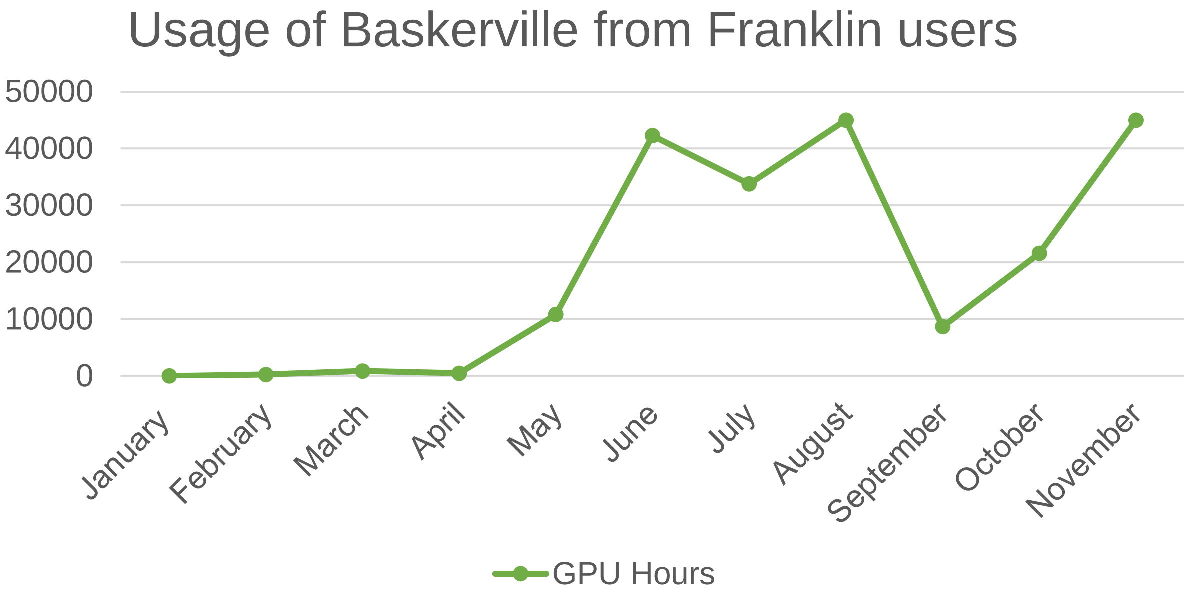
Open communication channels and support

There is a ticketing system in Franklin where users can email issues they experience when using RELION on Baskerville

The is also a baskerville-rse Slack channel which our users can use to directly contact the Baskerville team if there is an issue that needs to be addressed to them

Usage of Baskerville from Franklin Users since the beginning of the year (2023)





Conclusion and Future Plans

Things that helped:

- RELION can launch with a GUI made it more accessible to our users
- Use of Globus has facilitated fast transfers
- Joint meetings along all related groups, with dedicated resources for live testing accelerated the fixing of issues
- Offering training, documentation material and continuous support to our users

Things that do not help

- Difficult to write documentation on recommended compute options (for efficiency)
- Many manual steps required from the users
- Difficulty for users to interpret error outputs

Future Plans:

- Offer an automated system to run jobs on Baskerville (Transfer data & Compute & Return data)
- Create guides on recommended compute values

The Rosalind Franklin Institute

Thank you for your attention

The Rosalind Franklin Institute