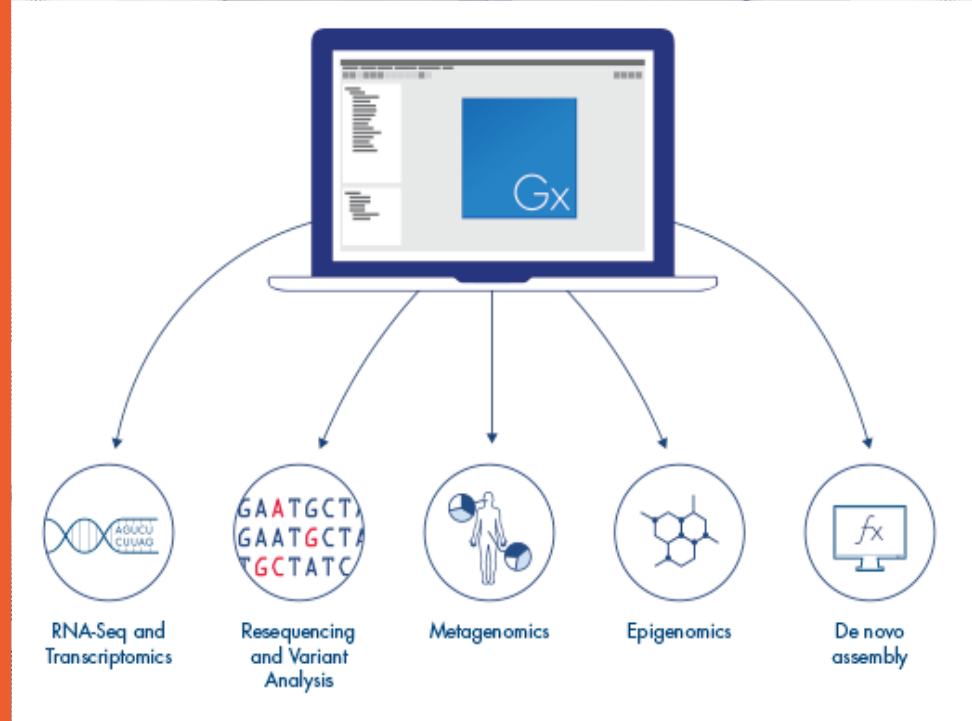


Introduction to the CLC Genomics Workbench on Artemis

Sydney Informatics Hub
Information Communications Technology

Tracy Chew, Rosemarie Sadsad

sih.info@sydney.edu.au



THE UNIVERSITY OF
SYDNEY

Course Overview

1. Preparation

- Introduction to CLC Genomics, Artemis and the Research Data Store (RDS)

2. Importing Data using the import-export directory

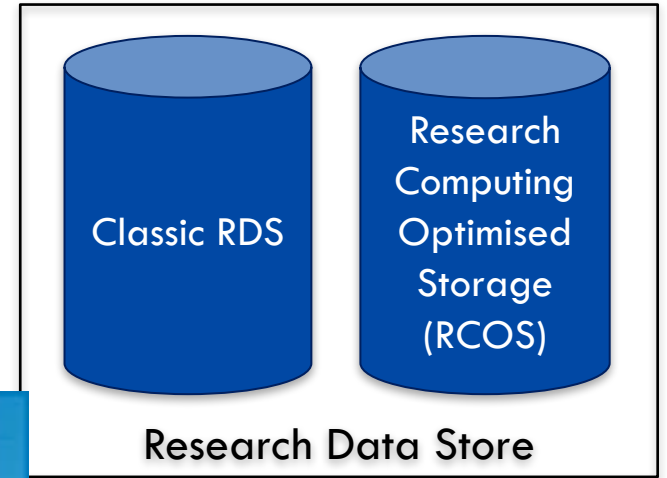
- Mapping network drives
- Moving data that is on the RDS (RCOS) into import-export using FileZilla
- Importing data into CLC Genomics and the DTQ nodes

3. Performing an Alignment in CLC Genomics

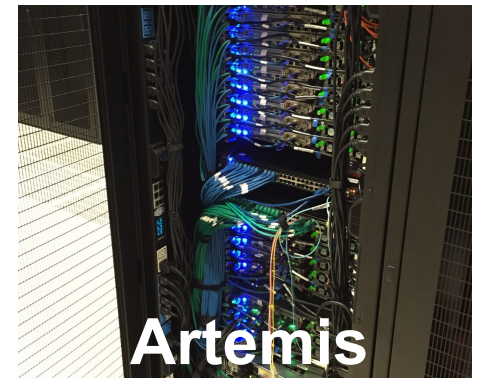
- Artemis compute nodes
- Checking job status on Artemis
- Viewing alignments

4. CLC Genomics subscriptions, DashR, PPMS booking system

5. Free time to explore CLC Genomics for new users



CLC Genomics Workbench



Training unikey

Today we will assign training unikeys for you to use in this course.

The training unikey is:

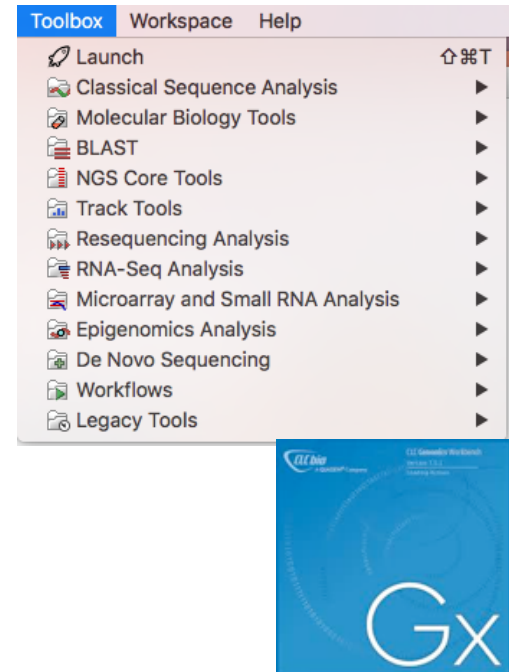
ict_hpctrainN

(N = 2–11, we will assign you a number)

CLC Genomics Workbench

The CLC Genomics Workbench is a user friendly program that provides comprehensive suite of bioinformatics tools.

New users can learn about subscriptions and user processes at the end of the course. You may also have some spare time at the end to explore CLC Genomics.



Today we will learn how to effectively use the Artemis server in CLC Genomics. QIAGEN provide some in depth tutorials on more specific analyses:

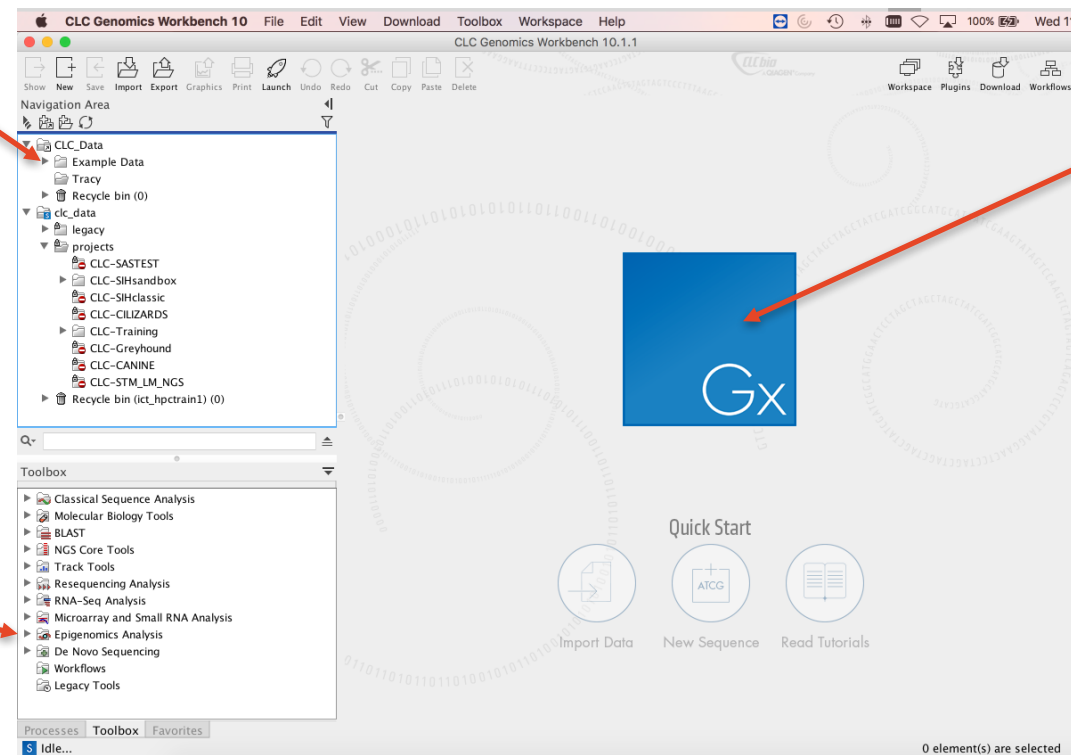
<https://www.qiagenbioinformatics.com/support/tutorials/>

CLC Genomics Workbench

Open the CLC Genomics Workbench (please see pg 4-6 of the [user guide](#) if you haven't already installed CLC and the Server plugin)

Navigation
area

Access to
the Toolbox
or status of
processes

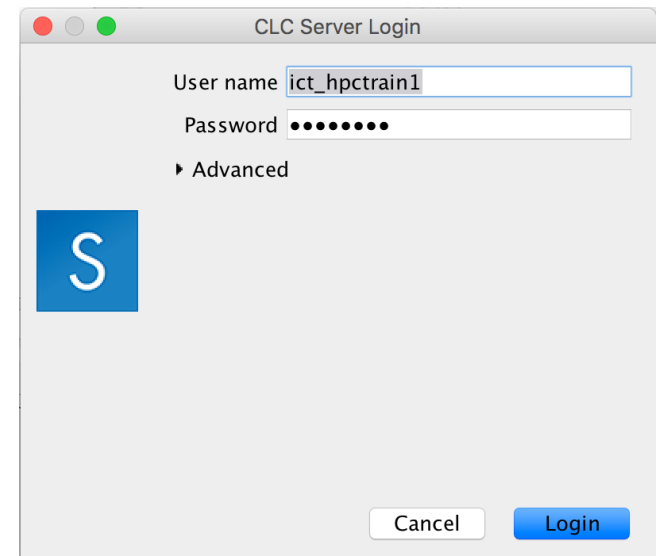
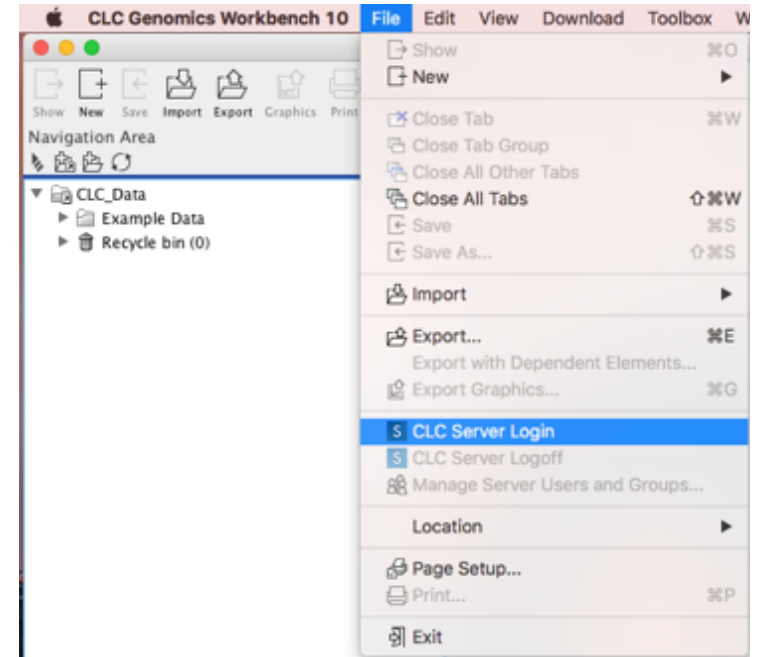


Main
viewing
screen

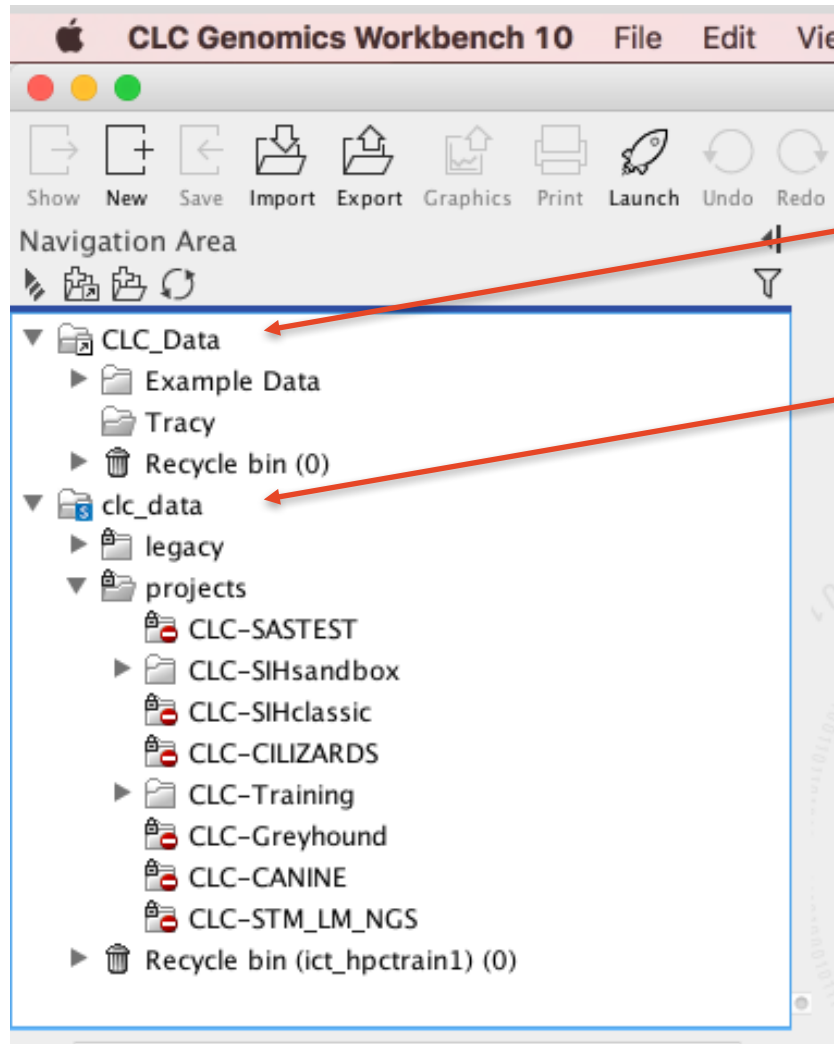
The CLC Server (Artemis)

- Connect to the Server by clicking:
 - File > CLC Server Login
- Enter your assigned training unikey details

Note: You must be connected to the University network



CLC Genomics Workbench



CLC_Data is on your desktop

clc_data is on the server
(Artemis)

- Legacy: data from pre-CLC on Artemis users. We will not use this.
- Projects: CLC-project (DashR). We will use CLC-Training.
- This only appears when you are connected to the server

Artemis HPC

The high performance computing (HPC) cluster gives you access to powerful compute resources:

3 Phases of Artemis:

- 7,636 cores
- 45 TB RAM

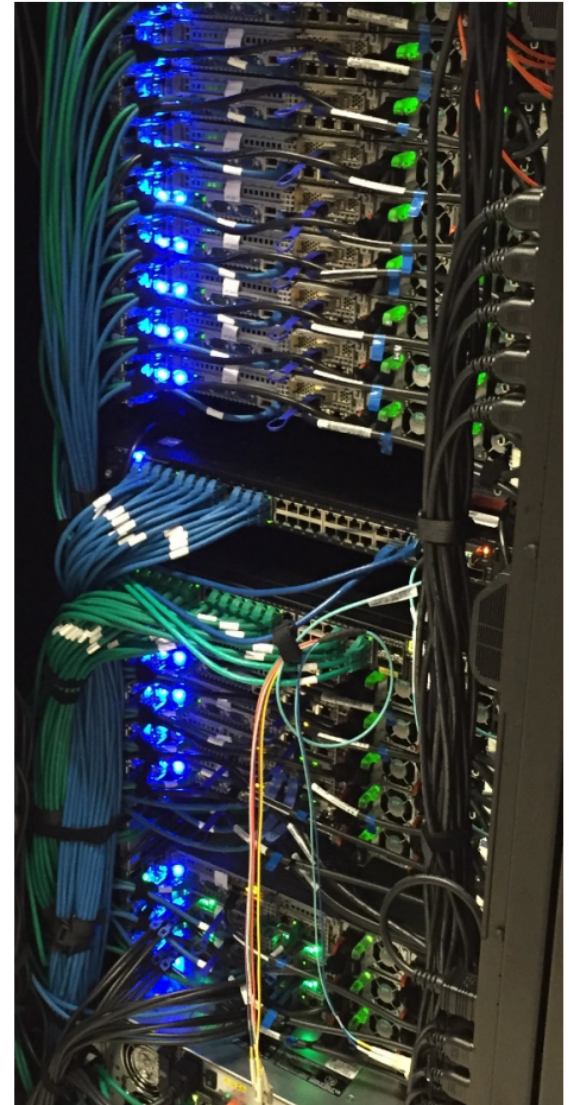
This will enable CLC users to run analyses with:

- Longer “walltime”
- Higher throughput
- Large input/output data
- Enhanced data security/permissions control

(Attend our Intro to Artemis course or read more here:
<https://informatics.sydney.edu.au/services/artemis/>)

This resource is shared across the University using a ‘job queuing system’ (PBS Pro job scheduling software)

- We can access Artemis to check the status of our CLC Genomics Workbench jobs



Connecting to Artemis

Artemis uses CentOS 6.9 which is a **Linux operating system**.

To access Artemis, you need:

- Connection to the University network
(connect via VPN if you are off-campus)
- A terminal client

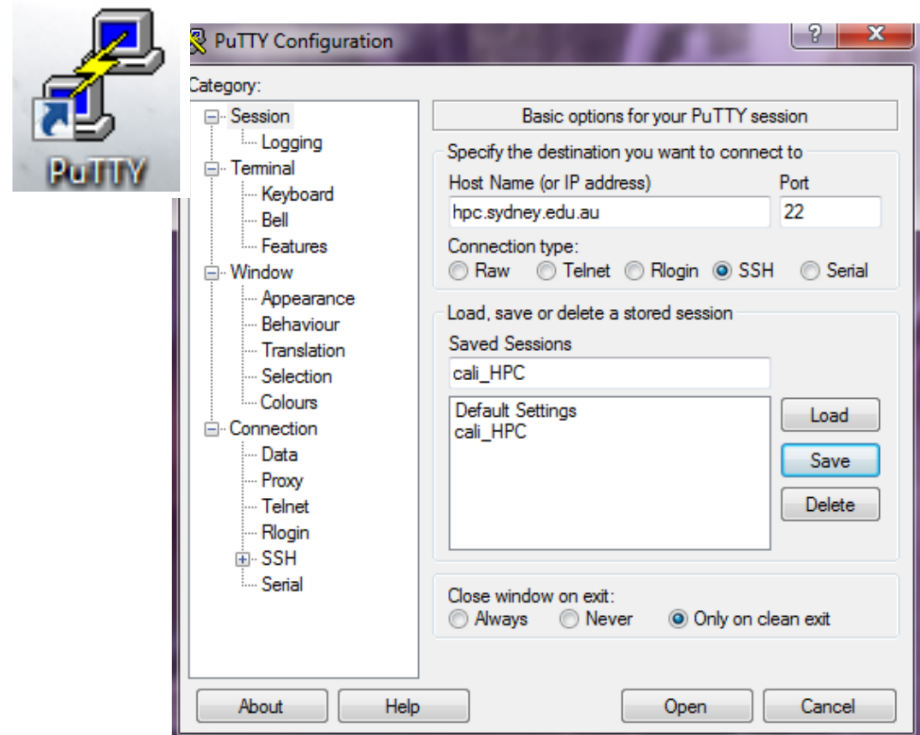
Mac	Windows
Use the existing terminal application	Putty

NOTE: you can also access the RDS through the terminal client

Terminal client

Windows users

- Go to: <http://www.putty.org/>
- Download and run 'putty.exe'
- In the configuration window, enter the following:
 - Under 'Host Name':
hpc.sydney.edu.au
 - Leave 'Port' as 22
 - Click 'Open'
 - At 'login as' enter your training unikey
 - Enter the training unikey password



Mac users

- Go to Finder > Applications > Utilities
- Double-click Terminal
- Type:

ssh ict_hpctrainN@hpc.sydney.edu.au

Training password

```
tracy — ict_hpctrain1@login2:~ — ssh ict_hpctrain1@hpc.sydney.edu.au — 80x24
tracys-mbp:~ tracy$ ssh ict_hpctrain1@hpc.sydney.edu.au
ict_hpctrain1@hpc.sydney.edu.au's password:
Last login: Fri May 18 12:49:54 2018 from vlan-639-10-17-2-128.staff.wireless.syd
ney.edu.au

***** ARTEMIS 3.0 *****

*** !!! LOGIN NODES ARE FOR JOB SETUP & SUBMISSION ONLY !!! ***

Use Interactive flag (-I) or defaultQ or small-express queue for job testing

NO ACTUAL COMPUTE WORK IS TO BE CONDUCTED ON LOGIN NODES

Any sustained compute or multicore/memory intensive process on login nodes
will be subject to being killed without notice.

!!! MEMORY MUST BE SPECIFIED FOR ALL JOBS - use the mem= PBS directive !!!

Please refer to the Quick Start User Guide to get started with Artemis 2.0:
https://informatics.sydney.edu.au/services/artemis/

[ict_hpctrain1@login2 ~]$
```

Research Data Store (classic RDS and RCOS)

Both **Research Data Store** (RDS) systems provide you with a safe place to keep important data. RDS systems are constantly backed up. Each RDS system is slightly different:

Classic RDS:

- More suitable if you work on your local desktop
- CIFS (Windows operating systems)

Research Optimised Compute Storage (RCOS):

- Good if you use Artemis HPC directly
- NFS (Unix/Linux operating systems)

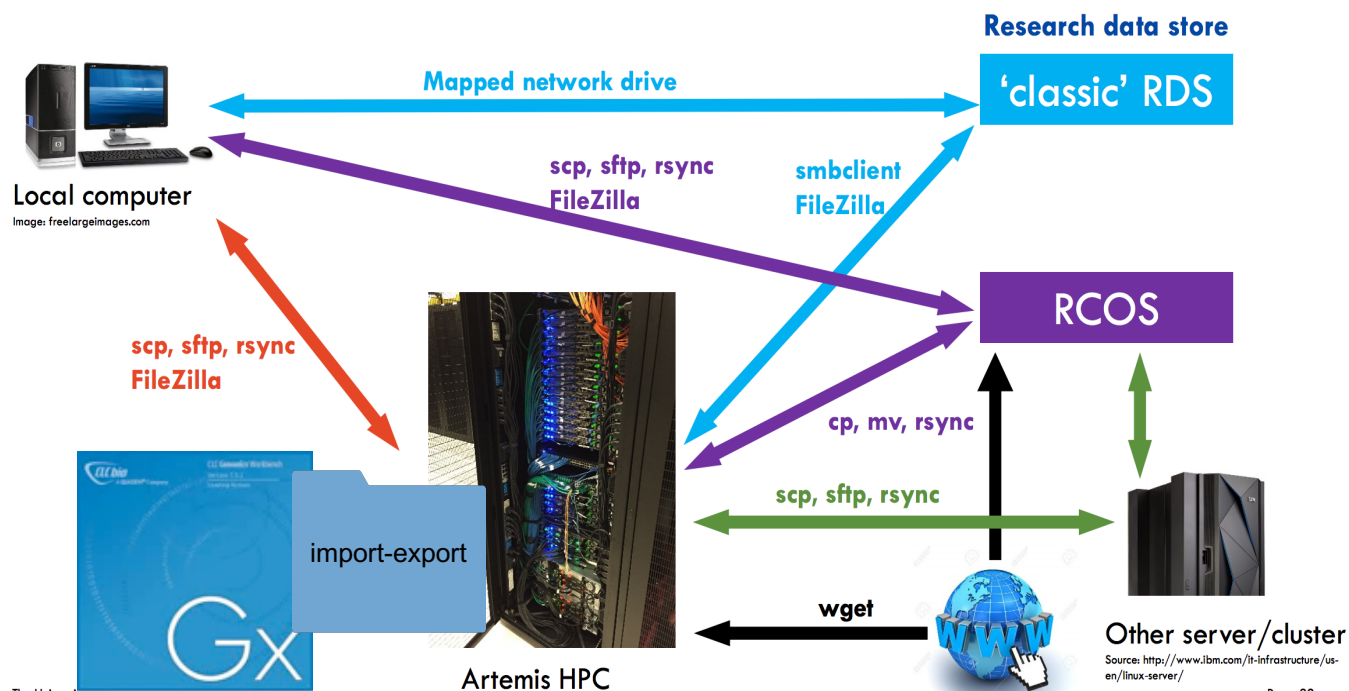
Data transfer methods

There are different ways of transferring data between the different systems involved with the CLC Genomics Workbench.

(Go to the Data transfer and RDS for HPC course to learn more!)

Today we will learn how to:

- Map RDS to local computer
- Use FileZilla
- Use Import-export



Today's exercise

A mutation in the *TYRP1* has an effect on canine coat colour.

We have NGS data of a Kelpie.

The task is to determine if this Kelpie is:

- Brown/red (TT); or
- Black (CT, CC)

Mammalian Genome 13, 380–387 (2002).
DOI: 10.1007/s00335-001-2147-2

**Mammalian
Genome**
Incorporating Mouse Genome
© Springer-Verlag New York Inc. 2002

TYRP1 and *MC1R* genotypes and their effects on coat color in dogs

Sheila M. Schmutz, Tom G. Berryere, Angela D. Goldfinch

Department of Animal and Poultry Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 5A8

Received: 27 November 2001 / Accepted: 20 March 2002



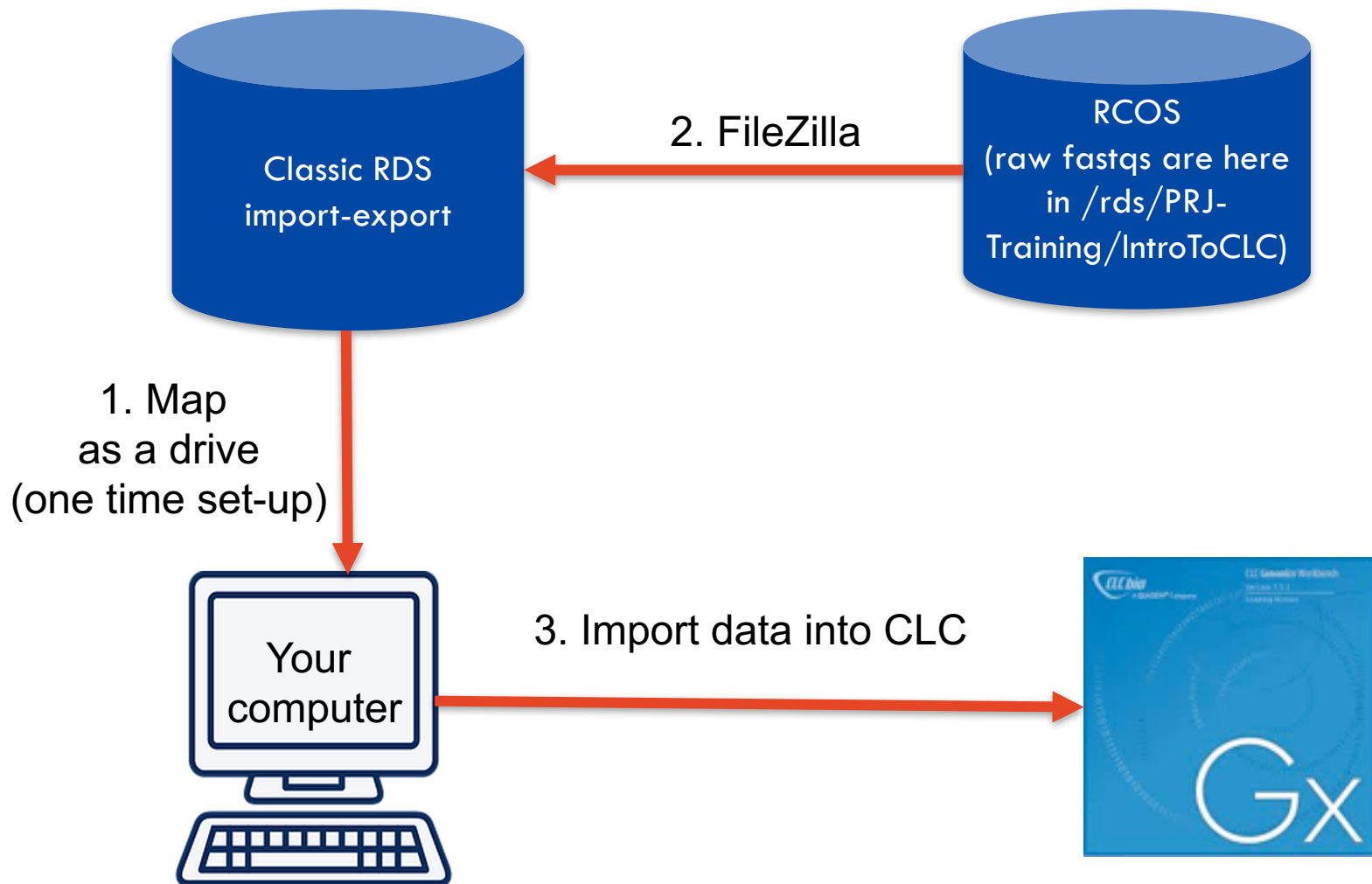
Today's exercise

Paired-end fastq data is stored on RCOS:

`/rds/PRJ-Training/IntroToCLC`

To get this data on CLC Genomics, we will:

1. Map import-export directory to our local computer
2. Use FileZilla to move the fastq's to import-export
3. Import the data into CLC-Training on CLC Genomics Workbench



The import-export directory

The import-export directory is on a classic RDS system.

It is useful for:

- Importing/exporting large data into and out of CLC
- Running data transfers in the background (meaning you can start your import/export, then close CLC and your computer)
- This saves your own computer resources to do other tasks while transfers occur

NOTE: Import and export functions in CLC include both a data transfer and **conversion** step that converts files into native CLC format (which can take extra time)

You can still import data **directly** from your computer (or any server mounted to it) to CLC Genomics Workbench. This is advisable only for small datasets (using the Workbench option).

1. Mapping the import-export directory

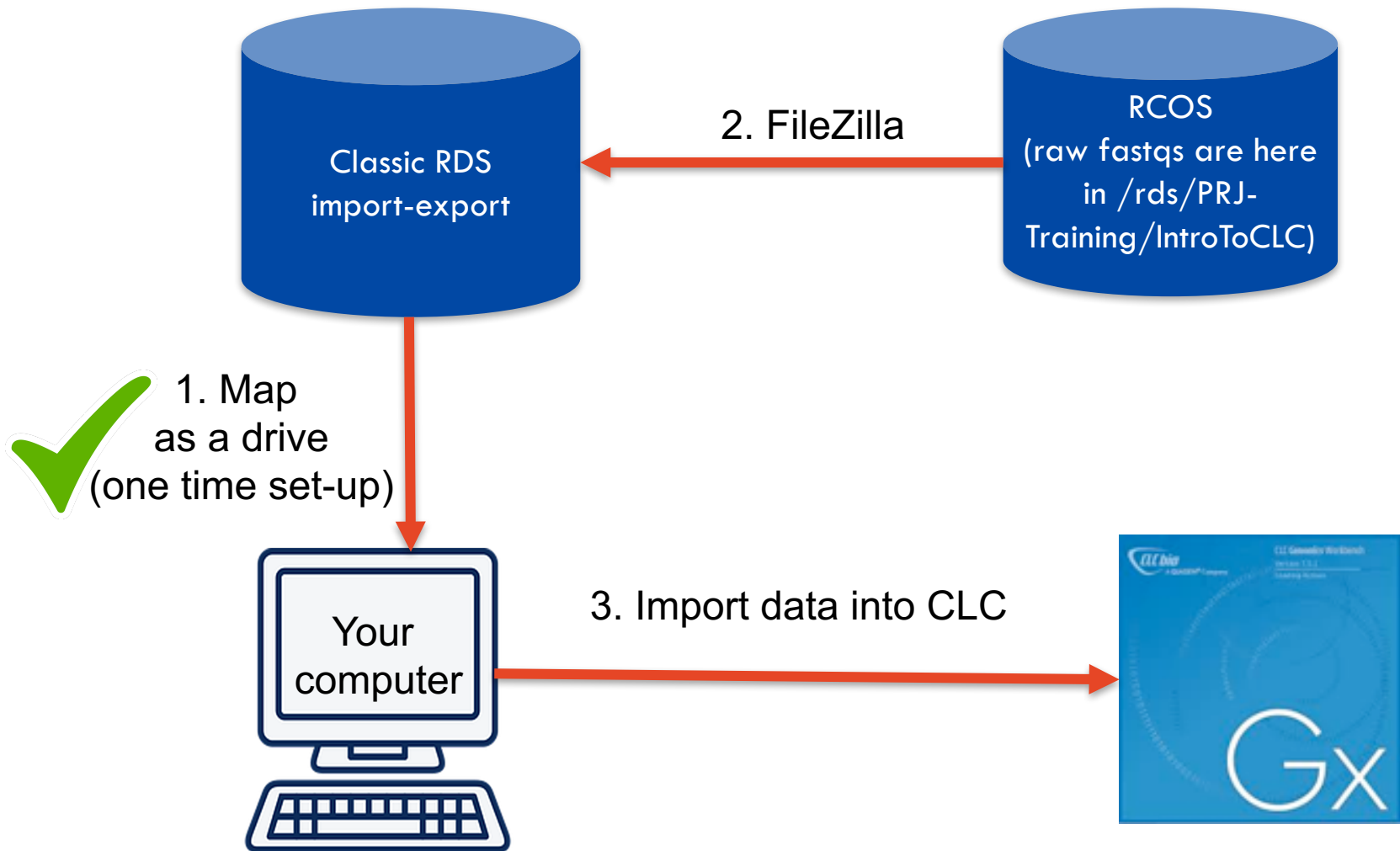
The import-export directory is on a classic RDS system.
Any Classic RDS can be mounted to your local computer.

Windows users (8/10):

- Click on This PC from the Desktop. On the Computer tab, click on Map network drive in the Network section.
- Choose a drive letter and enter your Classic RDS path:
`\\research-data.shared.sydney.edu.au\RDS-01\PRJ-CLC\import-export`
- Tick Reconnect at logon. Tick the Connect using different credentials box.
- Enter SHARED\ict_hpctrainN and the training password. Click Finish.

Mac users:

- Open a Finder window. Click Go > Connect to Server
- Type your Classic RDS address: `smb://research-data.shared.sydney.edu.au/RDS-01/PRJ-CLC/import-export`
- Click on the + button
(only perform this step if you want the network drive to display on your desktop when you log in).
- Click Connect. Select Connect as Registered User.
- Type SHARED\ict_hpctrainN in the Name field. Type the password in the password field.
- Click Connect. A finder window will open displaying your RDS folder



FileZilla

FileZilla can be used to transfer files between your local desktop and another server.

We can now use FileZilla to move our fastqs on RCOS (/rds/PRJ-Training) to classic RDS that has been mapped to your local computer (import-export).

FileZilla is a free file transfer client

- Go to <https://filezilla-project.org/>
- Download the client, not the server



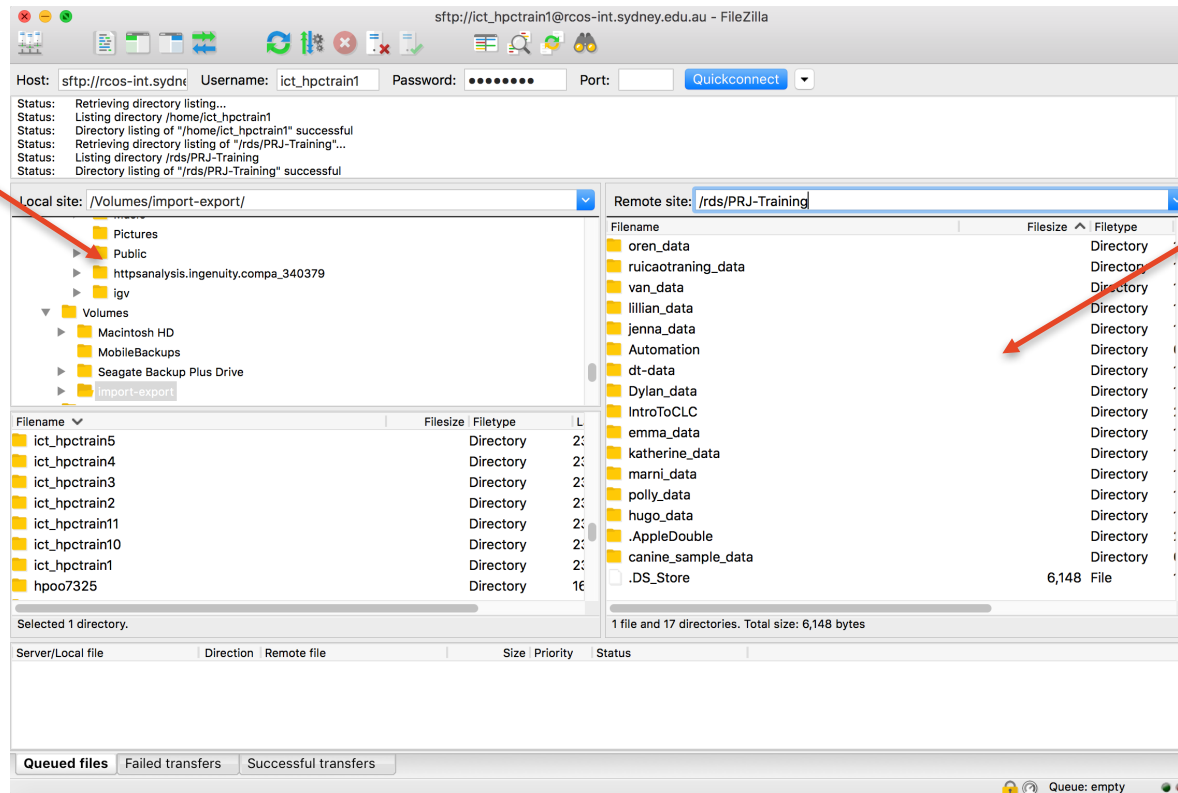
FileZilla

- Open FileZilla
- At host, type:
 - sftp://rcos-int.sydney.edu.au
- Username and password fields, type:
 - ict_hpctrainN and the training unikey password
- Leave port blank (default)
- Click Quickconnect



FileZilla

Your local
(including
anything
mounted to
your
desktop)
files will
appear
here



Your remote
(e.g. Artemis,
RCOS) files
will appear
here

FileZilla

- Navigate to the import-export directory on the local side
- Navigate to the PRJ-Training directory on the remote side
- One way is to type the path as below, or you can click into it (the “..” directory is to navigate back to the parent directory)

For local site, type:

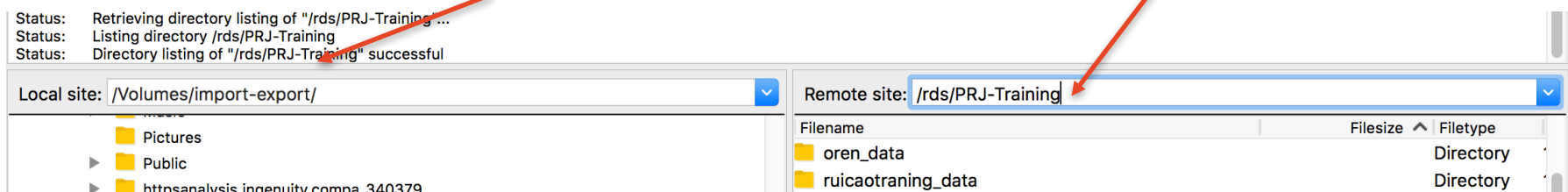
/Volumes/import-export (Mac) or

Z:\PRJ-CLC\import-export\

(Windows – your drive letter may be different)

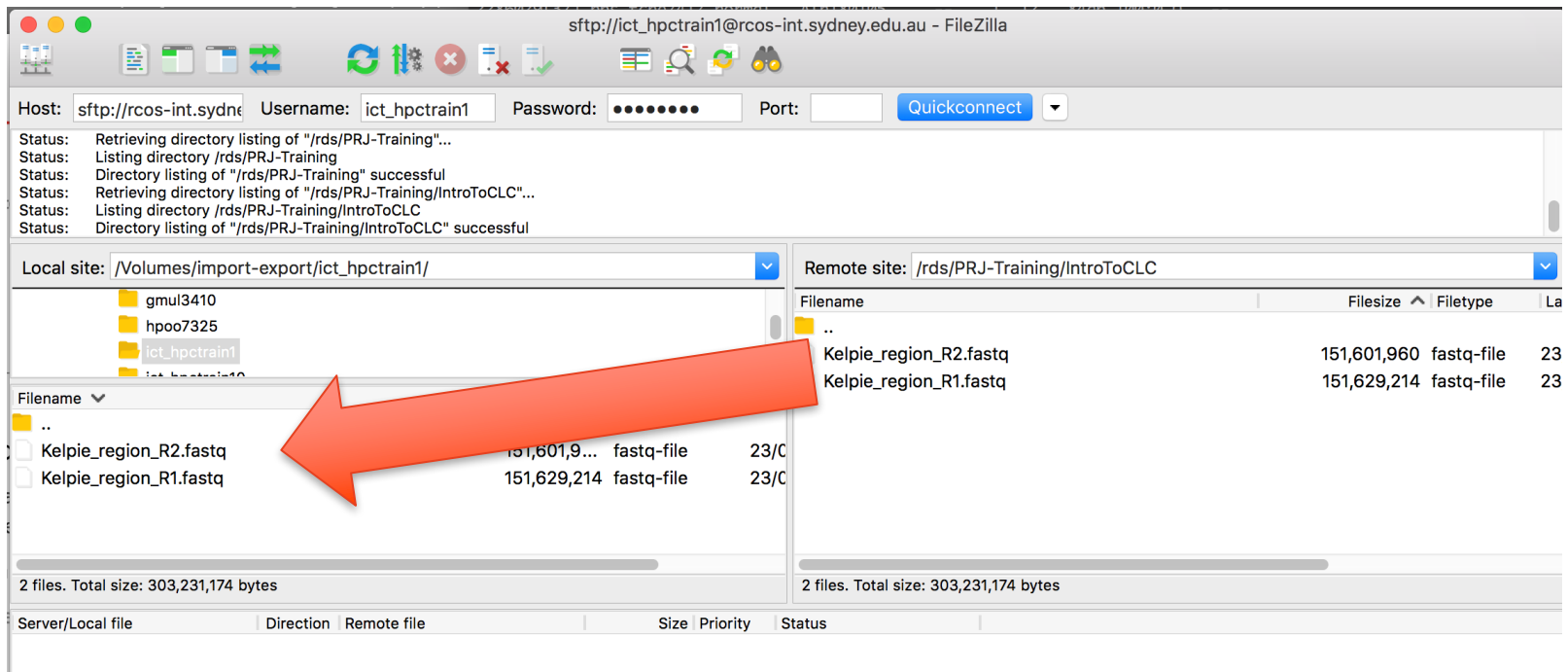
For remote site, type:

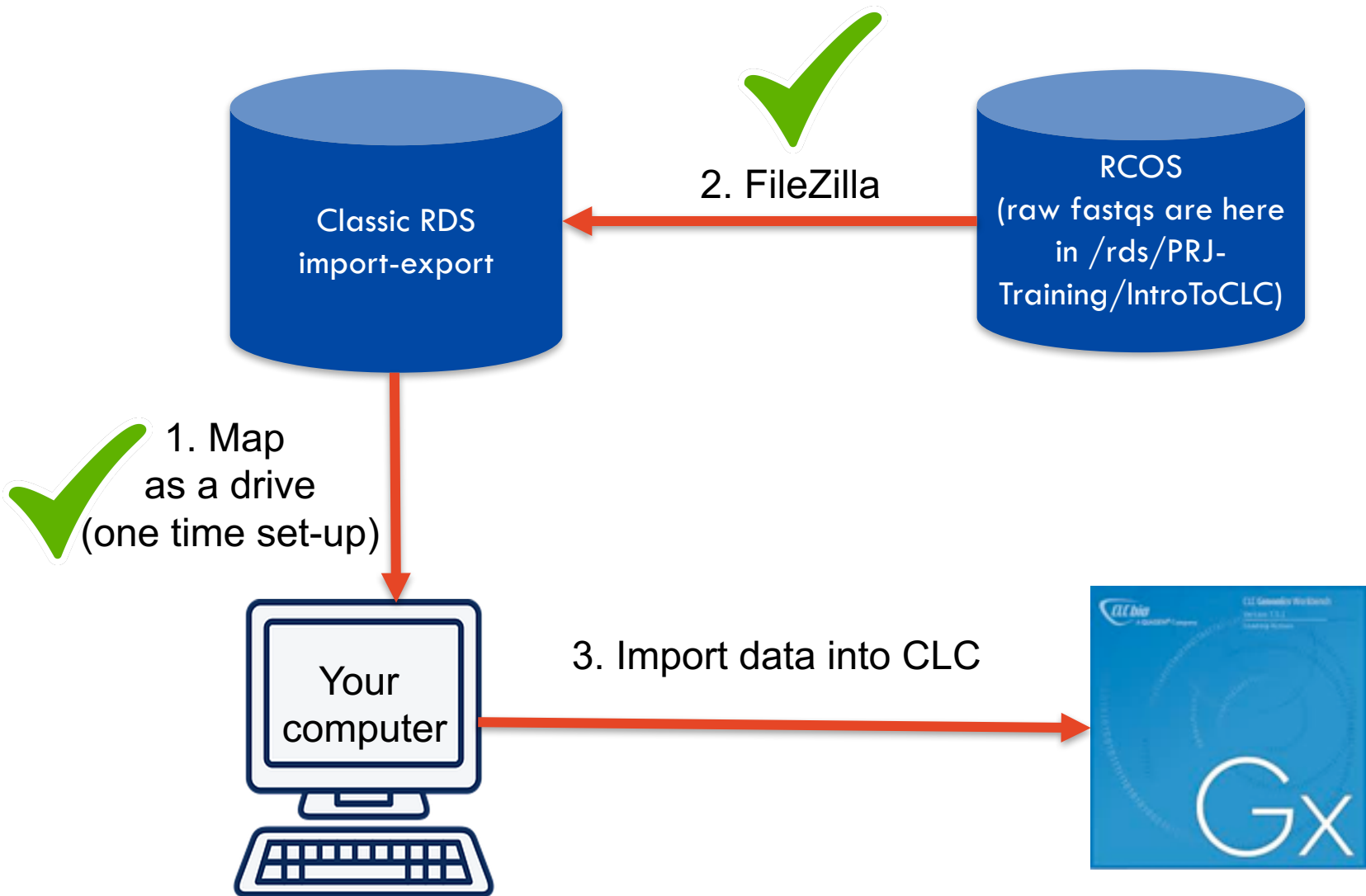
/rds/PRJ-Training



Transferring from RCOS to import-export


- On the local site, double click into your training unikey folder
- On the remote site, double click the 'IntroToCLC' folder
- Highlight the fastq files at the remote site, drag and drop them into the local site

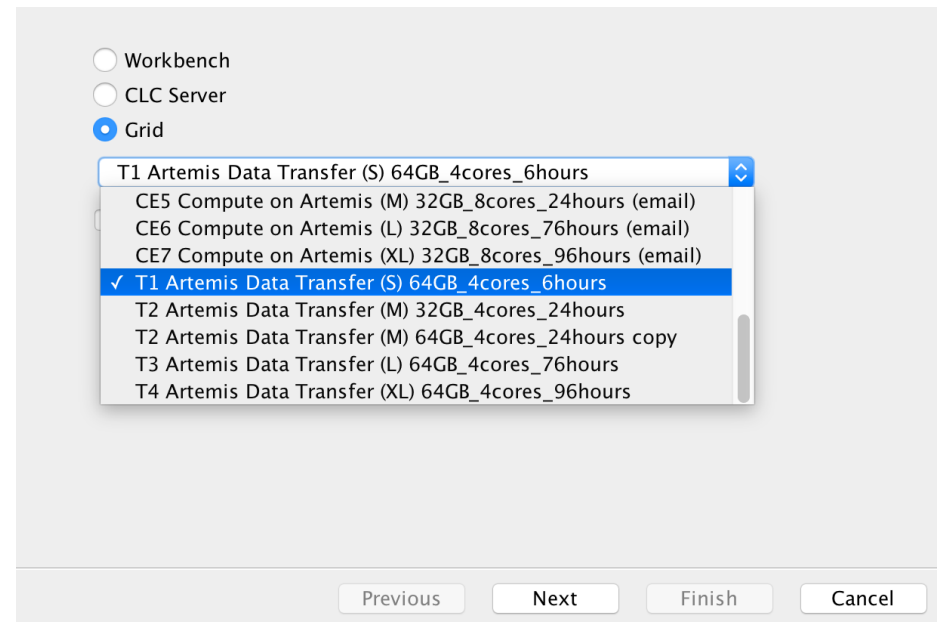




3. CLC Genomics Workbench – importing data

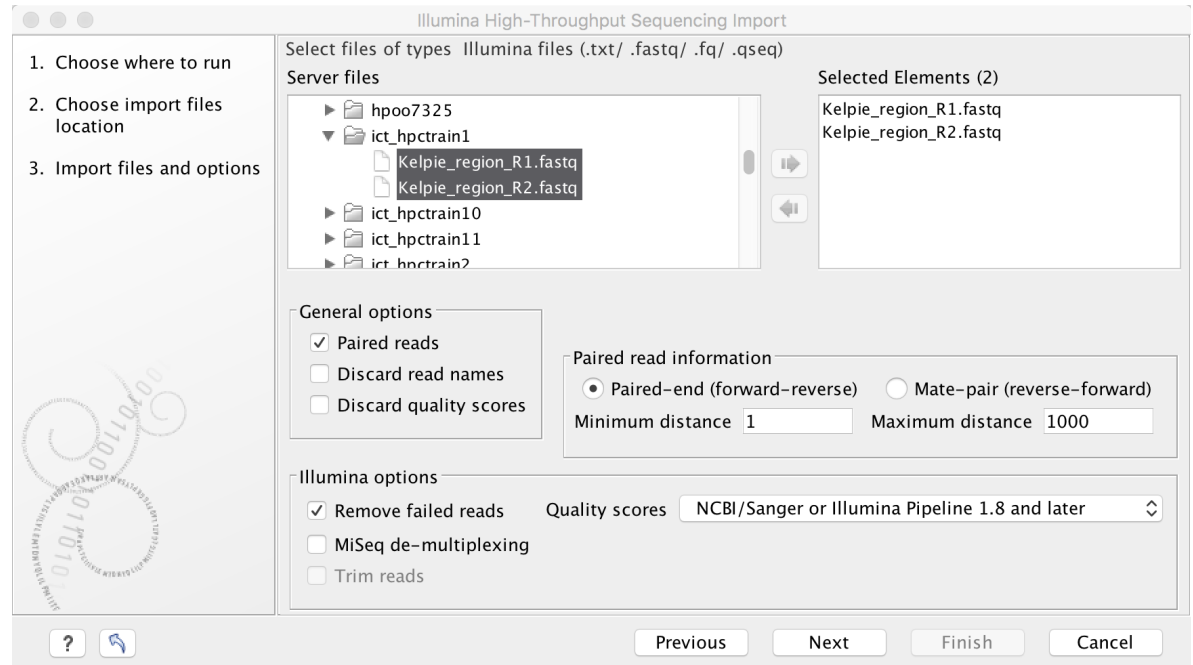
Now we are ready to import our fastq's on import-export into CLC-Training

- Go to the CLC Genomics Workbench
- Click import  > Illumina
- Leave Grid selected.
Notice the different nodes and computational resources
“Compute on Artemis” and
“Artemis Data Transfer”
- Select the most appropriate
(T1 Artemis Data Transfer (S) is sufficient for this exercise)
- Click Next



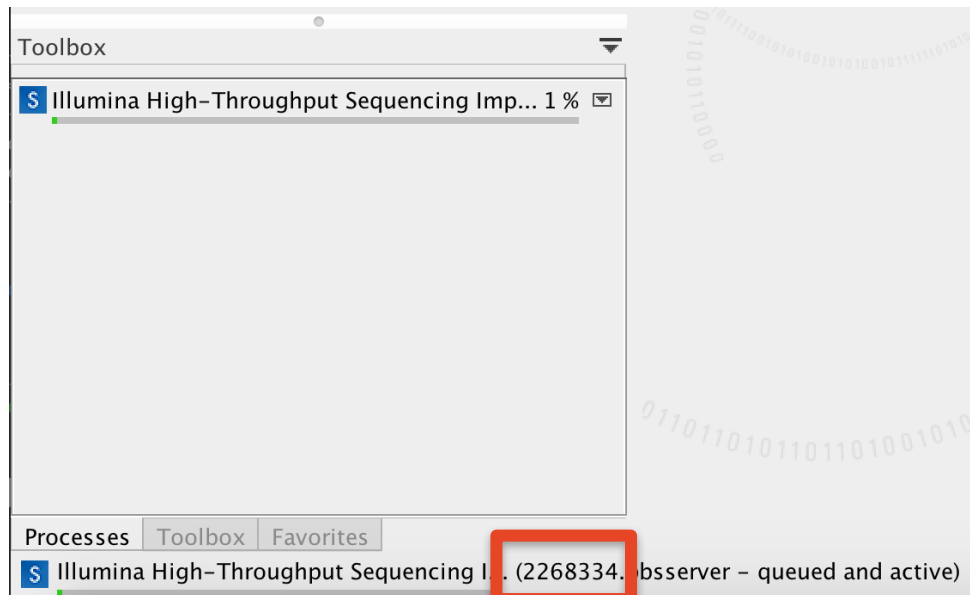
CLC Genomics Workbench – importing data

- Click 'On the server or place that the server has access to' (the other option is for files on your local computer)
- Under import-export, enter your training unikey directory
- Select both fastq files and click the right arrow to move them into selected elements
- Tick Paired reads
- Click Next



CLC Genomics Workbench – importing data

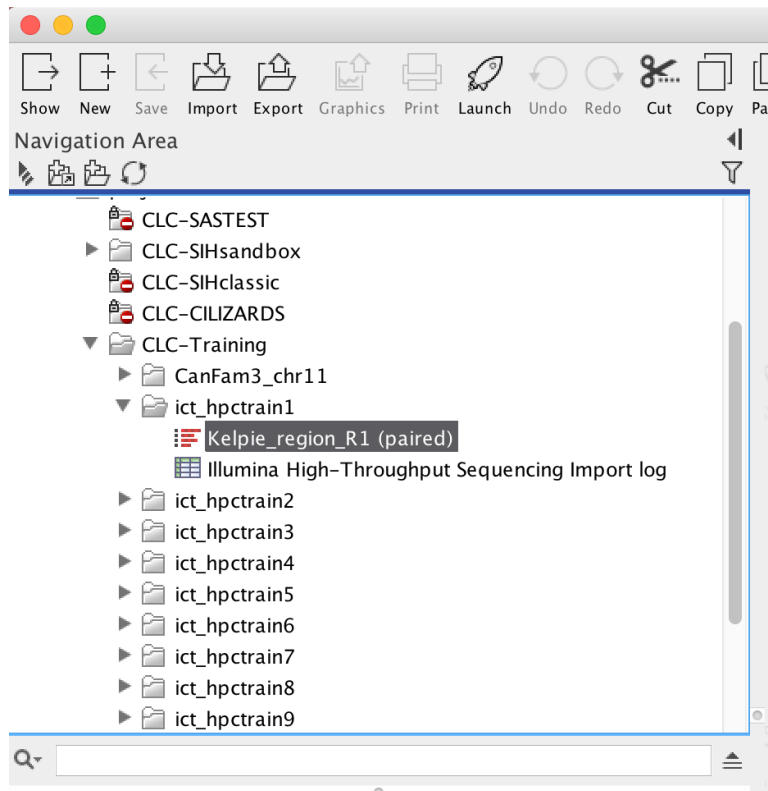
- Select Save option, click Next
- Under CLC-Training, click your training unikey directory and click Finish
- On the bottom left, you can see the your 'job number' and any processes that you are currently running

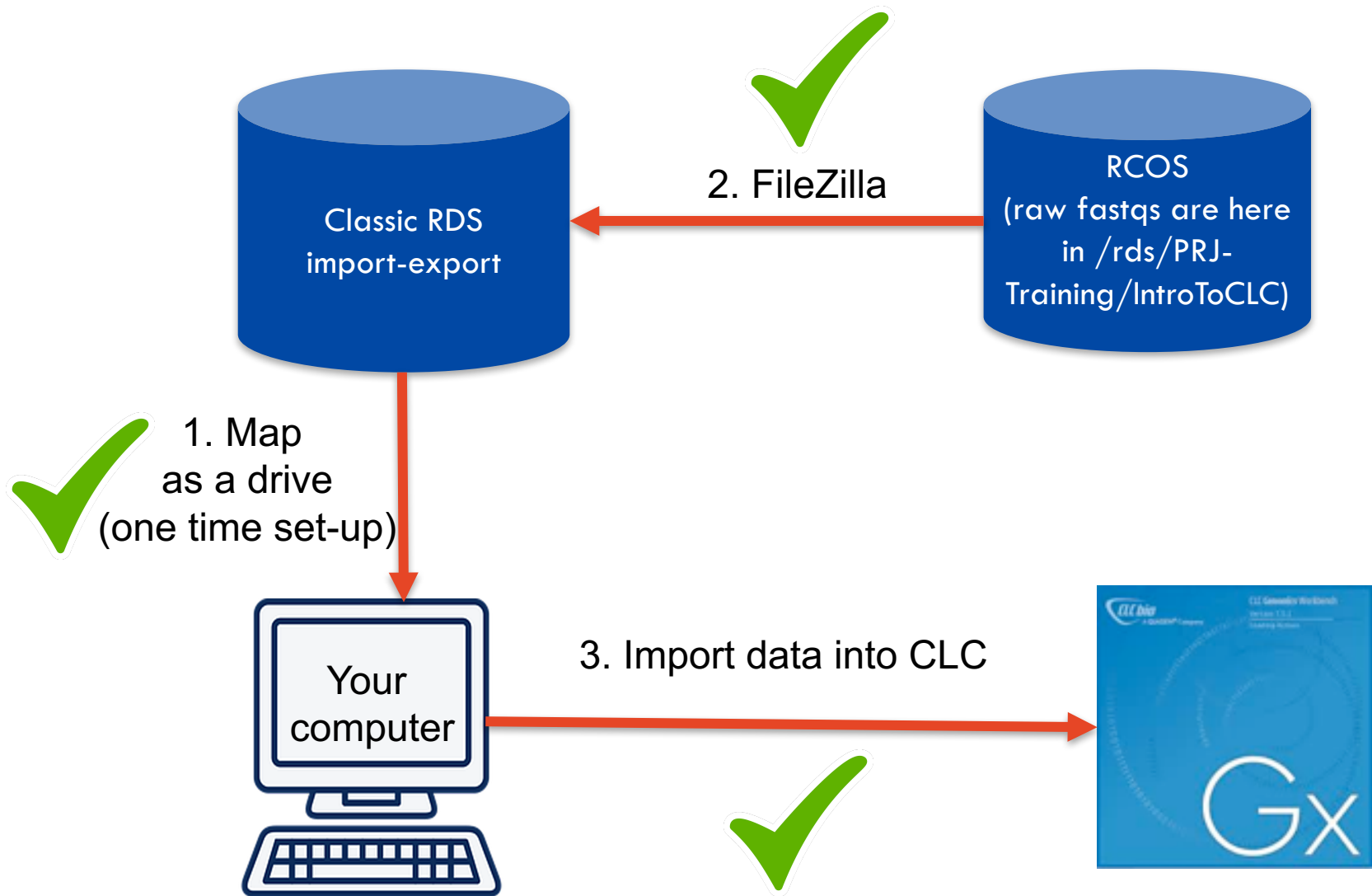


We will use this job number to check the job status on Artemis later

CLC-Training

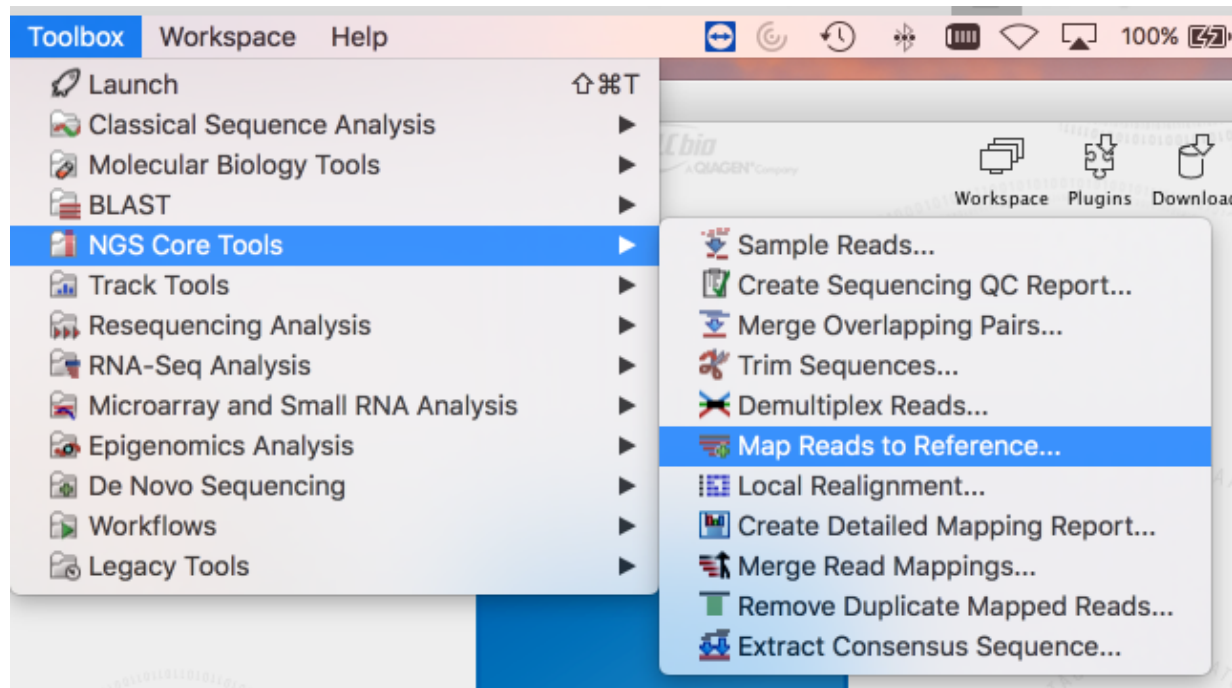
- You should now be able to see your raw data in CLC-Training
- Note: CLC puts both fastq files into a single (paired) file





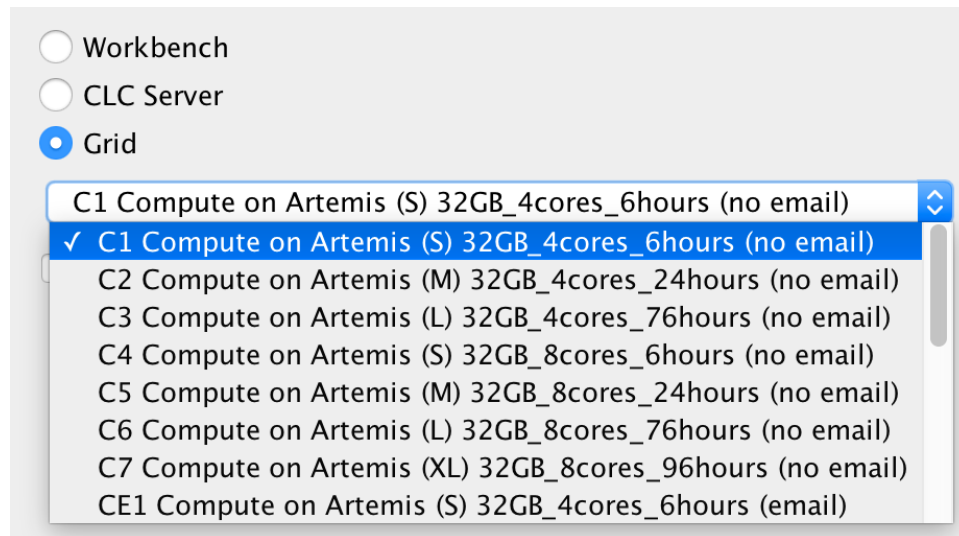
Aligning data using Artemis compute queues

- We will now align these raw reads using computational resources provided by Artemis
- Click Toolbox > NGS Core Tools > Map Reads to Reference



Aligning data using Artemis compute queues

- Select Grid
- Notice how there are 'Compute on Artemis' options with no email and ones with email. The email option notifies you of when your job has completed so you will know when to check back into Artemis
- The training unikeys are not attached to any email so we will select the small option with no email

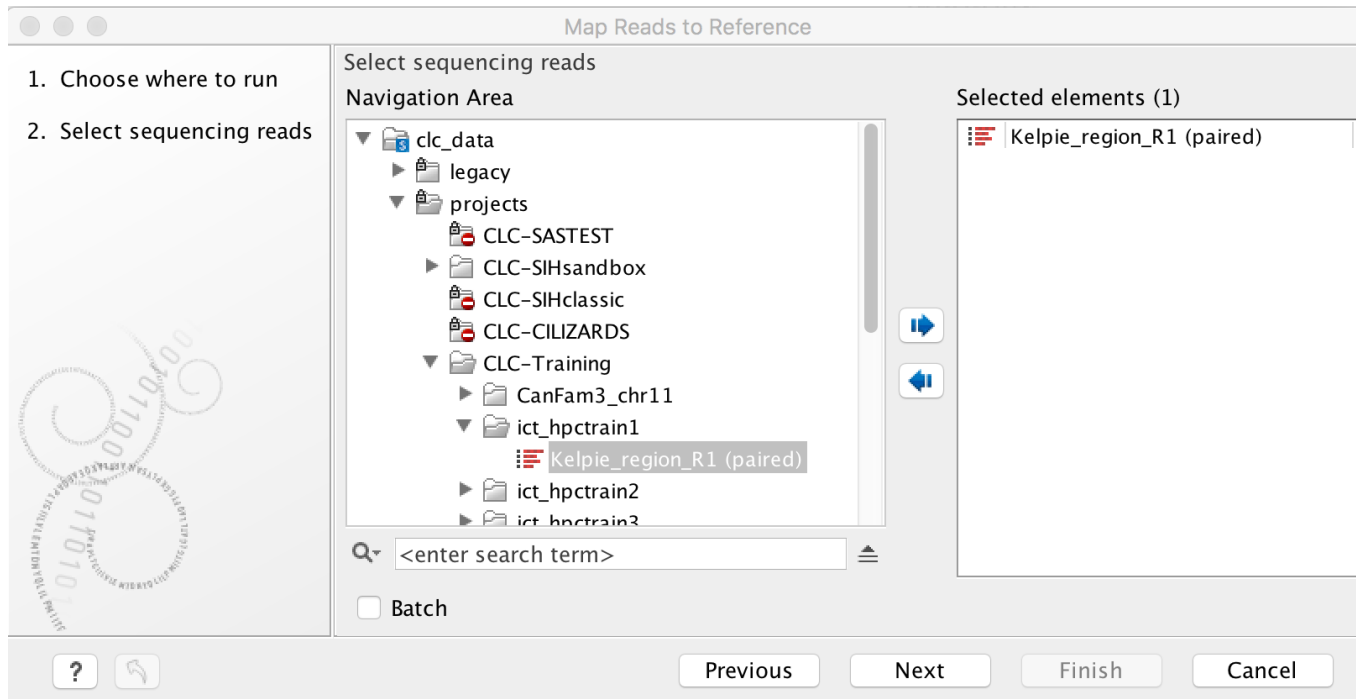


Queues

- Artemis uses PBS Pro Scheduler
- Selecting the most appropriate queue will ensure your process or “job” will start quickly and run to completion
 - Too few resources: jobs can fail
 - Too many resources: you may be pushed back in the queue and the starting of your job will be delayed
- This system is to ensure fair use of Artemis amongst all users in the University

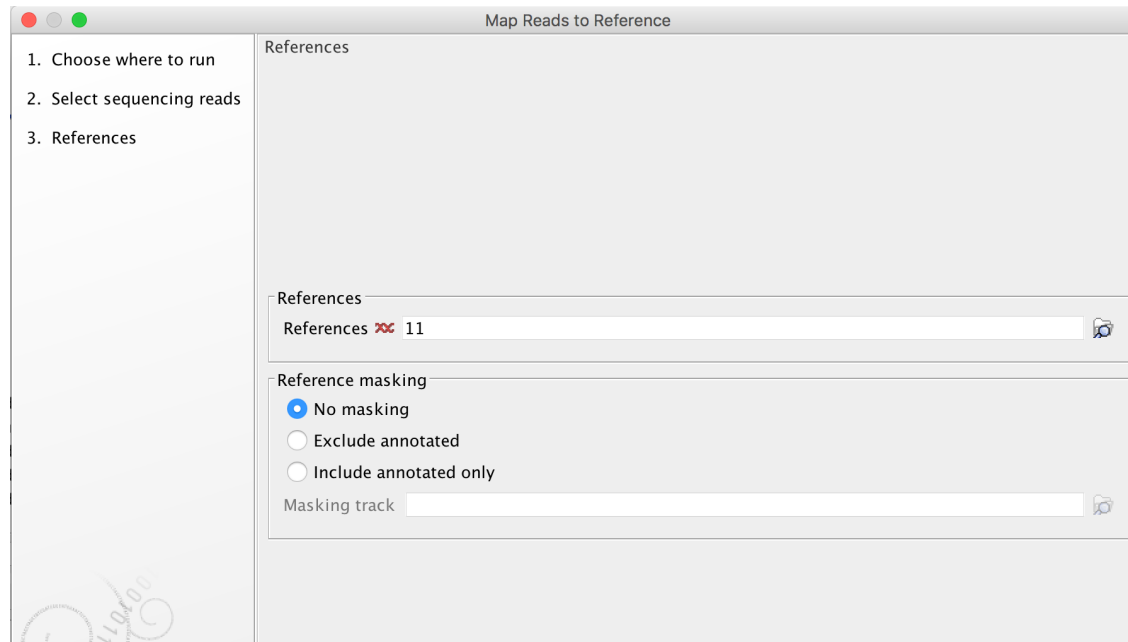
Aligning data using Artemis compute queues

- Select the raw sequencing reads (in your training unikey folder under CLC-Training)
- Select Next



Aligning data using Artemis compute queues

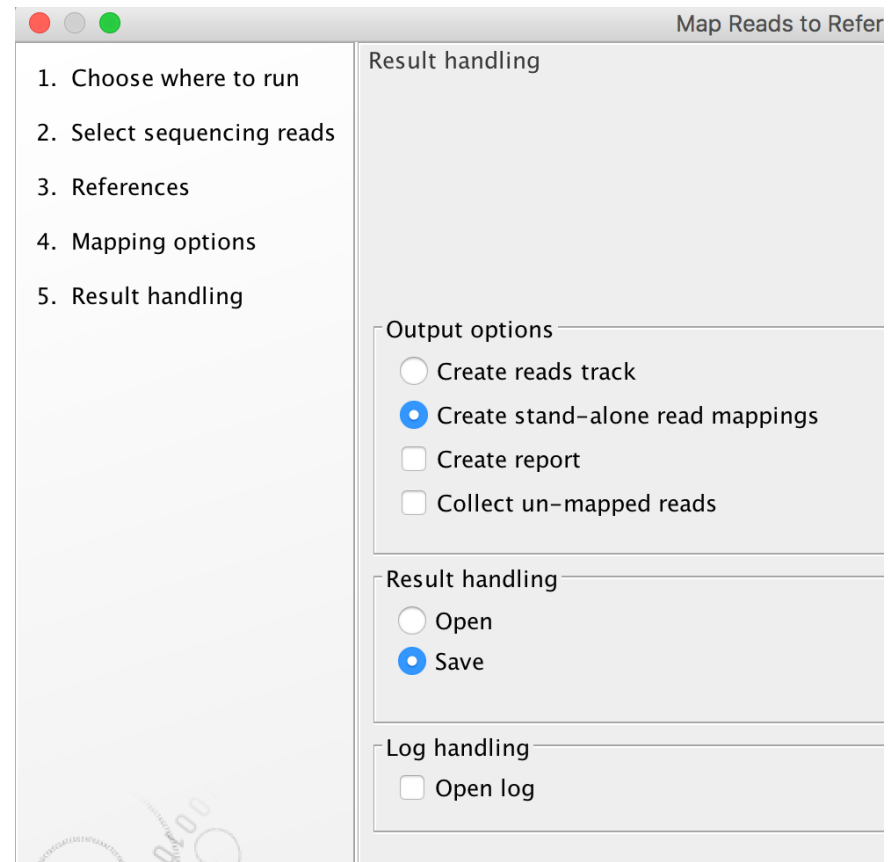
- Select the reference genome (11 > for chromosome 11)
- You can find the reference genome in CLC-Training > CanFam3_chr11



- Click Next until you get to step 5

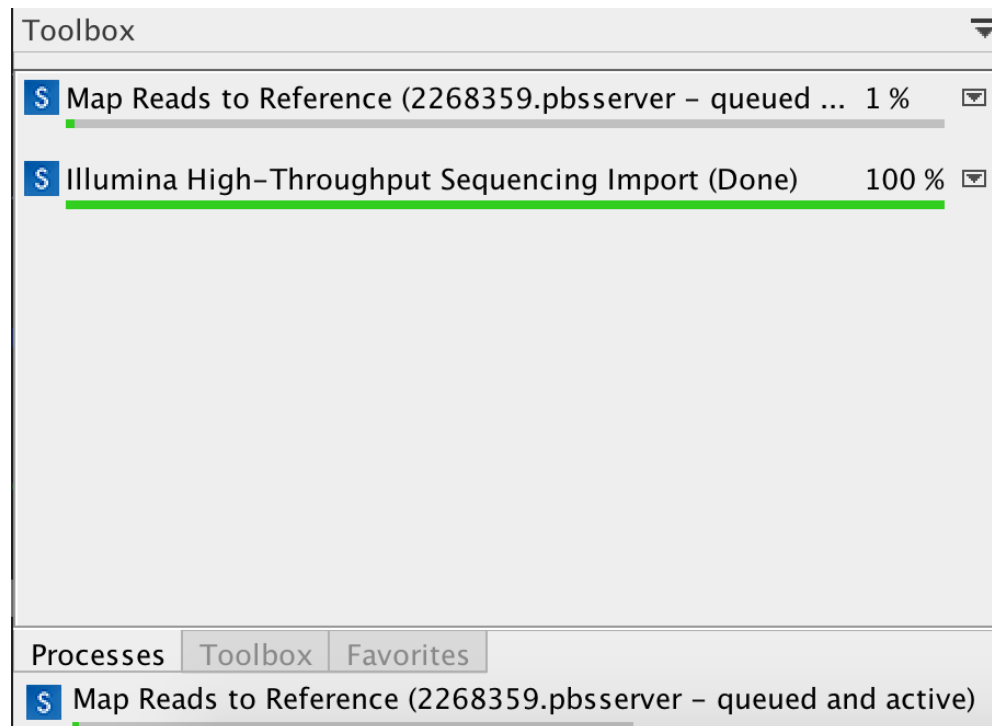
Aligning data using Artemis compute queues

- At step 5, select “Create stand-alone read mappings”
- Select Next
- Save to your unikey folder under CLC-Training



Checking the status of your job on Artemis

- Identify your job ID on CLC
- Go to your terminal where you have logged on to Artemis (putty for Windows users)



Checking the status of your jobs on Artemis

- At the command line, type:
 - `qstat <your job ID number>`

```
[[ict_hpctrain1@login2 ~]$ qstat 2268359
```

Job id	Name	User	Time Use	S	Queue
2268359.pbsserver	Map_Reads_to_R	svc_clc_nttgenom	0	Q	small

```
[[ict_hpctrain1@login2 ~]$
```

- Under 'S' (Status), the 'Q' indicates that the job is in queue. You may see 'R' (running) or 'F' (finished).
- You can also use this to view how long your job has been running for, mine ran for 1 minute 52 seconds at finish

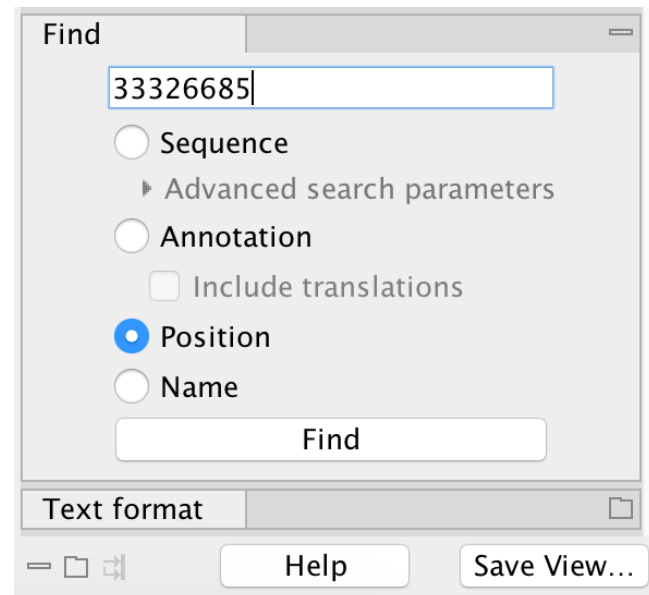
Job id	Name	User	Time Use	S	Queue
2268359.pbsserver	Map_Reads_to_R	svc_clc_nttgenom	00:01:52	F	small

Checking the status of your jobs on Artemis

- As we used compute nodes on Artemis, we could have closed CLC Genomics whilst this analysis ‘job’ was running in the background
- You can check the status of your job anytime on Artemis without having to book and log onto CLC Genomics Workbench (and subsequently using one of the 11 licenses)

Is this Kelpie black or brown?

- The mutation that affects coat colour is located in *TYRP1* at chr 11: 33,326,685
- Double click on the aligned file
- Navigate to the mutation by searching the above position in the right hand panel in “Find”



Viewing alignments in CLC Genomics

- Is this Kelpie black or brown?
 - Brown/red (TT); or
 - Black (CT, CC)



CLC Genomics Workbench requirements

System requirements

- Windows
- Mac
- Linux

Other requirements

- University of Sydney unikey
- DashR project
- CLC Genomics Workbench subscription
- Access to SIH PPMS

DashR

<https://dashr.sydney.edu.au/>

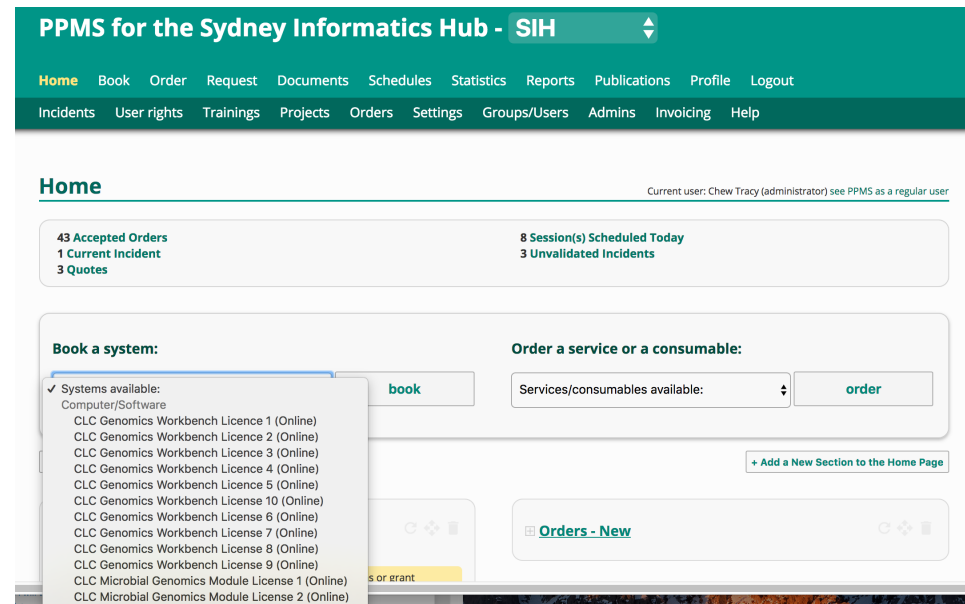
- A dashR project gives you access to Artemis and RDS (classic or RCOS)
- You can create a new one or nominate an existing project
 - We will create a CLC-Project folder for your project members
 - In this course, the training unikeys were members of the DashR “Training Project”. All members of one project can read/write in the CLC-Project directory.
- Subscribers are also added to ‘CLC Genomics Active Users’ project which gives you access to the license

CLC Genomics subscriptions

- Subscriptions to the CLC Genomics Workbench gives you full access to:
 - The CLC Genomics Workbench
 - Microbial Genomics Module plugin
- Subscriptions cost \$750 (incl. GST) for 6 months of access (>80% of the cost is subsidized by the University)
 - You can include multiple lab members in a single subscription
 - Only one user from one group subscription can use CLC at any one time
- We currently have 11 licenses that are shared amongst users across the University
- Contact sih.info@sydney.edu.au if you would like a free 4 week trial license

The PPMS booking system

- PPMS for the Sydney Informatics Hub is where you can order a subscription
- Subscribers then need to use the PPMS calendar to book the times that they wish to use CLC
- This ensures that the 11 licenses are shared fairly amongst subscribers



Other SIH services

We also provide access to Ingenuity Pathway Analysis

- Free if you have a University of Sydney unikey
- 1 license is shared University-wide

Training courses for

- DNA sequence analysis
- RNA sequence analysis
- Artemis HPC
- Data transfer and RDS for HPC
- Hacky Hour

Check our website for full details and to register:

<https://informatics.sydney.edu.au/services/training/>

Acknowledgements

DNA sequence sample data (sub-sampled) was contributed by Prof. Claire Wade's lab in the School of Life and Environmental Sciences.

We thank ICT in particular Stephen Kolmann for the feedback on this course and for providing ongoing support for CLC Genomics.

If you use any SIH service, please acknowledge us in your paper! This supports our work and helps us grow our facilities that we can provide to you. By doing so, you can also enter our regular SIH publication incentives where you have a chance to **win \$1,000!**

Please contact sih.info@sydney.edu.au if you have any feedback, comments or suggestions.

Thank you for your participation! 😊