Maximizing Your Time with a CBHDS Biostatistician/Bioinformatician

Data Transfer and Management

Please consider the following guidelines when using Virginia Tech's Advanced Research Computing (ARC) for data storage. The purpose of this document is to provide general guidelines for the use of ARC for data transfer and management at the Center for Biostatistics and Health Data Sciences (CBHDS). Note that you do not need to have any prior experience with high-performance computing (HPC) to follow this document.

For all CBHDS bioinformatics projects, we rely on ARC for the storage and data analysis. ARC is a unit within Virginia Tech's Division of Information Technology, providing centralized research computing infrastructure and support the university's research community.

ARC Services

VT ARC provides high-performance computing systems, large-scale data storage, big data processing, software, and consulting services. The data storage and sharing feature is designed for researchers from different disciplines or in different locations to collaborate on a project. ARC is very helpful when you have very large jobs to run (e.g., high resolution or large-scale datasets), many jobs to run (parameter sweeps, many datasets), or needs for specialty hardware (GPU, memory, high bandwidth storage, fast network, scale).

You do not need to have any prior experience with high-performance computing (HPC) to get started. ARC also provides introductory training sessions throughout the year via the TLOS Professional Development Network and there are computational scientists available in the ARC team to provide the necessary support.

For VT Users

ARC is available to all Virginia Tech faculty at no cost. Researchers and groups have the option to add computing costs to grants or contracts through ARC's Cost Center, for additional computing or storage resources. Departments and faculty can also purchase priority access to an ARC system for up to five years through ARC's Investment Computing Program.

Links to Additional Resources

- Office Hours
- Request a Consultation
- Create an ARC User Account
- New ARC Users Info
- Video tutorials
- Computing Resources
- Storage Resources
- Visualization Services
- ARC Faculty & Staff Directory

1. Getting Started with ARC

This section contains instructions for creating user accounts, requesting project storage space and computation allocations, managing projects and user lists, as well as transferring and managing data on ARC.

1.1 Creating an ARC user account

1) ARC accounts are based on your Virginia Tech PID (VT-PID) account and make use of centralized Virginia Tech authentication.

- a valid VT-PID is required and logins to ARC systems are authenticated with VT-PID **username and password** plus **DUO second factor authentication**.
- All current faculty, staff, and students at VT have a VT-PID.
- Researchers external to Virginia Tech can get a "sponsored VT-PID" that can be requested by a faculty member at Virginia Tech.

2) Any user with an active VT-PID can request access to ARC system. The account creation form will require you to confirm your acceptance of VT usage policies, and then create your account. Once your account is created, you can login into ARC. To become fully functional, you will need to have access to a "**Project**", its "compute allocation" account, and optionally additional storage (see **Figure 1** below).

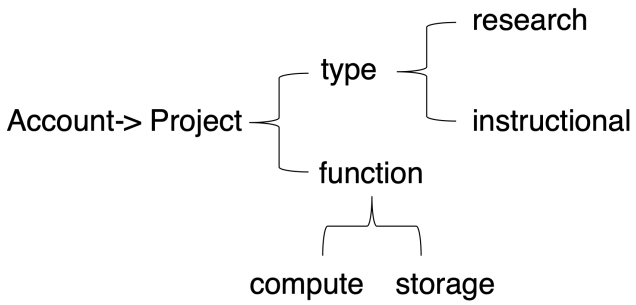


***Figure 1**. The relationship between project allocation type and function.*

1.2 Request Allocations

An allocation is a system time account requested and managed by a single person (e.g., a project PI). Many users (e.g., Co-PIs, Co-Investigators [Co-Is] or graduate students) can then be granted access to a single allocation. There are two types of projects (project allocations):

1) **Research Allocations**: for research projects and usually managed by the project's Principal Investigator (PI)

- typically granted for a single year and can be renewed annually for the length of the project.
- Multi-year research allocations may be granted through negotiation with ARC.
- Can be used for CBHDS data science projects.

Who is eligible for Research Allocations?

- a current faculty member or post-doctoral researcher at Virginia Tech.
- an employee of Virginia Tech and the Principal Investigator (PI) for a research computing-related project.
- an employee of Virginia Tech and the Co-PI or Co-I for a research computing-related project led by a non-Virginia Tech PI Adjunct professors must provide a letter from their department chair, indicating that they are qualified to lead an internal research project, before their project and allocation requests can be approved.

2) **Instructional Allocations**: for academic classes and are managed by the faculty member/instructor

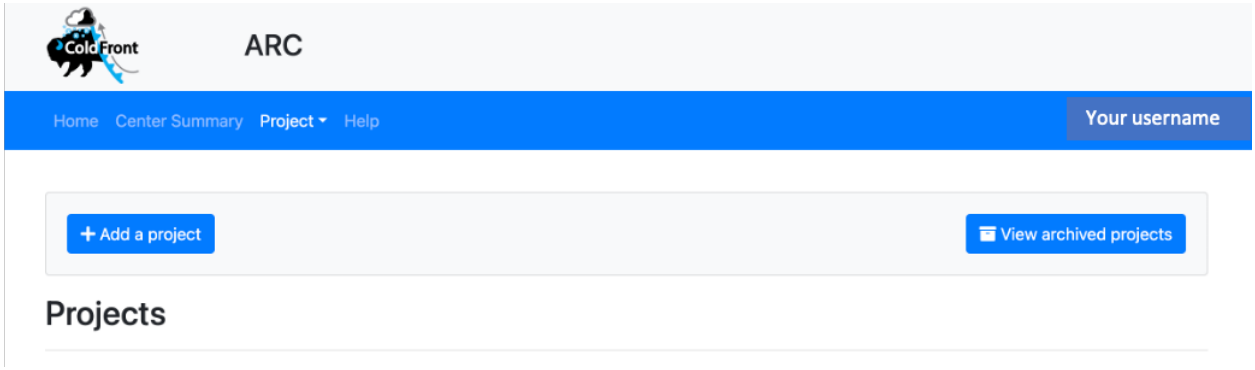
- typically smaller, available for shorter time periods
- limited to a select set of systems
- Can be used for training workshops

Once your ARC account is approved, you can go to ColdFront to create a project, and add users, allocations, and grant/publication information to it.

1) log into ColdFront using your ARC account username and password

2) from the home page, click "projects" and then "add a project" to create a new project

3) fill and submit the form to include information on title, project description, and department.

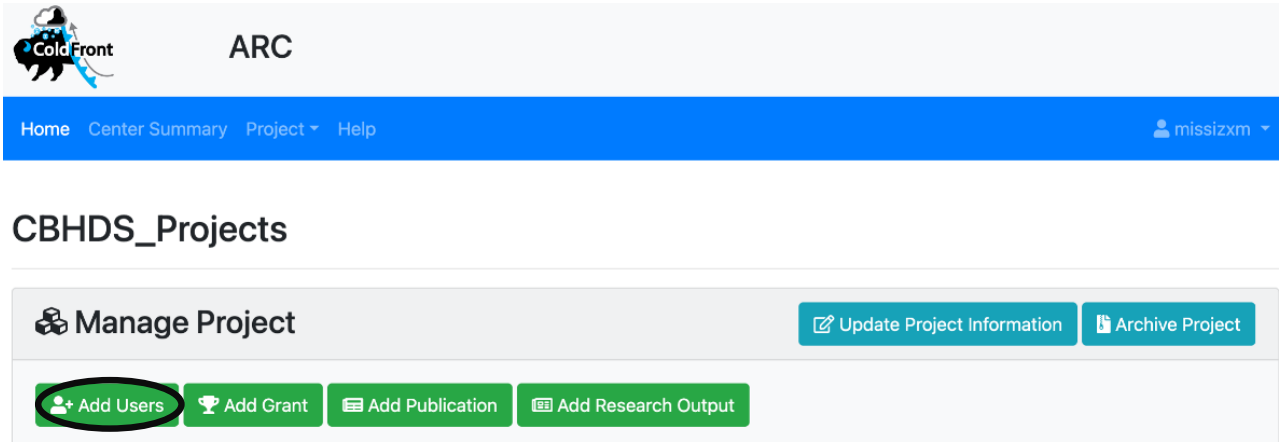


1.3 Set Up New Allocations

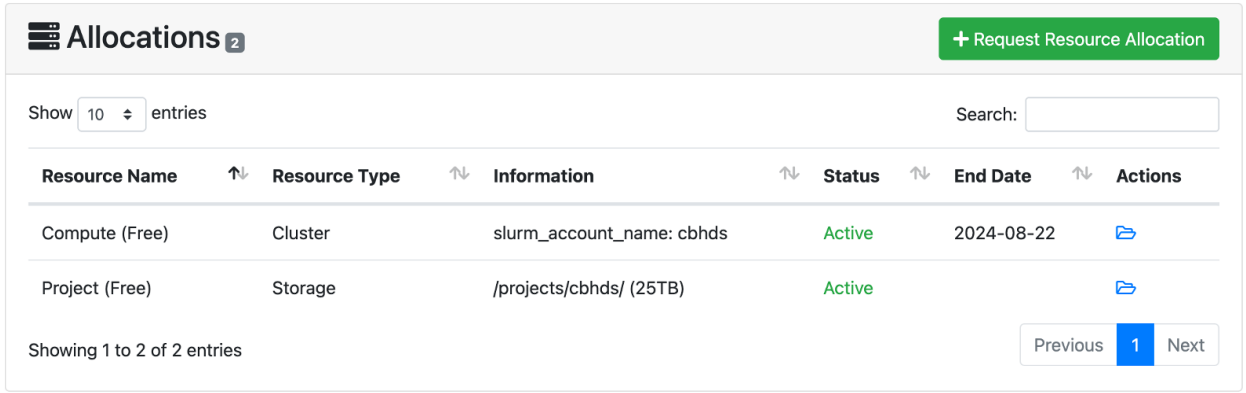
1) log into ColdFront using your ARC account username and password.

2) from the home page, select a project.

3) click “add users” and then you can search and add users using their email addresses.



4) click the “request resource allocation” and select the allocation type. Then fill out the form (on the same web page) and submit the request. Storage allocation is for project storage space, and compute allocation is for computation resources (time & power). ARC provides 0.6 million core-hour/month and 25TB storage for free to all PI. Extra storage space and computation time can be granted upon requests at a cost.



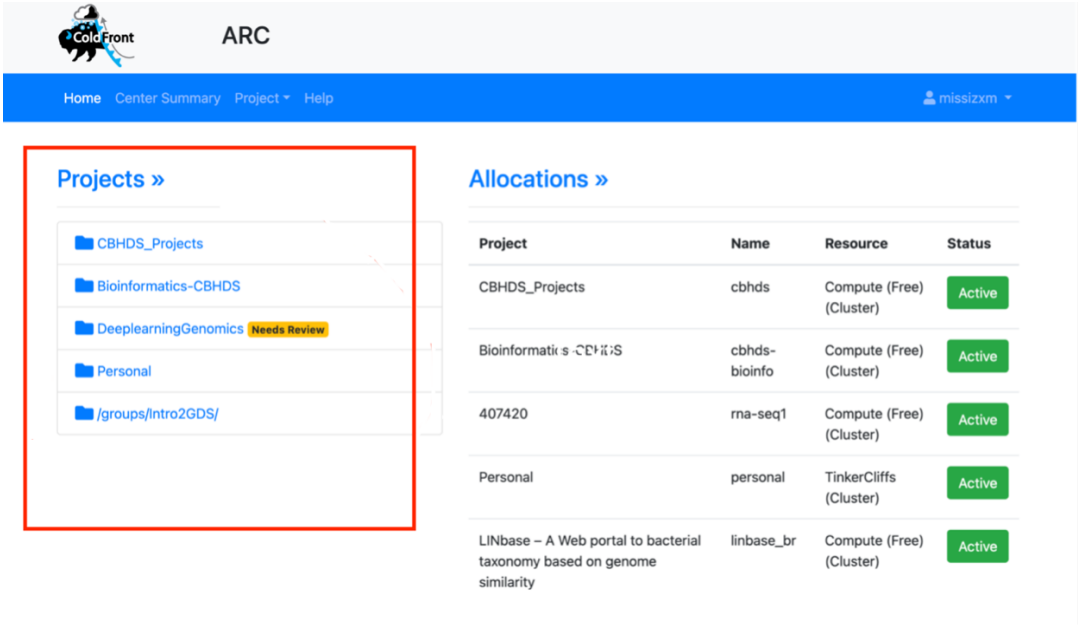
2. Getting Access to CBHDS Project Directories on ARC

1) Please download Virginia Tech's VPN if you are not connected to the campus network (eduroam). Instructions for VPN downloading, installing, and connecting are provided here.
    a. For Mac users: https://vt4help.service-now.com/sp?id=kb_article&sys_id=e08ee7361b7f3414688b2f82604bcbac
    b. For Windows users: https://vt4help.service-now.com/sp?id=kb_article&sys_id=d5496fca0f8b4200d3254b9ce1050ee5
2) Once you are connected to the VT network, please go to: https://ood.arc.vt.edu/pun/sys/dashboard/files/fs/projects/, where you will find a list of all the project directories on ARC.
3) Scroll down and you'll find 5 project directories related to CBHDS (see screen below with specific project names).

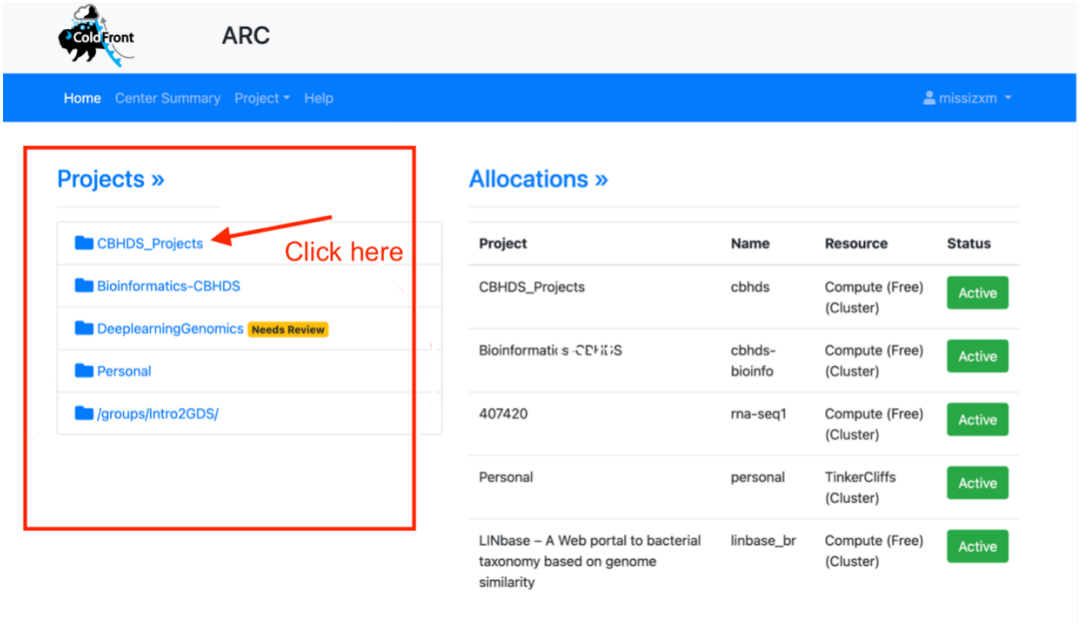
3. Manage ARC projects.

1) To manage ARC accounts and project directories (folders), please go: https://coldfront.arc.vt.edu/

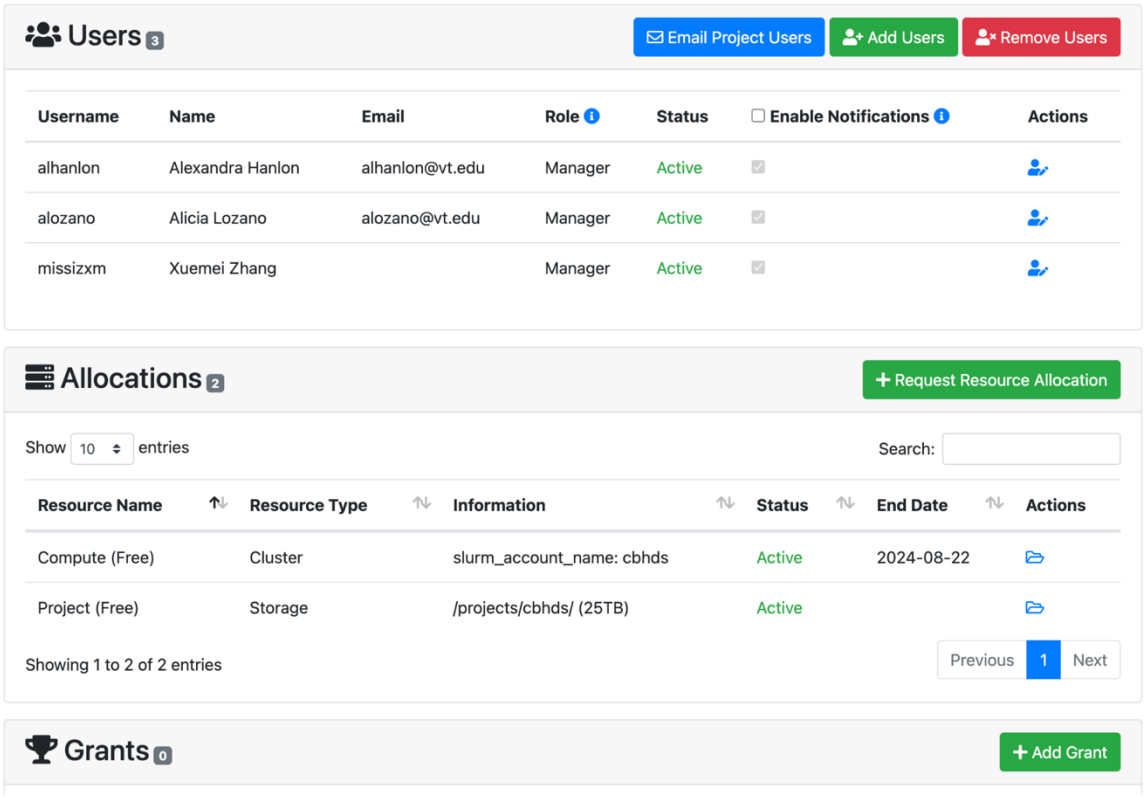
2) From the home page, you'll see a project list on the left (see screenshot below). All projects provided to you in this list are the ARC projects that you currently have access to.



3) To find the specific information of a project, click that project:



4) Once you enter the project page, scroll down to the “Allocations” section:



5) The third column labeled “Information” contains **the path of your project directory**:

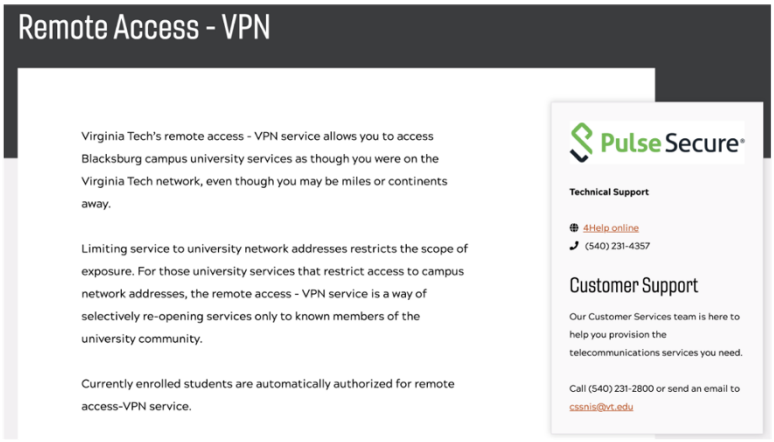
For example, “/projects/cbhds/”. “cbhds” is the project directory name.

6) Now, go to: https://ood.arc.vt.edu/pun/sys/dashboard/files/fs/projects/

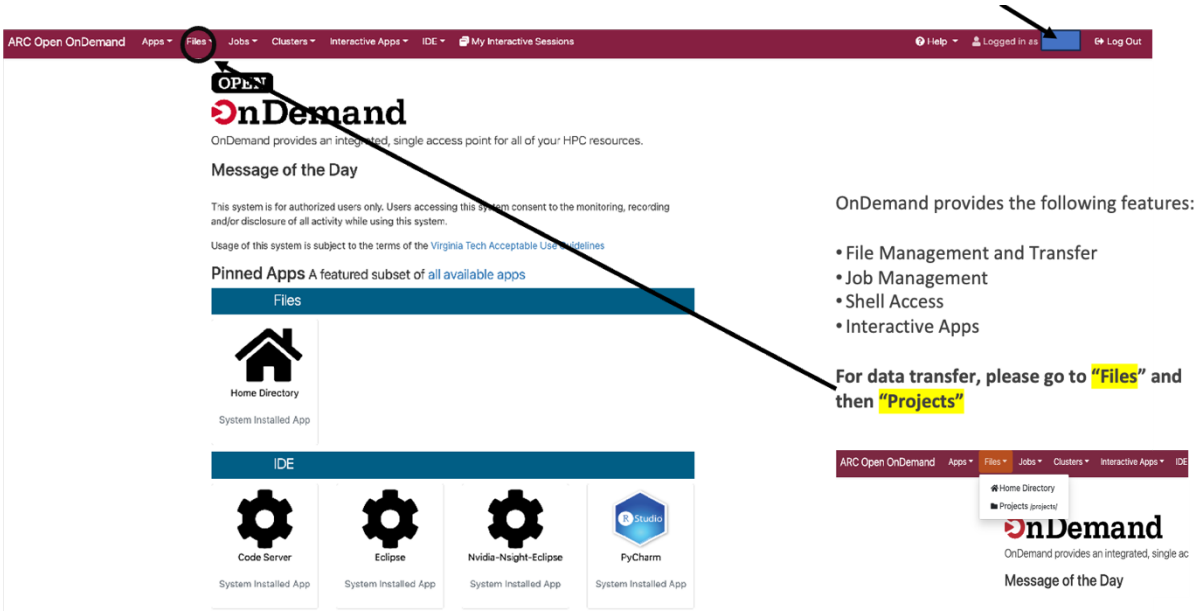
You will find the project directory name there.

4. Data Transfer and Management

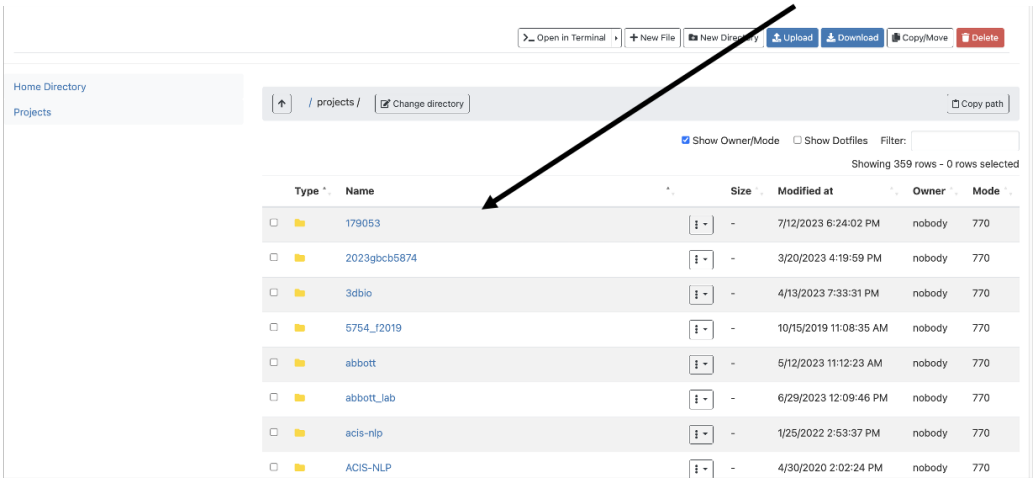
1) Go to https://ood.arc.vt.edu/pun/sys/dashboard. If you are not using campus network, please use PulseSecure. The instructions can be found here.



2) Once you log in using your Virgina Tech credentials (PID and password), you will see your username at the top right corner of the page (see in blue in screenshot below):



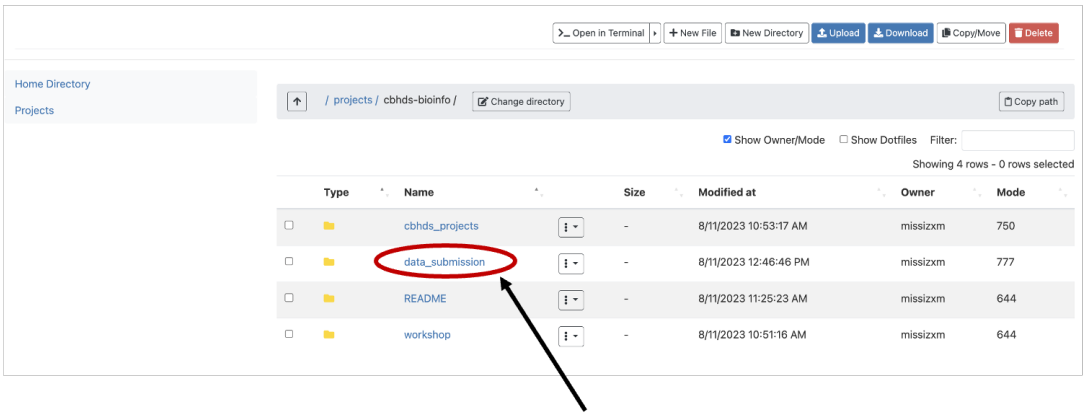
3) Once you enter the project folder (directory), you will see a list of project directories. To submit data to CBHDS bioinformatics, please go to “cbhds-bioinfo”.



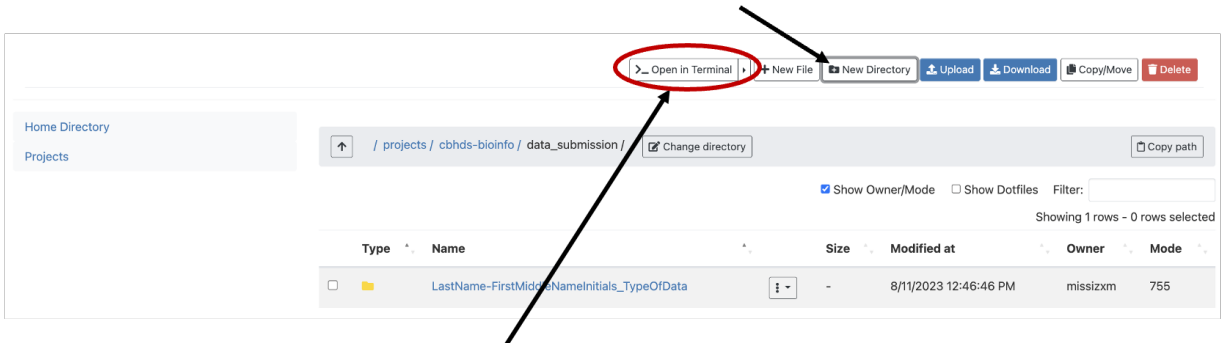
Now, please scroll down and find “cbhds-bioinfo”

4) Once you enter the “chbds-bioinfo” directory, you will see a few directories:
   - “README” contains data transfer and storage instructions, and overviews of different data analysis workflows
   - “workshop” contains all the workshop materials

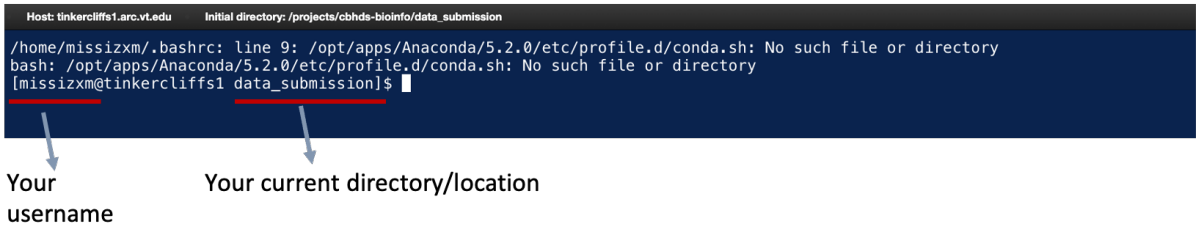
If you need to submit your raw data, please go to “data_submission”.



5) Once you enter the "data_submission" directory, you can click "new directory" to create a new directory. Please name this new directory use the format of last name and first & middle name initials, and the data type. For example: "Zhang-XM_RNASEQ"



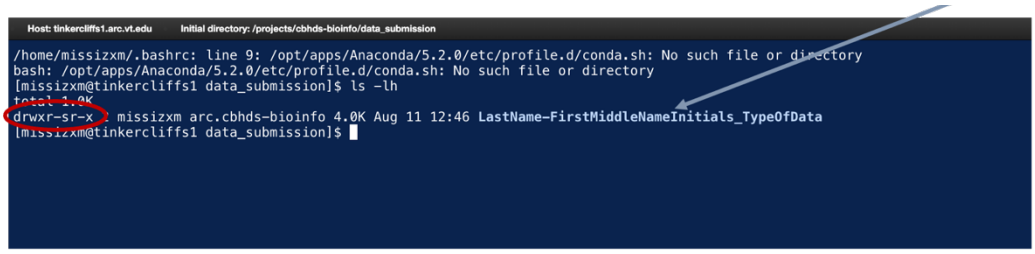
6) Next, please click **">_Open in Terminal".** Then you will see the Terminal window like this (see screenshot below):



Your username

Your current directory/location

7) In the Terminal Window, type “ls (space)-lh” and hit the “enter/return” key on your keyboard, and you should see your new directory in the list



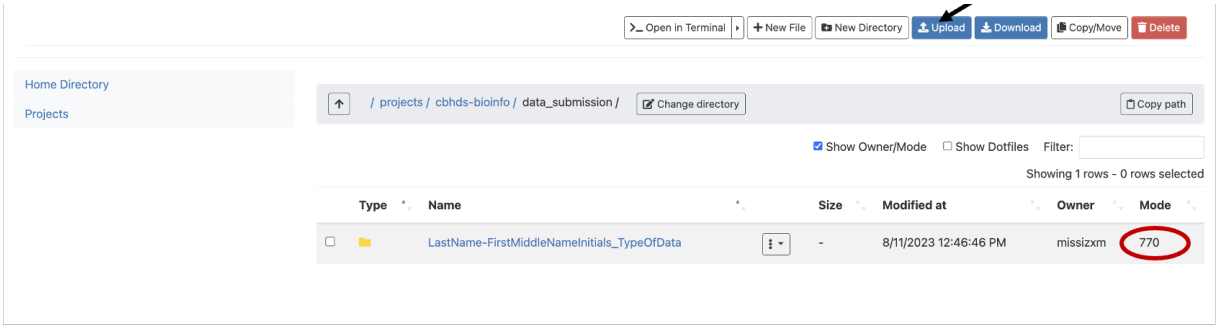
8) In the same Terminal Window, type “chmod (space) u=rwx,g=rwx,o-rwx (space)DirectoryName ”. Hit “enter/return”.

Type “ls (space)-lh” and hit the “enter/return”

Notice the output in the first column became “drawxrrws---” from “drwxr-sr-x”?

Now, everything in this new directory is only visible and accessible to you and our CBHDS bioinformatics team.

9) Go back to the Open On Demand ARC webpage to confirm that the mode of your directory is “770”. Now, click on the Upload Box in blue (see black arrow in screenshot below) to upload your data.



Note that It is highly recommended to change the directory permission (Step 7-9) before uploading your data. Changing the permission to “drwxr-sr-x” (“770” permission mode) will ensure that your data are only accessible to you and our CBHDS bioinformatics team.

If you have trouble changing the directory permission mode to “770”, or with any of the instructions provided in this document, please email Missi Zhang at missizxm@vt.edu.