Things you need if you are using windows:

1. Software called 'putty' to connect to Franklin cluster and type in the commands below. X11 forwarding needs to be enabled in order to display the venn diagram. See **Useful commands** and tips.
2. Software called 'winscp' to transfer files to Franklin cluster
3. Software called 'xming' to display x11 program. This is needed to display venn diagram.

There several different analyses you can run:

1. Trim low quality reads and adapter sequences then run quality control (`run_trim_qc.sh`)
2. Align trimmed reads to hg19 or hg38 or other reference genome and create tracks for visualization (`run_align_create_tracks_rna.sh`). Although, this can be run independently of #1, it requires all the files to be in the same path as if #1 has already been run.
3. Count reads in each feature and then performed differential analysis (`run_differential_analysis_rna.sh`). Again, this can be run independently but requires all files to be in their locations as output by the pipeline.
4. Instead of running each of the above steps separately, you can run the full RNA-seq analysis (#1-3) using `run_rnaseq_full.sh`
5. After running differential analysis, you can view a proportional Venn diagram of differential genes overlap between different groups (`run_overlap.sh`)

Notes:

- This pipeline only accepts fastq.gz files.
- User can run_create_tracks with only 1 replicate but to run full RNA-seq analysis, you need at least two replicates.
- Before running any of these two options, you need to prepare the directory and samples information file called samples.txt is required. More detailed steps are listed below.
- You can take a look at `/apps/opt/rnaseq-pipeline/scripts/project_ex` for examples of input files (i.e. samples.txt and fastq directory). `project_ex` folder also contains examples of output files upon completion of running RNAseq full analysis.
- You can copy the entire `project_ex` folder to your own folder, delete outputs folder and all the *.out files then run this as test files.
- In this document, there are also descriptions of output files.

Preparing directory and samples information

1. You need to have a project directory. Please avoid spaces in the directory name.
2. Inside the project directory, you must have a directory called "`fastq`". All fastq files for all replicates/samples should be in this directory. Inside fastq directory, you can have

subdirectories which then contain the fastq.gz files. See "`project_ex`" directory in `/apps/opt/rnaseq-pipeline/` for example of a project directory.
Inside the project directory, you must have a "`samples.txt`" file which contains your samples information. You can copy the "`samples.txt`" from "`project_ex`", open it with excel and change the text inside with your own samples' information. Do **NOT** save the samples.txt as xlsx file then convert it back to txt. Doing so will add extra invisible characters that are not compatible in Unix. When you're done, save it back as text file (tab delimited) with the same name and transfer it to your own project directory. Do **NOT** put empty line(s) in your "`samples.txt`" in between your samples information.

Inside "`samples.txt`":

a. **First column: "Groupname"** – type in the group name of the samples. Samples with replicates must have the same "Groupname". This name must be part of the fastq filename. This column cannot be empty.

Naming Rules:

1. **Do NOT** use punctuation characters in "Groupname". Punctuation characters are:
   !"#$%&'()*+,-./:;<=>?@[]^_`{|}~
   It's fine for the file name to have punctuation characters. Only part of the file name that is used as "Groupname" that need to have the characters removed (not folder name). Then match "Groupname" to the fastq file names.
2. **Do NOT** use a "Groupname" that starts with a number.
3. Different groups **cannot have** "Groupname" that is a complete part of another group. (i.e two "Groupname": "EFKD" and "EF" is not allowed as "EF" is a substring of "EFKD"). See example 3 below.

Examples of how to pick a "Groupname" :

- **Example 1:** if the file name is `iLuc-197_rep1_R1.fastq.gz`, remove the punctuation character "-" from the fastq file names so new file name becomes `iLuc197_rep1_R1.fastq.gz` then set "Groupname" to be iLuc197.
- **Example 2**: if file names are `iEF_714_FLAG_S3_rep1.fastq.gz` and `iEF-714_FLAG_S3_rep2.fastq.gz` which are two biological replicates, both need to have the same name in "Groupname". For these two files, you need to:
  i. Remove punctuation character "_" from part of their file names that are going to be used for "Groupname" so they become `iEF714_FLAG_S3_rep1.fastq.gz` and `iEF714_FLAG_S3_rep2.fastq.gz`.
  i. Have 2 lines in the `samples.txt` with iEF714 as "Groupname" and "rep1" and "rep2" under "Replicatename". See more information on 'Third column: "replicatename" ' below.

- **Example 3:** If file names are iEF_rep1_R1.fastq.gz and iEF714_rep2_R2.fastq.gz, remember that "iEF" and "iEF714" cannot be your "Groupname" as "iEF" is part of "iEF714". In this case, you can change iEF_rep1_R1.fastq.gz to iEFempty_rep1_R1.fastq.gz. That way, you can then have "iEFempty" and "iEF714" as their "Groupname". With this change, you no longer have a "Groupname" that is completely part of another group.
- **Example 4**: if file name is `0647_714R1_S57_L001_R1_001.sub.fastq.gz`
    1. Change file name to `0647_iEF714R1_S57_L001_R1_001.sub.fastq.gz`
    2. Have "Groupname" as iEF714.

b. **Second column: "Controlname"** - type in the name of the control/ background/reference sample. During differential analysis, the corresponding sample will be compared to its "Controlname". Fold-change will be calculated with "Controlname" as reference. "Controlname" has to be one of "Groupname". This name must be part of the fastq filename. This column cannot be empty. For control/background sample, please put NA for "Controlname".

The same naming rules for apply:

1. **Do NOT** use punctuation characters in "Groupname". Punctuation characters are:
   !"#$%&'()*+,-./:;<=>?@[]^_`{|}~
   It's fine for the file name to have punctuation characters. Only part of the file name that is used as "Controlname" that need to have the characters removed (not folder name). Then match "Controlname" to the fastq file names.
2. **Do NOT** use a "Controlname" that starts with a number.

Examples of how to pick a "Controlname":

- **Example 1**: If you have `iEF714_rep1_R1.fastq.gz` that need to be compared to `iEFempty_rep1_R1.fastq.gz` as a reference. For these two files you will have two entries in the samples.txt:
    1. one entry with "Groupname" as "iEF714" and "iEFempty" as "Controlname". This corresponds to your `iEF714_rep1_R1.fastq.gz`
    2. Another line with "Groupname" as "iEFempty" and "Controlname" as "NA".

c. **Third column: "Replicatename"** – type in the name of the replicates. This name must be part of fastq filename. This column cannot be empty. Sometimes the replicates are denoted as R1 or rep1. Make sure the correct replicate notation is typed in the samples.txt. Make sure "**Replicatename**" is not exactly the same as "**string_pair1**" and "**string_pair2**"

Example of how to pick a "Replicatename":

- **Example 1:** If you have `A8_20-0647_iEF714R1_S57_L001_R1.fastq.gz` and `A8_20-0647_iEF714R2_S57_L001_R1.fastq.gz` then "Replicatename" will be "R1" and "R2". Make sure when you fill in "string_pair1" it is "_R1" and **not** "R1"
- **Example 2:** If you have `A8_20-0647_iEF714Rep1_S57_L001_R1.sub.fastq.gz` and `A8_20-0647_iEF714Rep2_S57_L001_R1.sub.fastq.gz` then "Replicatename" will be "Rep1" and "Rep2". It is case sensitive.

d. **Fourth column: "spikename"** – type in NA for all rows.
e. **Fifth column: "email"** – type in the email address of the person running the analysis. Slurm job emails will be sent this email address.
f. **Sixth column**: "**string_to_identify_sample**" – this string will be used to identify files that below to a particular sample.
This could have punctuation characters.

Examples of how to pick a "string_to_identify_sample":

- **Example 1**: if your file names are:
  - `A8_20-0647_iEF714R1_S57_L001_R1_001.sub.fastq.gz`
  - `C8_20-0649_iEF563R1_S59_L001_R2_001.sub.fastq.gz`
  - `H7_20-0646_iEF197R1_S56_L002_R2_001.sub.fastq.gz`
  - `A9_20-0655_iEF197R2_S65_L001_R2_001.sub.fastq.gz`

  The "**string_to_identify_sample**" would be iEF714R1, iEF563R1, iEF197R1 and iEF197R2 respectively.

g. **Seventh column**: "**string_pair1**" – each fastq file will be a pair of file that only differs in one string. Usually this string is _R1_ for pair1 reads and _R2_ for pair2 reads.
h. **Eight column**: "**string_pair2**" – similar to "**string_pair1**" but for pair 2 reads.

Examples of how to pick "string_pair1" and "string_pair2":

- **Example 1**: If your file names are:
  - `A9_20-0655_iEF197R2_S65_L001_R1_001.sub.fastq.gz`
  - `A9_20-0655_iEF197R2_S65_L001_R2_001.sub.fastq.gz`

  Therefore the "string_pair1" would be "_R1_" and "string_pair2" would be "_R2_"

- **Example 2**: If your file names are:

- A9_20-0655_iEF197R2_S65_L001_R1.fastq.gz
- A9_20-0655_iEF197R2_S65_L001_R2.fastq.gz

Therefore the "string_pair1" would be "_R1" and "string_pair2" would be "_R2"

How to

Trim adapter sequences and low quality reads then run quality control (QC) without running full analysis.

After preparing the directory and samples information as detailed above, to trim adapter sequences and low quality reads then run QC without running the full analysis:

1. Go into the project directory.
    1. For example: if your project directory is called "my_project" then type in:
    ```
    cd my_project
    ```
2. To trim and run QC only, type in:
    ```
    bash /apps/opt/rnaseq-
    pipeline/scripts/project_ex/scripts/run_trim_qc.sh &>
    run_trim_qc.out &
    ```

    Check "run_trim_qc.out" file for progress and messages. **Do Not** change this output filename.

    Type in help to get more information:
    ```
    bash /apps/opt/rnaseq-
    pipeline/scripts/project_ex/scripts/run_trim_qc.sh help
    ```

Align reads to reference genome hg19 (or hg38 or other reference genome) and create normalized tracks for visualization without running the full analysis.

After preparing the directory and samples information as detailed above, to align and create tracks without running the full analysis:

Note: everything needs to be in the same directories as if run_trim_qc.sh had been run.

1. Go into the project directory.

    For example: if your project directory is called "my_project" then type in:
    ```
    cd my_project
    ```

2. To align to **hg19** and create tracks only, type in:

```
bash /apps/opt/rnaseq-
pipeline/scripts/run_align_create_tracks_rna.sh &>
run_align_create_tracks_rna.out &
```

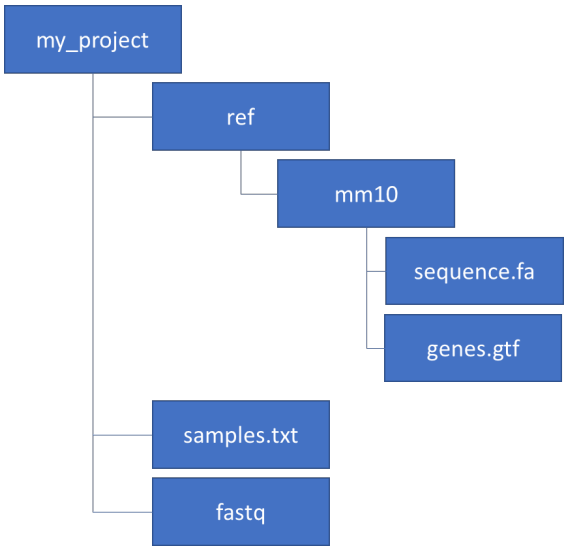
Check "run_align_create_tracks_rna.out" file for progress and messages. **Do Not** change this output filename.

3. Or to align to **hg38** and create tracks only, type in:

```
bash /apps/opt/rnaseq-
pipeline/scripts/run_align_create_tracks_rna.sh genome=hg38 &>
run_align_create_tracks_rna.out &
```

Check "run_align_create_tracks_rna.out" file for progress and messages. **Do Not** change this output filename.

4. Or to align to **other reference genome** and create tracks only:
Place reference genome sequence file (.fa or .fasta) and genome annotation (.gtf) inside a directory with the genome name. Place this directory under a directory named "ref", inside your project directory.
For example, to align to genome=mm10, the following would be the file structure:



Then type in:

```
bash /apps/opt/rnaseq-
pipeline/scripts/run_align_create_tracks_rna.sh genome=mm10 &>
run_align_create_tracks_rna.out &
```

Check "run_align_create_tracks_rna.out" file for progress and messages. **Do Not** change this output filename.

5. Type in help to get more information:

```
bash /apps/opt/rnaseq-pipeline/scripts/project_ex/scripts/
run_align_create_tracks_rna.sh help
```

Run feature count and differential analysis without running the full analysis.

After preparing the directory and samples information as detailed above, to count features and run differential analysis without running the full analysis:

Note: everything needs to be in the same directories as if run_align_create_tracks_rna.sh had been run.

1. Go into the project directory.

   For example: if your project directory is called "`my_project`" then type in:

```
cd my_project
```

2. To count features and run differential analysis, type in:

```
bash /apps/opt/rnaseq-
pipeline/scripts/run_differential_analysis_rna.sh &>
run_differential_analysis_rna.out &
```

   Check "`run_differential_analysis_rna.out`" file for progress and messages. **Do Not** change this output filename.

3. To get more information, type in:

```
bash /apps/opt/rnaseq-
pipeline/scripts/run_differential_analysis_rna.sh help
```

- ## Run full RNA-seq analysis and obtain lists of differential genes (commonly used)

After preparing the directory and samples information as detailed above, to run the full analysis which includes trim, qc, alignment, tracks creation, feature counting and differential analysis:

1. Go into the project directory.

   For example: if your project directory is called "`my_project`" then type in:

```
cd my_project
```

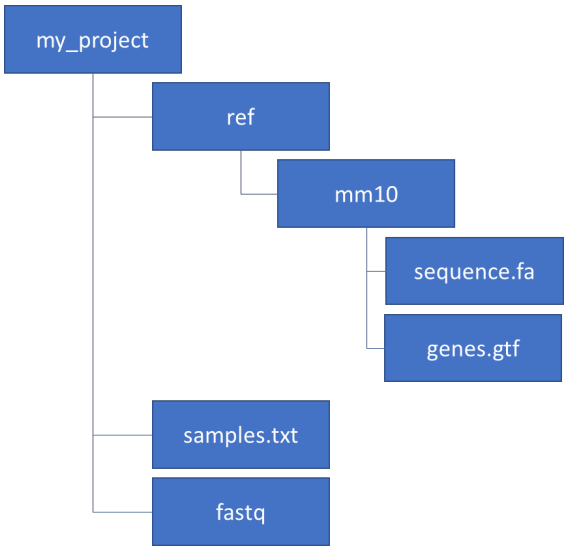
2. To align to **hg19** and run full RNA-seq analysis, type in:

```
bash /apps/opt/rnaseq-pipeline/scripts/run_rnaseq_full.sh &>
run_rnaseq_full.out &
```

3. Or to align to **hg38** and run full RNA-seq analysis, type in:

```
bash /apps/opt/rnaseq-pipeline/scripts/run_rnaseq_full.sh
genome=hg38 &> run_rnaseq_full.out &
```

4. Or to **align to other reference** genome and run full RNA-seq analysis, place reference genome sequence file (.fa or .fasta) and genome annotation (.gtf) inside a directory with the genome name. Place this directory under a directory named "ref", inside your project directory.
   For example, to align to genome=mm10, the following would be the file structure:



   Then type in:

```
bash /apps/opt/rnaseq-pipeline/scripts/run_rnaseq_full.sh
genome=mm10 &> run_rnaseq_full.out &
```

5. Or to specify time limit of 1 day, type in:

```
bash /apps/opt/rnaseq-pipeline/scripts/run_rnaseq_full.sh time=1-
00:00:00 &> run_rnaseq_full.out &
```

   Use this command only if you really need the time limit. Any commands described above, can also be used with the time option. This command is usually used for times when the cluster is scheduled for maintenance within a certain time limit. **Note**: your jobs will be killed if they are not finished within the time limit. The format for time is `time=DD-HH:MM:SS`

6. Or to get more information, type in:

```
bash /apps/opt/rnaseq-pipeline/scripts/run_rnaseq_full.sh help
```

   Check "**`run_rnaseq_full.out`**" file for progress and messages. **Do Not** change this output filename.

Output folders:

In case of errors, you should first read `*.out` in your project folder. For example, if you are running `run_rnaseq_full.sh`, `run_rnaseq_full.out` will contain message about the error and where to look for more information.

After running full RNA-seq analysis, you will have an **outputs folder** with all the results of running the analysis. Inside outputs**, diff_analysis_rslt** folder will have the results of differential analysis. **RNA-seq differential_analysis_report.html** which contains detailed report of the entire analysis is a good place to start reading.

bw_files will have the normalized tracks for visualization

fastqc_rslt will have the quality control reports

logs will have logs of programs ran. Details of programs ran will be here

STAR_2pass will have the alignment files produced by running two-pass STAR. The **final bam** files are in "**STAR_2pass/Pass2**". Feature counts used as input for differential analysis are also stored here.

trim will have the trimmed fastq files.

View interactive area-proportional Venn Diagram.

Once you have run differential analysis, you can run Venn diagram app. The app will allow user to change FDR, fold-change, labels, etc. Run the app using the following steps:

1. Go into the project directory.

   For example: if your project directory is called "`my_project`" then type in:

   ```
   cd my_project
   ```

2. To generate Venn Diagram, type in:

   ```
   bash /apps/opt/rnaseq-pipeline/scripts/run_overlap.sh
   ```

   Note: There is no '&' at the end of the command!!

   This will load shiny app, then it will ask for your Franklin cluster's password. After you typed in the password, it will take a while before a firefox with interactive Venn diagram appear.

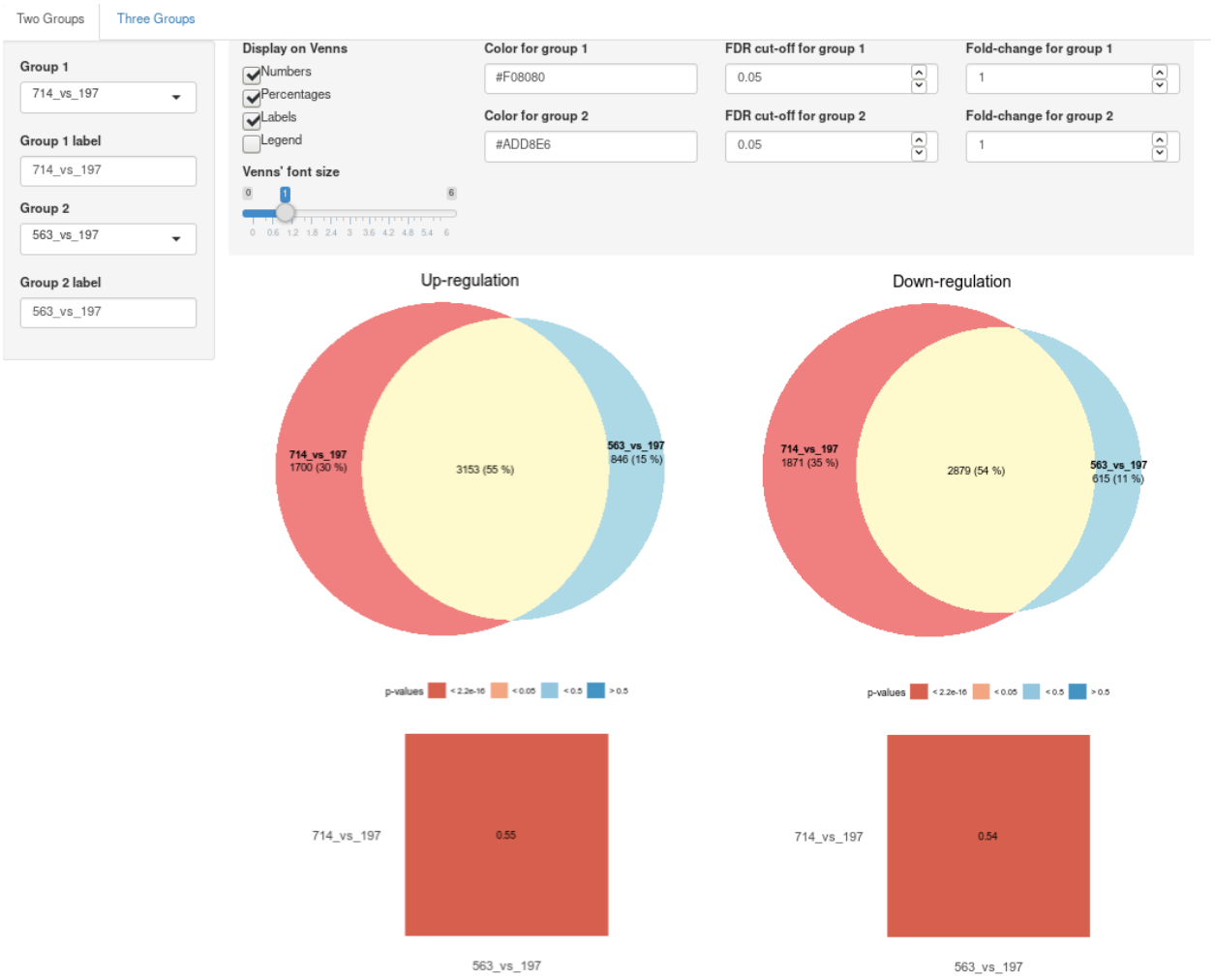
Please wait patiently ☺, it may take a while for firefox to appear. After firefox appears, it may take some time for the page to be loaded. During this time, firefox will be unresponsive. After loading, if firefox is slow, simply exit out of firefox and restart `run_overlap.sh`.

If it fails with the following error: “connection to … closed by remote host” simply try to re-run the above command.

If a window appears asking you to create new profile, please click on “create new profile”.

Note: you will need to have an x11 display server such as Xming running on your computer (see Useful commands and tips).

Screenshots of the venn diagram app:



Overlap analysis with more than 3 groups:

Useful commands and tips

- To see where you are currently, type:
  `pwd`
- To change directory type:
  `cd directory_path`

For example:

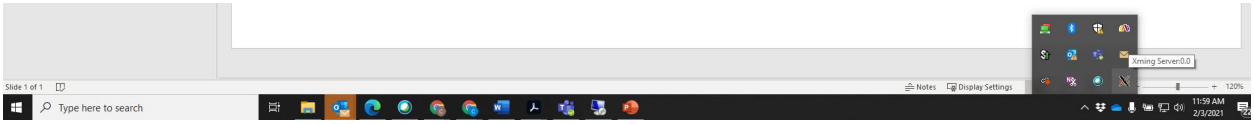
```
[cxt050@r1pl-hpcf-log01 ~]$ pwd
/gpfs0/home/gdlessnicklab/cxt050
```

Typing '`pwd`' let me know that I am currently in "`/gpfs0/home/gdlessnicklab/cxt050`"

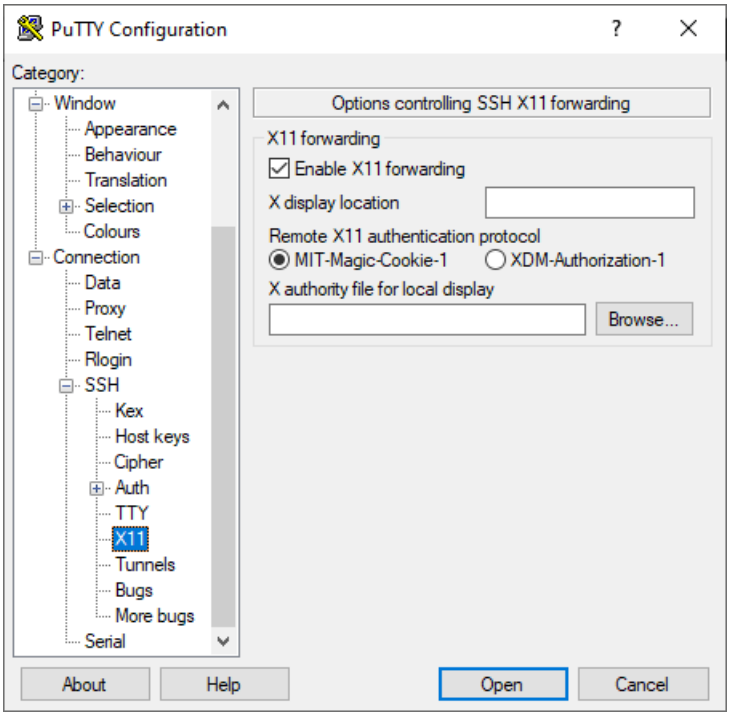
To go into '`/home/gdlessnicklab/share/data/`', type:
"`cd /home/gdlessnicklab/share/data/`"

```
[cxt050@r1pl-hpcf-log01 ~]$ cd /home/gdlessnicklab/share/data/
[cxt050@r1pl-hpcf-log01 data]$ pwd
```

```
/home/gdlessnicklab/share/data
```



If you are using putty, make sure x11 is also checked (enabled):



Have questions? Email cenny.taslim@nationwidechildrens.org