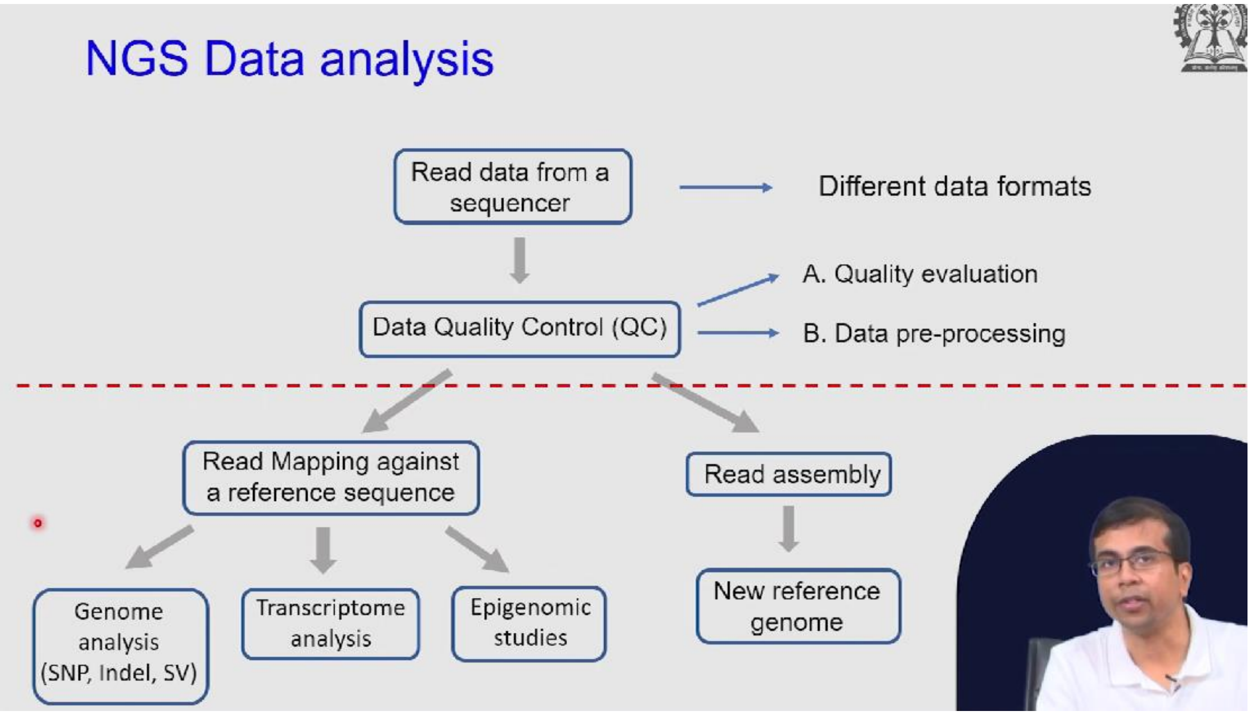
Next Generation Sequencing Technologies: Data Analysis and Applications

Hands-on Setting up the system

Dr. Riddhiman Dhar, **Department of Biotechnology**

Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the course on Next Generation Sequencing Technology Data Analysis and Applications. So far, we have discussed data formats, we have discussed data quality control, or QC, and we have discussed data preprocessing. So, in this class and the next few classes, we will have some hands-on exercise. So, I will show you how we actually do this analysis. So, this is something you can do along with me, and you will learn yourself how to analyze this data and get a feel for the data. So, all these steps we will do one by one. So, in this class, we will first set up the system where we can do the data analysis, and then we will discuss some basic Linux commands. So, what I realize is that many of you might be familiar with the Windows system, but you might not be familiar with Linux, Mac, or unique space systems. So, this is something that will give you an introduction to Linux commands, ok? So, this is very important because majority of the tools with few exceptions major majority of the tools they work on Linux right. They need this command-line interface, which only any unique space operating system can provide. So, these are the keywords that we will see: uniques, Windows subsystem for Linux, and shell commands. So, this is the flow chart for NGS data analysis, right? So, you have the data that you get from the sequencer, and you have different data formats.

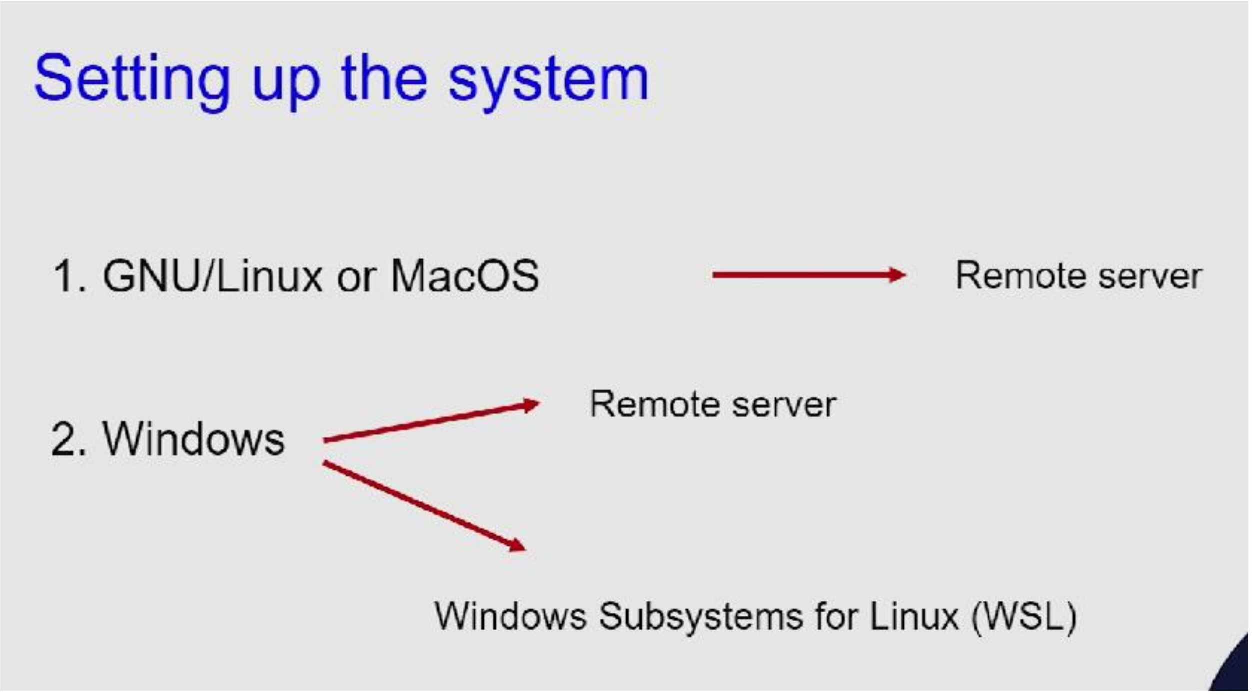


NGS Data analysis

This is something you have seen before, and I have shown you before. The next step is data quality control. Now, once you get the data, where do you store the data, and how do you actually look at the data? And then you start processing this data along these lines. So, for data quality control, we have talked about two steps: quality evaluation and data preprocessing. So, this is up to this point we have discussed so far, and in this class, in the next few classes, we will cover up to this point. We will set up the system, we will look at the data, and we will do the quality control and preprocessing. So, you will learn how to do this with your own data. So, these are the five steps that will be covered in this class and the next few classes. So, first, we will set up the system for doing the data analysis. We will talk about basic shell commands. These are the commands that are heavily used in any unique Linux space system.

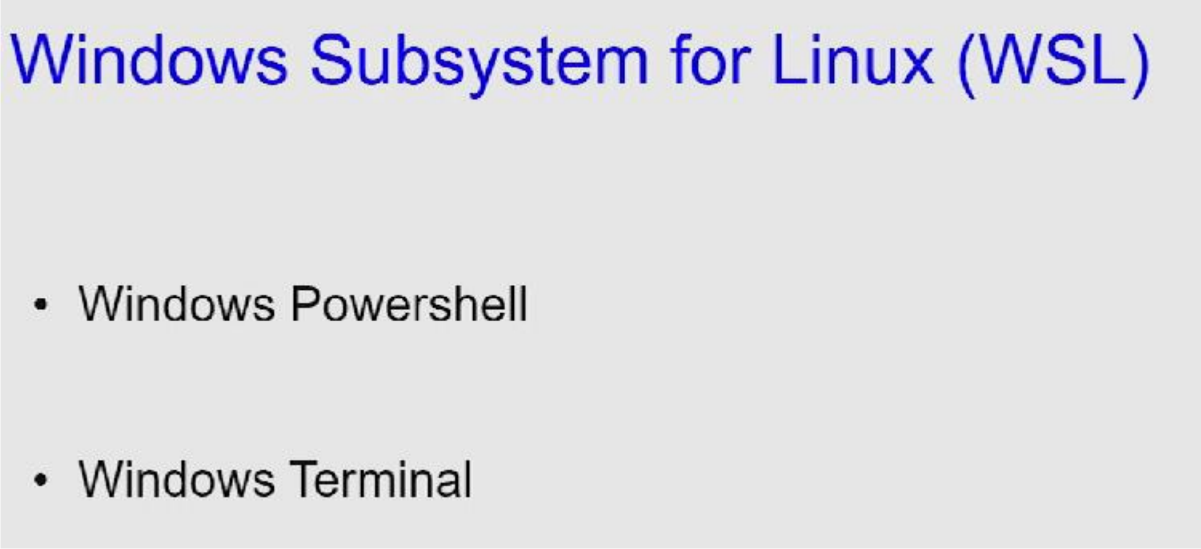
Then, in the third part, the actual data will come out. So, where we get the data from any sequencing data, we will work with a real data set. We will download the data, and I will show you where you can download the data. Even if you do not have your own data, you can download publicly available data sets, and you can start doing this kind of analysis. This is how you would learn much better than just learning the theory. We will explore the data; we will see what the data looks like, whether there are any sort of insights that we can get just by exploring the data, and how we can actually explore these big data sets. So, they require some special ways of navigating.

Then we have the fourth point, which is that quality control will actually use a quality control tool to actually apply to the data that we have downloaded or your own data, and we will actually get those statistics that we have talked about in the theory class. And finally, if we will do data trimming, if it is required for the data, it may or may not be required, but if it is required, how do we actually go about it? So, this is something that will be on the agenda. So, the first thing is setting up the system right. So, if you are working on a Linux or Mac OS system, then you are ready to go. You have the terminals on Linux and Mac, and you can work on that, or, if you might be, I guess most of you are working on Windows. So, this is where you have to get yourself a little bit familiar with the Linux-based system. So, the good thing is that now you can run Linux inside Windows. You do not have to have dual boot systems; you can install Linux inside Windows, and I will talk about this in this class. So, you can work on these two systems, your laptop or desktop, or sometimes we analyze these big data sets because they often take a lot of memory, RAM, and storage space. So, that might not be available with your system.



Setting up the system

So, we set up remote servers. These are powerful computers that can analyze where you can do this analysis, and they can actually provide you with a very high amount of RAM. For example, if you need 50 GB of RAM, this is available with servers. So, we can connect to these remote servers, and we can actually do this analysis as well. Now again, on the remote server, we have this Linux-based system. So, this will be there, and we have to know the basic shell commands.

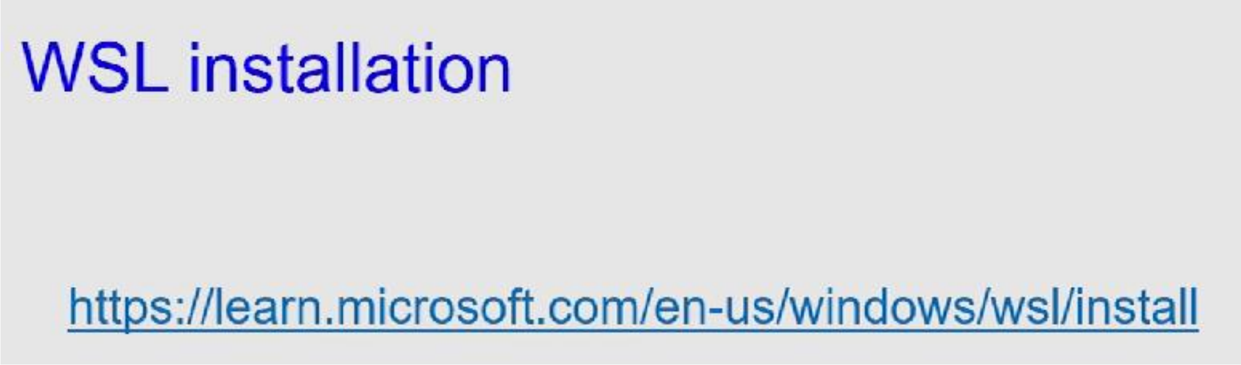


Windows Subsystem for Linux (WSL)

- Windows Powershell
- Windows Terminal

In Windows, we can connect to a remote server and do all the analysis. This is a preferred way if you are dealing with a lot of data, but we can also run this inside our system on our computer with something called Windows subsystems for Linux. This is a Linux that was developed by Microsoft and runs within Windows. In short, it is called WSL. So, this is something I will talk about for those who have a Windows computer, desktop, or laptop and who would like to do this analysis on their desktop. So, for running WSL right, you need something called Windows Power Shell or Windows Terminal, ok? So, I will show you in a moment what this power shell or terminal would look like, and then you can go on and install WSL on Windows, ok? So, here it is, right? So, here is the power shell. So, if you go to your Microsoft Store, you can actually search, for example, for Windows Terminal, and you can simply check this gate, and this will be installed in your system. So, you can also use a power shell. So, these are two options that you have.

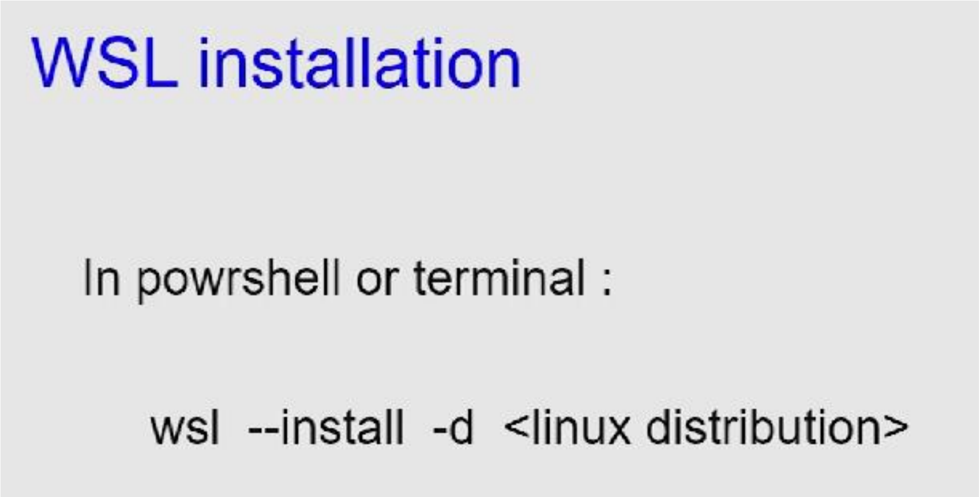
So, I have this installed. So, here, I will just open this so that I can show you right southward, ok? So, this has started, and what I will do is actually go into the shell, which is called bash. So, this actually takes me to the shell, and then I will talk about how this is comparable to the terminal that you have in Linux. Once you have once you are here, right so, this shell actually would not be there.



WSL installation

https://learn.microsoft.com/en-us/windows/wsl/install

So, once you are here, you can type something called WSL. Maybe what I will do is I will kind of do so, I mean WSL minus minus ok. So, this is something you would have to type, and this is something that is actually also given right. So, coming to the next slide, right, you can actually take two parts, and I will discuss them in a moment about how we can do this. So, one is that you can directly install WSL by typing a command, and for this, you need certain versions of Windows. You can see the prerequisites by going to this link and you can read all the instructions, prerequisites, etcetera. So, if you are running Windows 10 in certain versions or Windows 11, you can do this directly, right? This is the simplest one, right? Or you can do something called ah.



WSL installation

In powrshell or terminal :

```
wsl  --install  -d  <linux distribution>
```

So, this is the simplest way of doing this. You can type WSL minus minus install, and you can simply type this minus the distribution of the Linux that you want. So, there are many such Linux distributions that will be available right now. So, you can have Ubuntu or other types of Linux, and you can choose whichever you like. So, for this class, I will be choosing mostly Ubuntu. So, I have this system installed, but I can also show you how to install it in WSL. So, you can specify the distribution, and this would be installed on a computer.

Manual installation

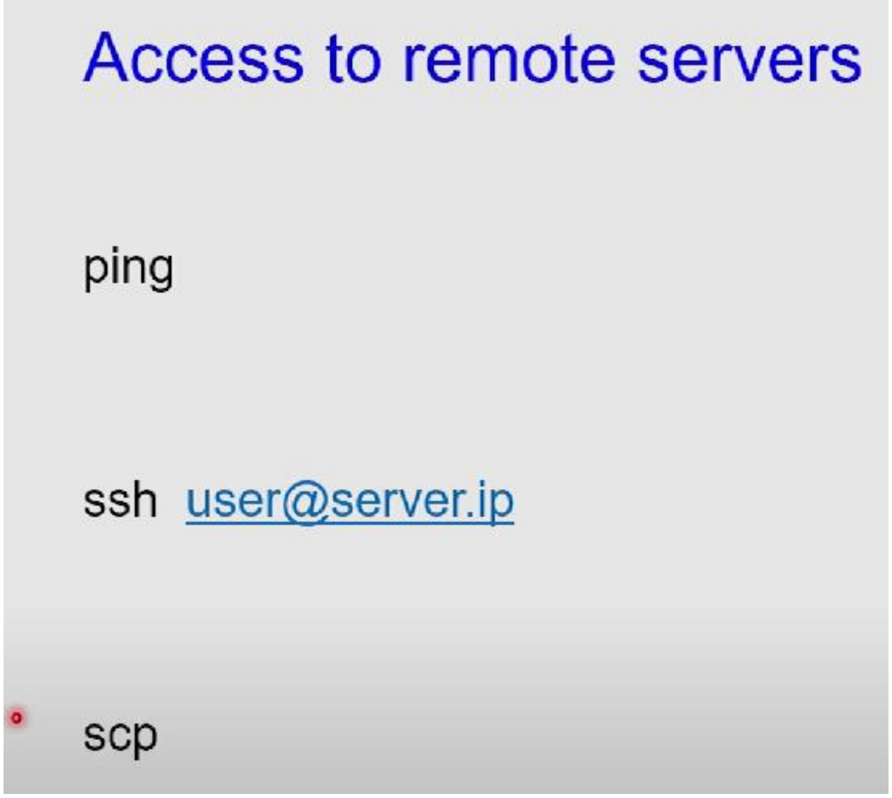
https://learn.microsoft.com/en-us/windows/wsl/install

The other way is for older versions, you can have to go through something called the manual installation step, and this will be again a step-by-step process. You have to first enable the Windows subsystem for Linux right in your system. Again, all the instructions are given here, and I will show you again in a moment where these instructions are. You have the first step, which is the command that you would have to run in the terminal. Then, in step 2, you have to check the requirements for running WSL version 2. So, whether you fulfill the requirements, Step 3: You would have to enable the virtual machine feature again. If you do not understand what all these steps mean, just follow the instructions that are given in this link. So, this is step 3 and step 4 that you can upload. You can now download the Linux kernel update package. So, that will actually give you the WSL 2 Linux kernel, and then you can set this WSL 2 as your default version, and finally, you can install your Linux distribution. So, this is something you see: there are a lot of steps and much more complexity, right? So, the first option is actually the best one. So, this is something I will show you in a moment, and if you want to go through the installation process right now, this is what you have to do.

So, you have these 2 steps: if you are doing the direct installation, this is what you will get. If you have this install command, you can simply type it in the terminal and you will get Ubuntu or any other Linux version installed, or you can follow these manual installation steps for older versions of WSL, and you have to follow these different steps again. You have to type these commands in the power shell or in the terminal. So, let us follow this right. Let us start with this one right. We can simply type WSL minus minus installed, and we can type the version right that we want. So, we said we want Ubuntu 22.04, right? Let us see what actually happens. So, once you type the distribution name properly, this will install the Ubuntu version. So, in my case, this is already installed, but this is what you would have to do, and as you can see, this is launching Ubuntu and it is in your system. If it is not installed, it will take some time, but it will show you that it is installing, configuring, etcetera. So, you need to wait a little bit, and then you will get this window where you will see that Ubuntu is installed. Once this is installed, you can actually come here in the terminal. You can directly work on that Ubuntu shell, but I will recommend that to come to this terminal because, as you know, you are kind of oriented better towards the file system of this computer. So, here you will see in these settings that Windows Ubuntu is actually enabled. So, we will run this here, right? We can simply run this, and you will get this Windows shell OK. The other option is something called bash OK. This is a shell where we run all the commands; this is equivalent to a terminal in a Linux or Mac system. So, this is what we require for running all the tools for data analysis, ok, and this is where we will build all this analysis. I will show you all the steps, etcetera, ok. Now, going back to the manual installation, you might want to try if, if you have old systems or older versions of the WSL, you just have to type these commands. You just have to copy these commands into the terminal, ok. So, you just come here and type these commands in the terminal before the bash, ok? So, ok. So, because the bash will come only when you have the Linux install, right? Ubuntu install, for example. So, once you have this, I will talk about one more step right before we actually go into all the details of the commands, shell commands, etcetera. So, as I said, sometimes we just want to access remote servers. We just want to check if we have access to remote servers, that is, those that are running Linux. We can take all the analysis to that remote server because, as I said, sometimes you need a lot of RAM and a lot of storage, and that might not be available on your local computer. So, you want to do this analysis on a remote server. Now, connect to this remote server.

So, one of the commands that we use is called ping. So, I will actually ping one server on our server where we can run this. So, simply type the IP of that server. So, wherever you are, So, if you have access to a remote server, there must be an IP or a name right where you can access it. So, I am just using this ping. So, ping actually checks whether there is network access to that server, right? So, if that server is online, okay. So, if it is online, you will get this kind of reply with time, etcetera, but if it is not online, you will see that the lost package will be all the packages that should be lost. Here, you see the package sent equals 4; it equals 4 and lost equals 0. So, the server is online.

So, we can access it remotely, ok? Now I will come to that in a moment: how do you actually access the server remotely, and how do you actually run the analysis there? So, this is the first command to check whether the server is online. The next command, which is then How do you access the server? is actually using the command SSH. If you have access to the server, you will have a right user name that will be given to you. You will have an account on that server, and at that rate, the server IP is okay or the server name is right. So, this is what we are going to do right here.



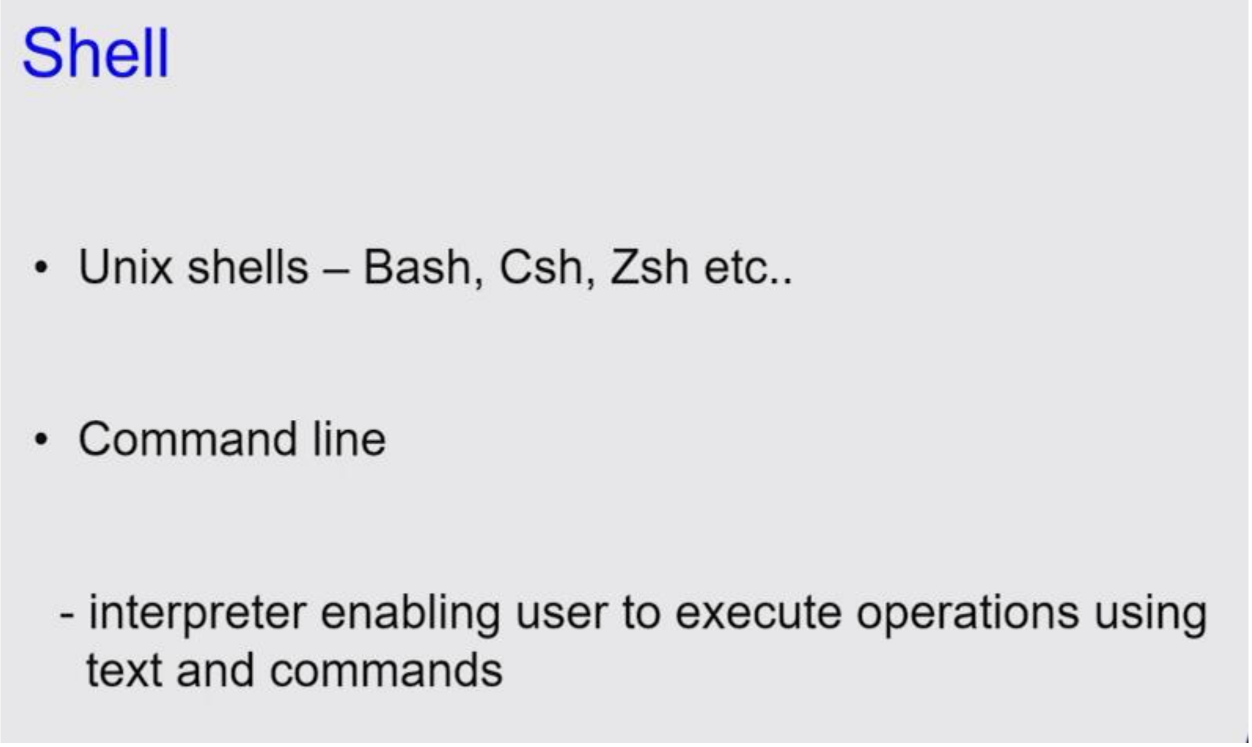
Access to remote servers

ping

ssh user@server.ip

scp

So, the SSH and I are going to access our server, and this is the IP of the server. So, I know this is the IP of the server. So, your IP could be anything else, and I have this user name and that account on that server. So, I can go and access that server, ok? So, the first time you do this, it will ask for this kind of storage of this key, etcetera. That is ok; you can say yes, and then you have to give that password. So, this is an account that will have a password that will be created by the administrator of that server. So, you have to give the server, and now we have logged in, and it gives you some information about the server, right? You can see here that the server is running this Ubuntu 17 version, and we have some updates that can be done, and it also gives information about the last login of this account. So, what we will do is now we want to run this. So, you can see this has actually changed, right? So, if you notice this path has actually changed to this, add bioserver, ok? We can simply say exit, and we come back to this path. So, this path before this command, where we type command right, will change once we access the server, ok? Now, how do we actually communicate with this server? How do you install tools? We will discuss this when we talk about shell commands, etcetera, step by step. How do you get data from the server? We usually use scp. This is a function we will discuss again once we get a bit into a bit more detail about accessing the command lines, commands, etc. in a moment. So, as I said, like all these commands we run, we run on shell, ok? So, the unique shells are there, and many shells are there. So, in this case the shell is bash. So, that is why I type this command, bash. Here you can see right, and the moment we try to type bash, it changes again. This part actually changes the path before it, which means we are now accessing the Linux system. So, once we are accessing this Linux system, we can now start the commands right on the terminals. So, we can use all the command-shell commands that we use in the Linux system, ok? So, here now, I will just clear the screen, and this is the shell where we will run all the basic shell commands. And this is required for doing all sorts of analysis.

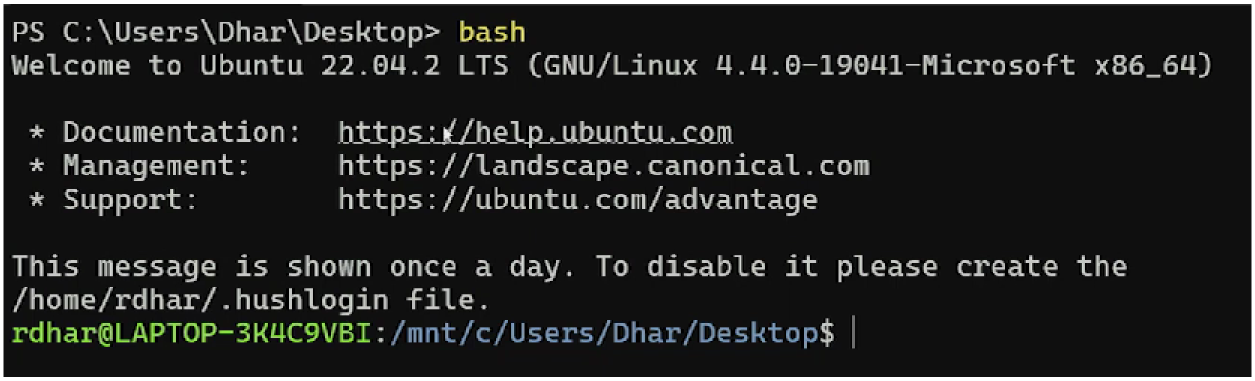


Shell

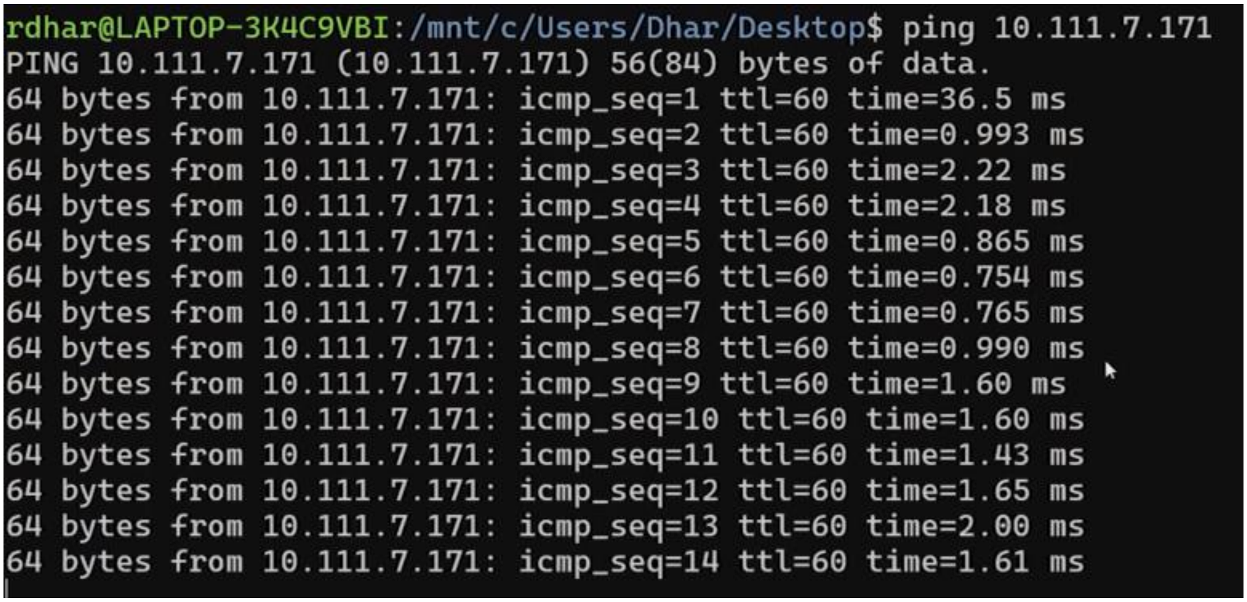
- Unix shells – Bash, Csh, Zsh etc..

- Command line

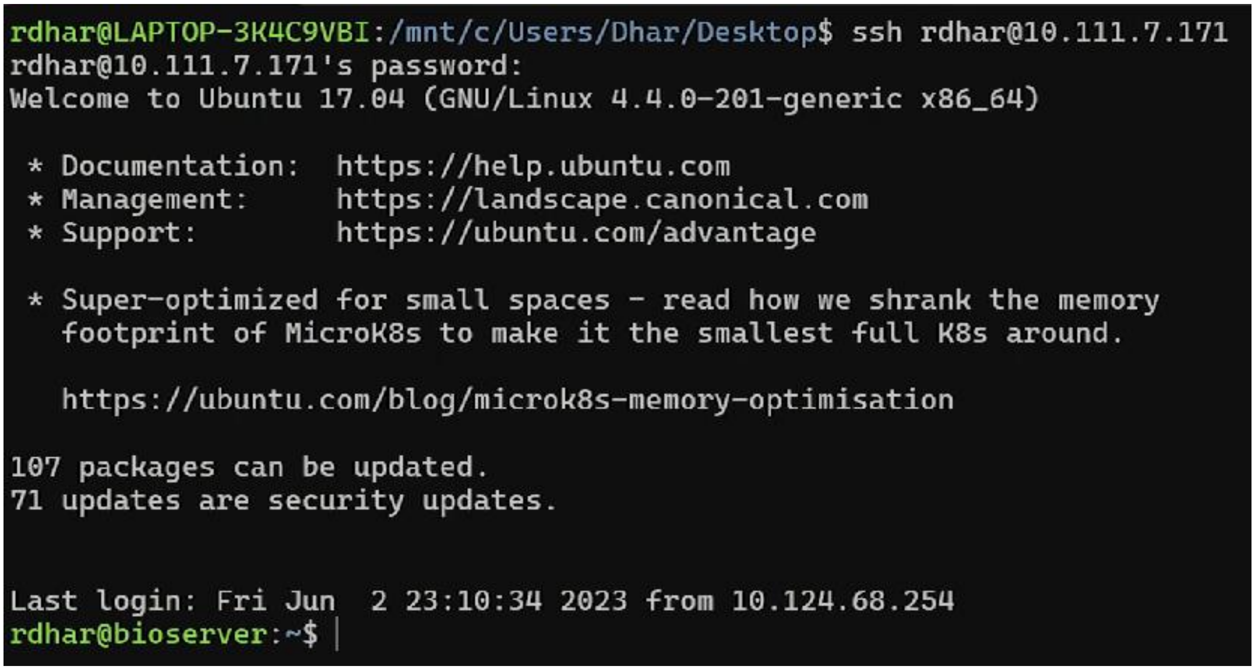
  - interpreter enabling user to execute operations using text and commands



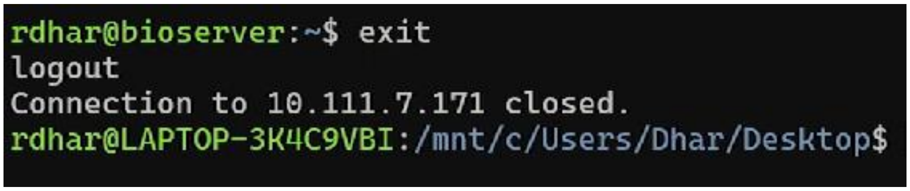
```
PS C:\Users\Dhar\Desktop> bash
Welcome to Ubuntu 22.04.2 LTS (GNU/Linux 4.4.0-19041-Microsoft x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

This message is shown once a day. To disable it please create the
/home/rdhar/.hushlogin file.
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop$
```



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop$ ping 10.111.7.171
PING 10.111.7.171 (10.111.7.171) 56(84) bytes of data.
64 bytes from 10.111.7.171: icmp_seq=1 ttl=60 time=36.5 ms
64 bytes from 10.111.7.171: icmp_seq=2 ttl=60 time=0.993 ms
64 bytes from 10.111.7.171: icmp_seq=3 ttl=60 time=2.22 ms
64 bytes from 10.111.7.171: icmp_seq=4 ttl=60 time=2.18 ms
64 bytes from 10.111.7.171: icmp_seq=5 ttl=60 time=0.865 ms
64 bytes from 10.111.7.171: icmp_seq=6 ttl=60 time=0.754 ms
64 bytes from 10.111.7.171: icmp_seq=7 ttl=60 time=0.765 ms
64 bytes from 10.111.7.171: icmp_seq=8 ttl=60 time=0.990 ms
64 bytes from 10.111.7.171: icmp_seq=9 ttl=60 time=1.60 ms
64 bytes from 10.111.7.171: icmp_seq=10 ttl=60 time=1.60 ms
64 bytes from 10.111.7.171: icmp_seq=11 ttl=60 time=1.43 ms
64 bytes from 10.111.7.171: icmp_seq=12 ttl=60 time=1.65 ms
64 bytes from 10.111.7.171: icmp_seq=13 ttl=60 time=2.00 ms
64 bytes from 10.111.7.171: icmp_seq=14 ttl=60 time=1.61 ms
```



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop$ ssh rdhar@10.111.7.171
rdhar@10.111.7.171's password:
Welcome to Ubuntu 17.04 (GNU/Linux 4.4.0-201-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Super-optimized for small spaces - read how we shrank the memory
   footprint of MicroK8s to make it the smallest full K8s around.

   https://ubuntu.com/blog/microk8s-memory-optimisation

107 packages can be updated.
71 updates are security updates.


Last login: Fri Jun  2 23:10:34 2023 from 10.124.68.254
rdhar@bioserver:~$ 
```



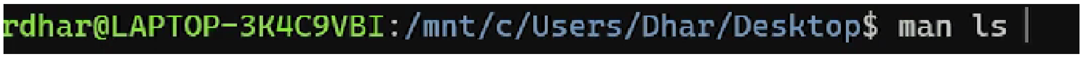
```
rdhar@bioserver:~$ exit
logout
Connection to 10.111.7.171 closed.
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop$
```

These are the basic commands for day-to-day life for managing data and looking at system usage, whether the tool is installed or not, etcetera. So, as I said, you do not have to run bash always; you can run something else, and this gives you the command line. So, the command line is actually an interpreter that enables us to execute operations using text and commands. We can give certain text and give certain commands, and those will be executed by the command line. Now, we will talk about some of the shell commands that are very useful for running anything in Linux. So, we will start with the very basic ones. So, and then we will go into more complex ones in this class in the next one. And then I will also show you how to actually access the server and how you actually download or communicate data with the server. Because if you are generating your own data, you want to put that data on the server, and all the analysis results, some of which you may want to download to your system for further analysis, visualizations, etcetera, are okay.

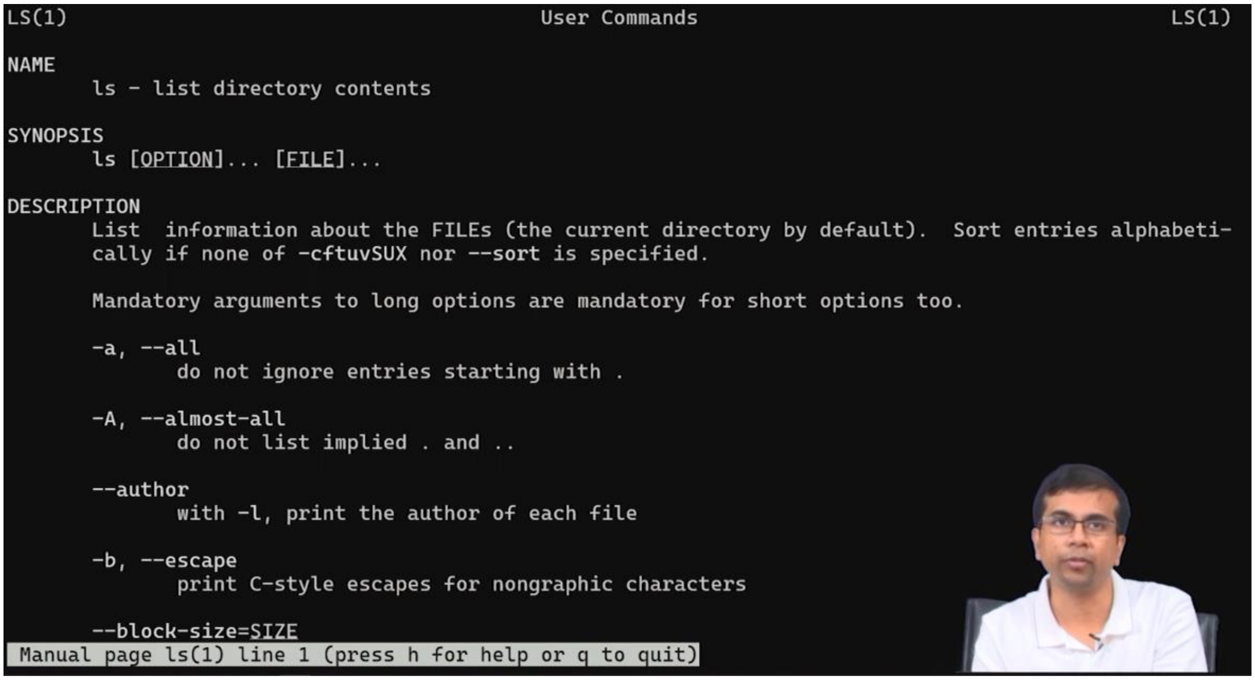
Basic shell commands

man

ls



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop$ man ls
```

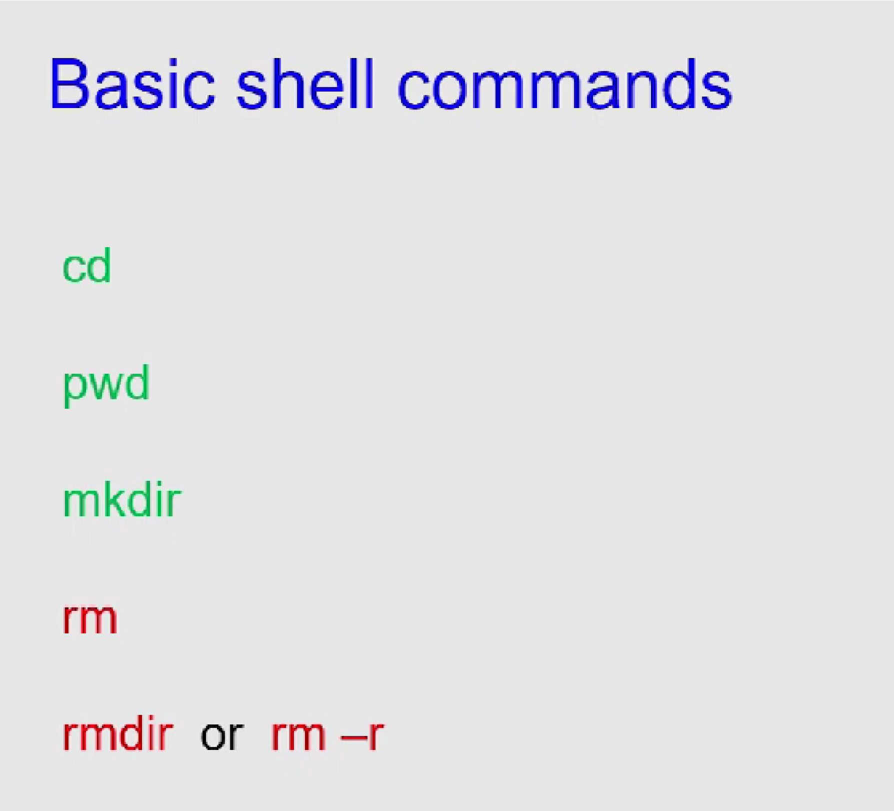


```
LS(1)                                  User Commands                                  LS(1)

NAME
       ls - list directory contents

SYNOPSIS
       ls [OPTION]... [FILE]...

DESCRIPTION
       List  information about the FILEs (the current directory by default).  Sort entries alphabeti-
       cally if none of -cftuvSUX nor --sort is specified.

       Mandatory arguments to long options are mandatory for short options too.

       -a, --all
              do not ignore entries starting with .

       -A, --almost-all
              do not list implied . and ..

       --author
              with -l, print the author of each file

       -b, --escape
              print C-style escapes for nongraphic characters

       --block-size=SIZE
 Manual page ls(1) line 1 (press h for help or q to quit)
```

So, the first command is called man ok. So, this actually gives you a manual for any other shell commands, ok? So, let us introduce the second command, and then we can use this man command, ok? So, the second command is right; it just leads all the file folders that are present in the current directory, ok. So, we will come to the directory structure in a moment, ok. So, let us type something called ls here; it is giving me all the files, etcetera, that are here or folders that are here in this directory. Now, if we want to test the man's function, So, what we do is man's right. So, it kind of tells us what we will do, okay? In addition, this will list directory contents; this is the LS function. Then you have some sort of description, ok? Then this description is actually followed by some signs, right? You can see this minus small m minus capital A minus B, etcetera.

These are options that you can use with this command, ok? So, if you want to use these options, you have to say something like ls space minus A ok. We will use this in some cases, as you will see in a few moments. So, for example, you have this minus D, ok? So, it will just list directories that are there, ok, not the contents of the directories, etcetera. So, how do you quit this? It says right; you can just type q, and this will quit ok.

So, let us try this minus D, and it does not find anything right. So, no directory, etcetera, ok? Now, the next command, ok, is actually associated with this directory structure. So, this is a very important part because, well, this is something that might not be used in the Windows system. In the Windows system, we just click and go into these subdirectories, etcetera, go on, move on, etcetera.



Basic shell commands

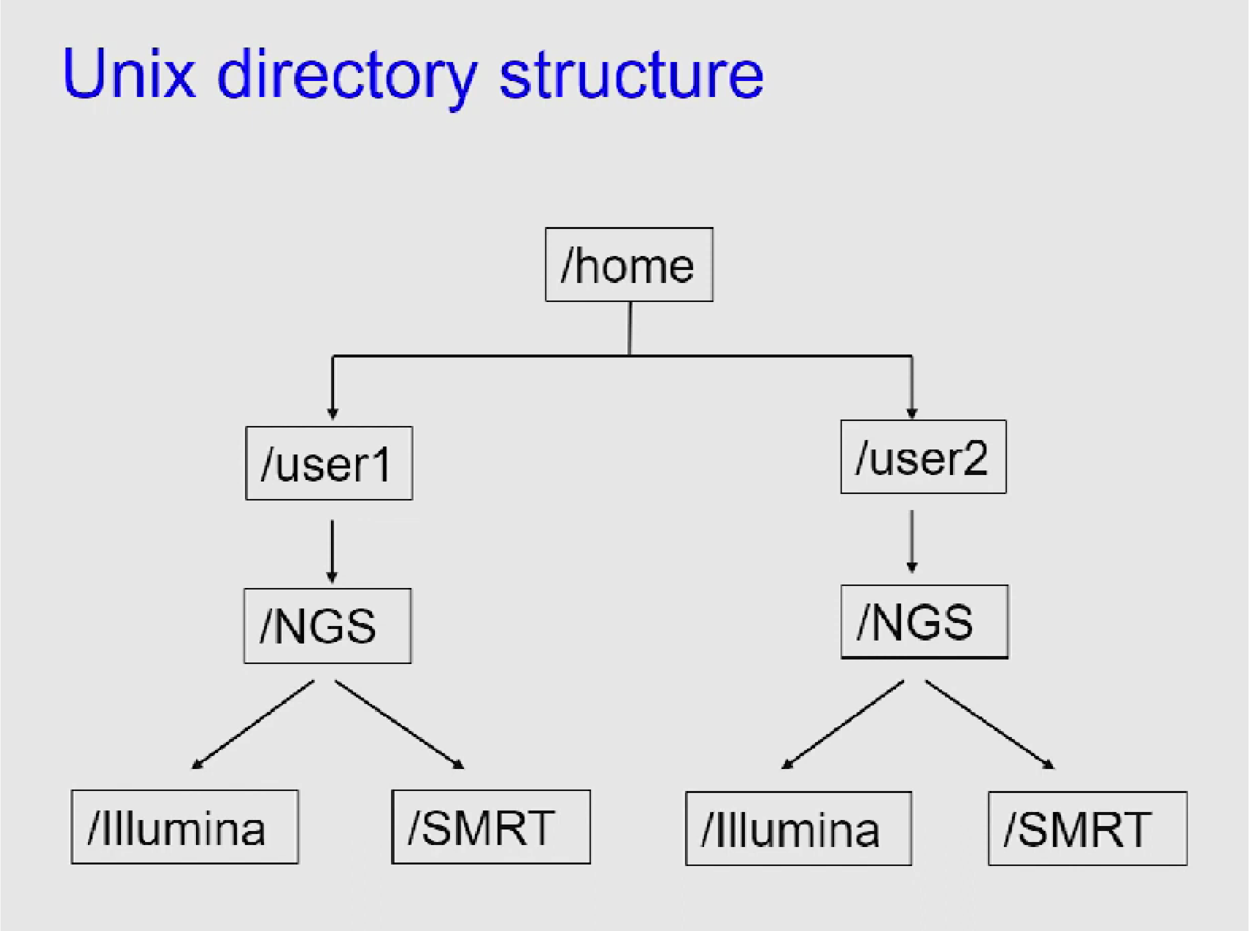
cd

pwd

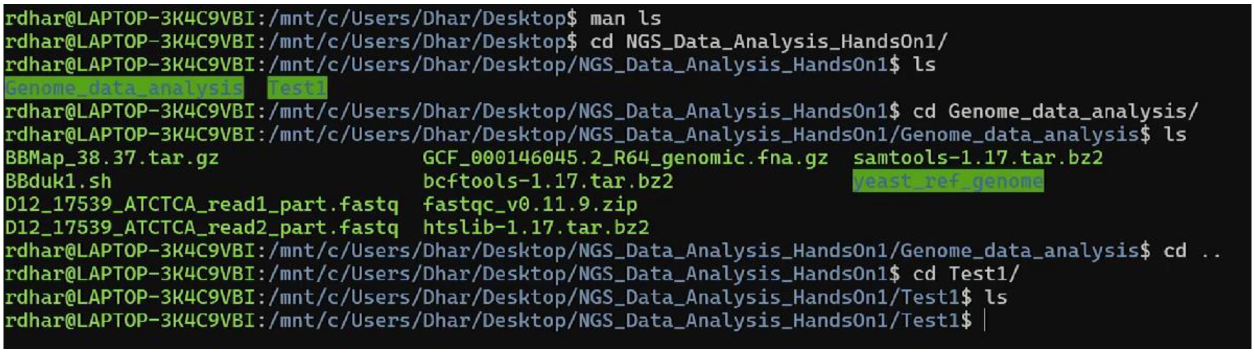
mkdir

rm

rmdir or rm –r

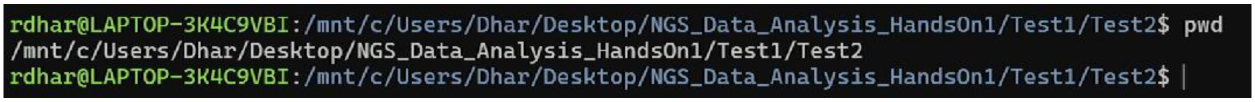


Unix directory structure

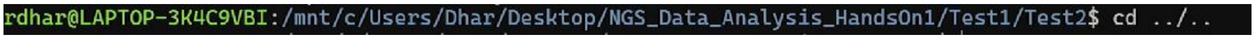


```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop$ man ls
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop$ cd NGS_Data_Analysis_HandsOn1/
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1$ ls
Genome_data_analysis  Test1
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1$ cd Genome_data_analysis/
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Genome_data_analysis$ ls
BBMap_38.37.tar.gz                  GCF_000146045.2_R64_genomic.fna.gz  samtools-1.17.tar.bz2
BBduk1.sh                           bcftools-1.17.tar.bz2               yeast_ref_genome
D12_17539_ATCTCA_read1_part.fastq  fastqc_v0.11.9.zip
D12_17539_ATCTCA_read2_part.fastq  htslib-1.17.tar.bz2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Genome_data_analysis$ cd ..
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1$ cd Test1/
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ |
```

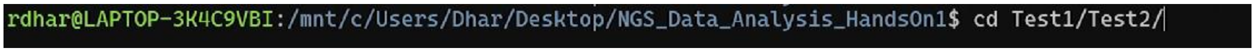
You can do that with some Linux systems, but that is really inefficient. In the terminal, you can actually change directories much faster with this command line. Then the first command that we see on the screen is actually the command for doing that. So, cd actually means change directory, okay? So, let us change the directory. There is a folder called desktop in this path, ok? So, I have changed the directory to that folder. Now, this blue part here actually tells me my current directory, ok? It is a kind of change when I change directory, and I can change further right. I can go on to say something like I have a directory called NGS_Data_Analysis_HandsOn1. So, I will change the directory. So, I will move into that directory here, ok, and then again, you see this path has changed, right? So, it is telling me where I am wrong. Now, you can also use another command that will tell you where you are. It is called PWD OK. So, the pwd means pwd is the path present marking the directory OK, and notice the directory structure is now OK. So, slightly different from the Windows directory structure is how their directories are written. So, here is the directory structure you have: slash mnt, slash c, etcetera.



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1/Test2$ pwd
/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1/Test2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1/Test2$ |
```



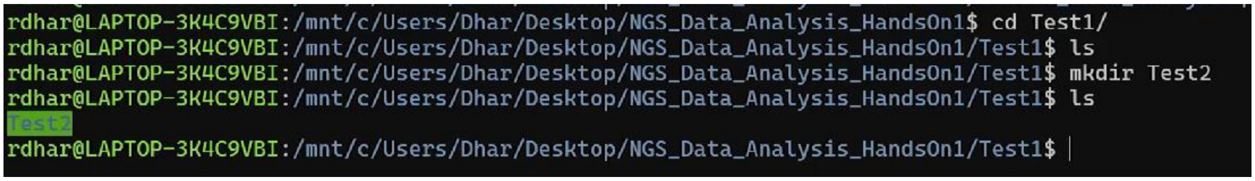
```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1/Test2$ cd ../..
```



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1$ cd Test1/Test2/
```

Therefore, the full path of the directory where you are. Now, I will discuss this directory structure in a moment and how you can actually navigate this directory structure very easily. Now, one of the functions that you can use is to create a directory using the command line itself. So, in Windows, you will just right-click and create a new directory, but in Linux, to be more efficient, we use the command line for most of the tasks, and the command is called mkdir. So, we can do something called mkdir test 1, ok? So, it is a new directory that we are creating. So, mkdir stands for make directory, and we have created this test. Now, let us say we do not want this vector, right? We want to delete this vector, right? So, again, you can go and do this with the mouse, but again, if you want to do this through the command line, the command is called rmdir ok, remove directory and the directory name is ok. One of the things I should mention is that you should be very careful when using these RM commands, because this is irreversible. Once you have removed it you cannot get it back. Now, the good thing with rmdir is that it will not remove the directory if you have a file inside the directory. If

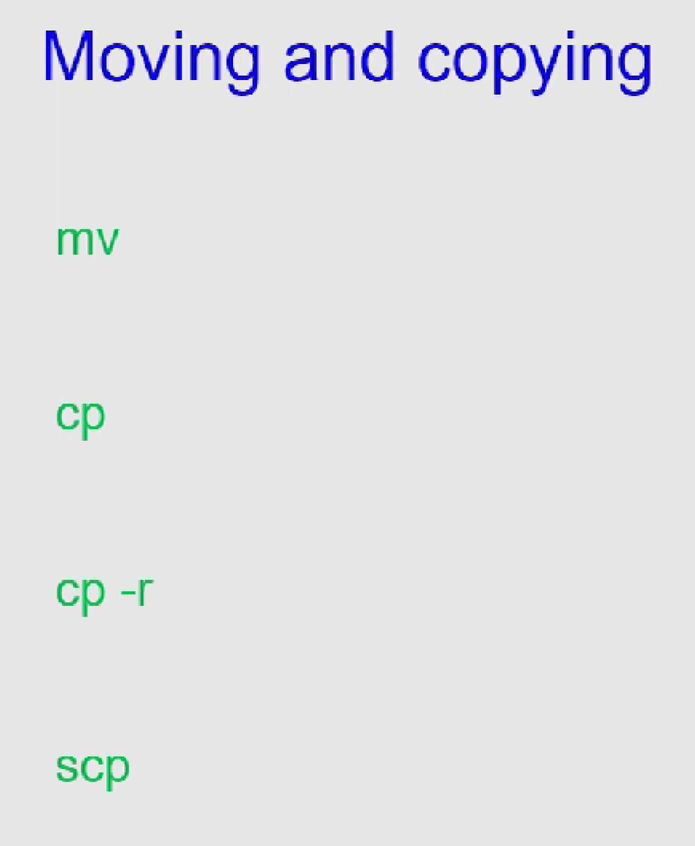
rmdir will not work, you will have to use something else, which is called rm minus r. So, that is why I have written this option here. You can see that you have rmdir or rm minus r ok. So, if you have data and you want to delete the directory along with the data, then you will use this rm minus r command ok. So, again, let us create this right, and you can create it. So, let us now move in inside this directory test 1. We can create another one right directory inside this test 1; let us call it test 2 right, and then we want to now come back to the original directory and delete this test 1 directory.



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1$ cd Test1/
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ mkdir Test2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls
Test2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ |
```

Now, we have test 2 inside, ok, and we will see how this rmdir and rm minus r work. Now, how do you go back? How do you go up? I have shown you right, with a CD, you can go inside. Let us say we have moved into desktop, we have moved into NJ data analysis, we have now created test 1, we have moved into test 2, etcetera, right? How do you actually go up? How do you come out of the directory again? In Linux, you can do that in the terminal, which is called cd dot dot. Ok, cd dot dot will bring you back to the NGS_Data_Analysis directory. If you want to now say I do not, it might seem very tedious, right? You want to go back one step at a time, but if you want to let us say go into test 1 right, you remember that now you have test 1 inside this directory and then we have test 2 inside test 1 ok. If you want to go inside test 1 and then to test 2 right instead of 2 steps, can we do this in a single step? The answer is yes, we can. So, you can simply say cd test 1 test 2 ok you can actually have as many steps as you want ok it is not limited to just one step at a time you can do test 1 test 2 you can have test 3 test 4 etcetera right. So, let us do this, okay? So, from hands-on 1 directory we have moved into test 2 in one step. Now can we come out again in one step to hands on 1? Now say yes, ok, and the command is cd dot dot right. We want to go back and go up in the directory, and we have to do it two times because we want to come back to hands on 1. So, once we do this two times again, we are coming back to this hands-on directory, ok? So, this is the directory structure that you have to remember or kind of visualize in your mind when you are navigating this directory. So, at first, it might seem a bit challenging and it might be forgotten, but you always have a list to see what you have and which directory, and then over time, you will get used to the directory structure. You will have this in mind, right? So, the other command that is there is the rm command, and as you probably can guess this is a remove file command. So, if you have a file, you just remove that file by saying rm test 1 dot txt some file name, and that will delete that file. These commands I have shown in red in the presentation because you would have to be very careful. Once you have used this command, there is no going back; you have deleted the file or the folder. So, there is no going back again; you said you did it by mistake and you have lost the data, ok? So, these are commands that you should use very carefully. Now, let us delete this directory that we have created.

So, rm minus minus r test 1 test 2 is right. So, if you want to delete, let us say test 1 along with its contents; we just say RM minus 1 r test 1 ok. So, this will delete this directory, ok? So, just to illustrate the directory structure now, just come back here. So, we have this directory structure that I have just shown you, right? This is what you see. You have this directory here. These examples are given: home user 1; user 2; NGS Illumina SMRT. So, if you want to come from home to NGS, you will see cd user 2 NGS right, and if you want to go back to home from NGS, you say cd dot dot slash dot dot slash right 2 dot dot slash ok and this is what we have just demonstrated.



Moving and copying

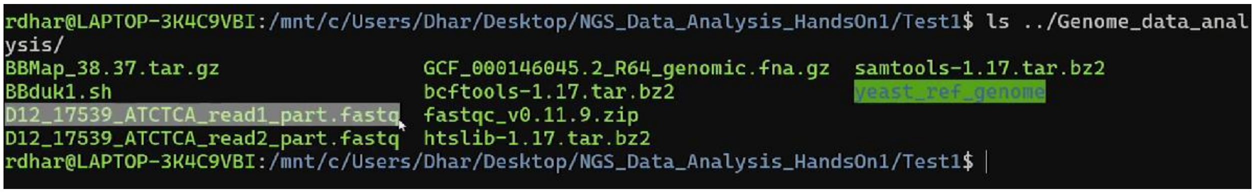
mv

cp

cp -r

scp

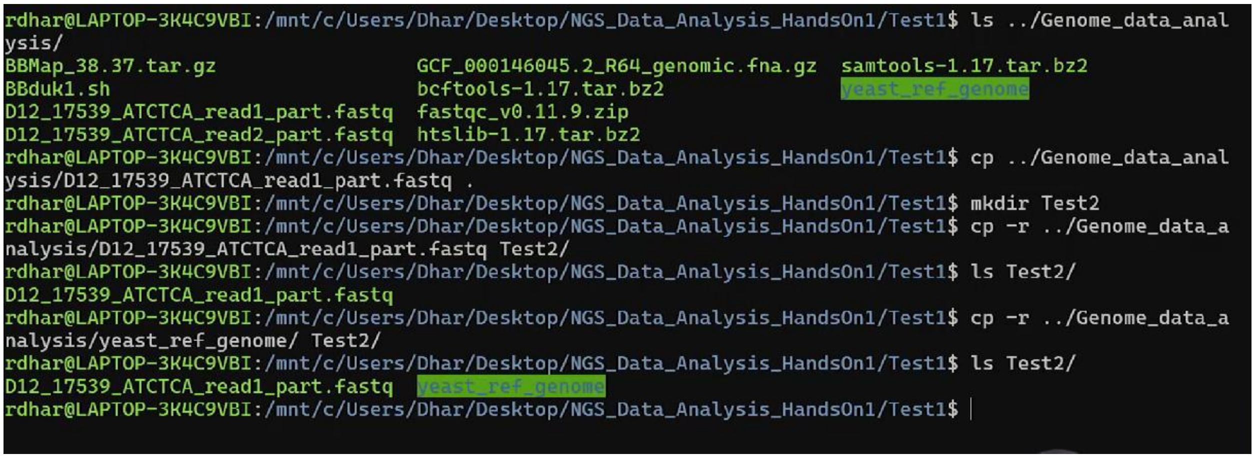
There are other very important commands, right? So, one is called MV. So, it will move a file or a folder to another location, and we also have something called cp, which is the copy command, and then you have the cp minus r command. So, the cp command is used on files, and the cp minus r command is used on folders. So, if you have a folder of data you want to copy recursively, right every file in that folder, you will do s through cp minus r. There is another version of CP, which is the scp. I just mentioned that we use this scp for data transfer through the server to the server or from the server; this is called secure copy. So, the data that is transferred is encrypted, okay? If cp does not encrypt the data, scp does ok. So, that is why, for communication with the server, we always use SCP. And then we have some other commands, which are called cat head tail. So, I will just mention that I will just show you these commands: mv, cp, etcetera, how they work, and scp, maybe when we connect to the server.



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls ../Genome_data_anal
ysis/
BBMap_38.37.tar.gz                 GCF_000146045.2_R64_genomic.fna.gz  samtools-1.17.tar.bz2
BBduk1.sh                          bcftools-1.17.tar.bz2               yeast_ref_genome
D12_17539_ATCTCA_read1_part.fastq  fastqc_v0.11.9.zip
D12_17539_ATCTCA_read2_part.fastq  htslib-1.17.tar.bz2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ |
```

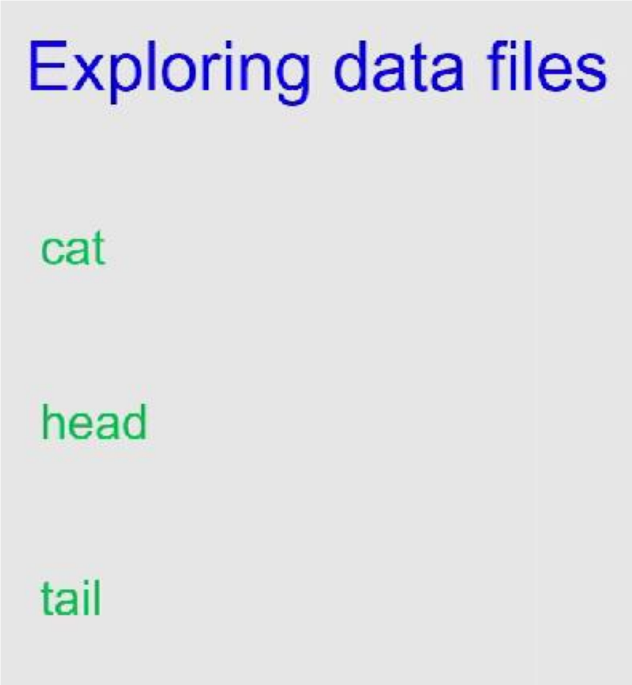
Then we have the data file exploration, right? So, let us say we have a file ng data file. To see what you have in that file, right some of the contents of that file. So, we use this command, cat. So, cat, will you show the full content of the data in the terminal? Sometimes this is not very good, right? If you have a big file, you do not want to see the full data. So, what you want to do is say you want to see the first few lines or the last few lines, and in that case, you use something called a head or tail.

So, head means it will show you the first 5–6 lines of the file, and tail will show you the last few lines of that file. So, that is good; you just get an idea of the data format, etcetera. And you can also use something called more or less; these are also commands that will show you the contents of the files. So, let us try some of this, and then we will understand it better. So, let us create this MKDR Test1 again because we will use this command. So, we can use it right. So, here in the LS, you can see we have this genome data analysis right, and inside this genome data analysis, we have some files right that we can copy. So, again, how do you copy this data? Let me say cp right, we can take any data from any file that you want. Let us say we want this fastq file. We have this fastq file, ok. And we want to copy these two tests, ok? The format is that you have to give the command, then the file name that you want to copy, and the destination. So, if the destination is same, you just would not copy it, right? If we want to save it as a new name, then you have to give the new name, but if you want to copy the file to another destination, you can simply say give the destination, and you get this test done, ok? So, copy and move are different, right? So, to copy, you will just make a copy move, which will move the file.



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls ../Genome_data_anal
ysis/
BBMap_38.37.tar.gz                 GCF_000146045.2_R64_genomic.fna.gz  samtools-1.17.tar.bz2
BBduk1.sh                          bcftools-1.17.tar.bz2               yeast_ref_genome
D12_17539_ATCTCA_read1_part.fastq  fastqc_v0.11.9.zip
D12_17539_ATCTCA_read2_part.fastq  htslib-1.17.tar.bz2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ cp ../Genome_data_anal
ysis/D12_17539_ATCTCA_read1_part.fastq .
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ mkdir Test2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ cp -r ../Genome_data_a
nalysis/D12_17539_ATCTCA_read1_part.fastq Test2/
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls Test2/
D12_17539_ATCTCA_read1_part.fastq
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ cp -r ../Genome_data_a
nalysis/yeast_ref_genome/ Test2/
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls Test2/
D12_17539_ATCTCA_read1_part.fastq  yeast_ref_genome
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ |
```

So, it is kind of like a cut-and-paste ok? So, the format is exactly the same with CP and MV. Now, let us check, like, whether this file has been copied okay. So, how do you check this? The command is right, and again, we are playing with the directory structure here, right? So, dot dot slash because we need to go up where the Test1 directory decides right. So, that is why we have gone to Test1, and we can check this file, and we see that this file has been copied. Now, let us move to Test 1 because we want to try reading that file. We will not use this cat command because it will add all the data to the terminal. So, let us move in there, and I will clear the screen for better visibility. We will not use the cat command because the full data will be displayed. We can use this command, head, okay, because this will show us only a part of the data, right? You see the first few lines are shown, and similarly, you can see this with the tail command OK and the first few lines. So, these two commands are really useful when you are working with these NGS datasets because you do not want to concatenate the full data in the terminal because that will overlap the terminal with this huge dataset.

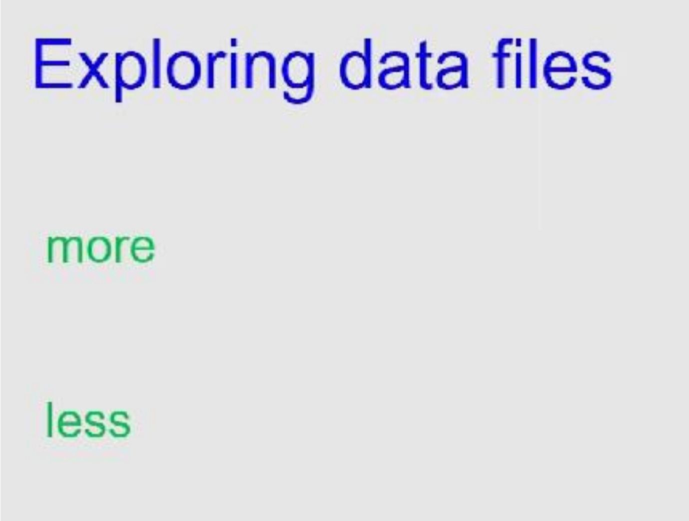


Exploring data files

cat

head

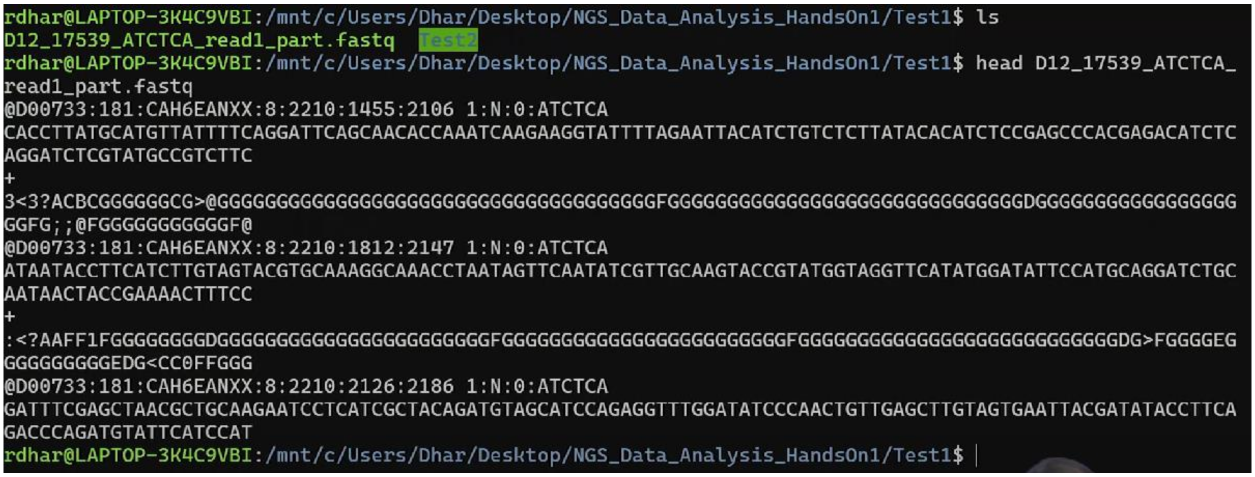
tail



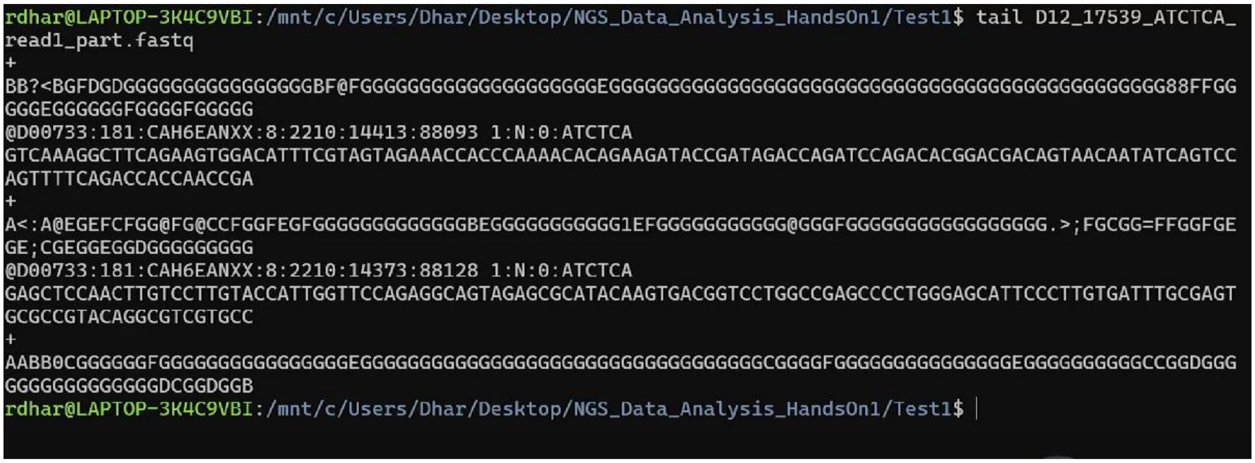
Exploring data files

more

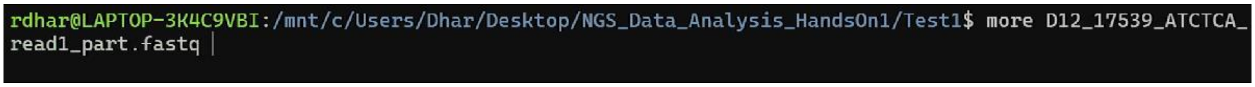
less



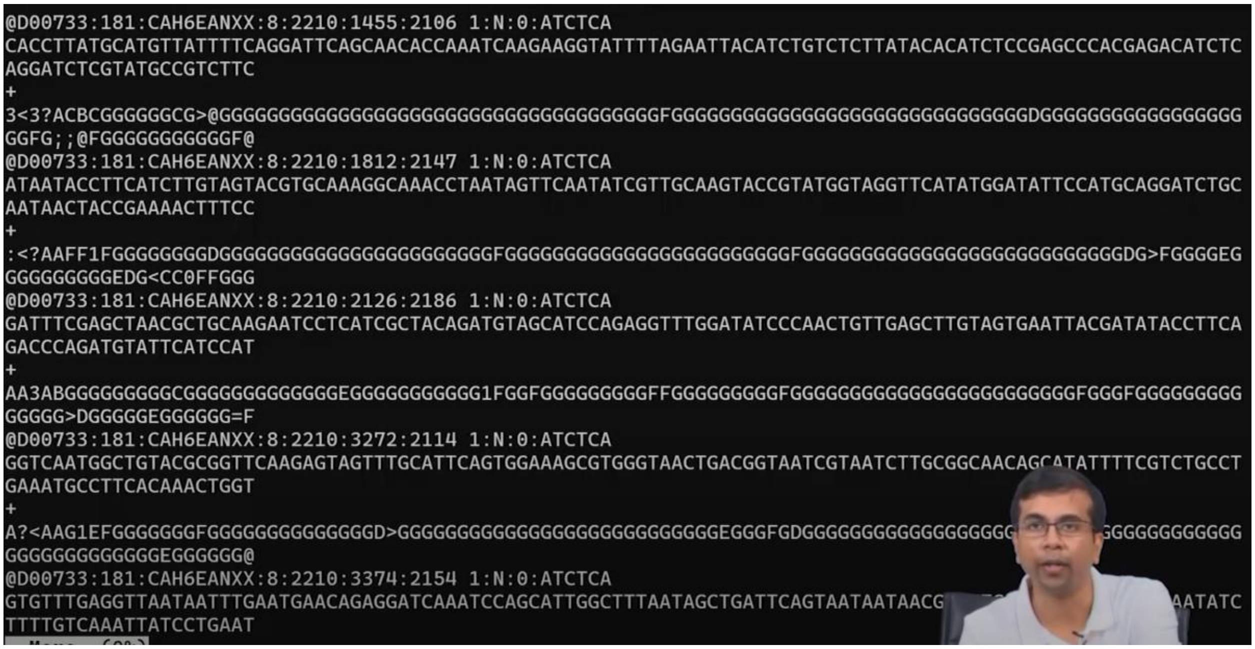
```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls
D12_17539_ATCTCA_read1_part.fastq  Test2
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ head D12_17539_ATCTCA_
read1_part.fastq
@D00733:181:CAH6EANXX:8:2210:1455:2106 1:N:0:ATCTCA
CACCTTATGCATGTTATTTTCAGGATTCAGCAACACCAAATCAAGAAGGTATTTTAGAATTACATCTGTCTCTTATACACATCTCCGAGCCCACGAGACATCTC
AGGATCTCGTATGCCGTCTTC
+
3<3?ACBCGGGGGGCG>@GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGDGGGGGGGGGGGGGGGGG
GGFG;;@FGGGGGGGGGGGF@
@D00733:181:CAH6EANXX:8:2210:1812:2147 1:N:0:ATCTCA
ATAATACCTTCATCTTGTAGTACGTGCAAAGGCAAACCTAATAGTTCAATATCGTTGCAAGTACCGTATGGTAGGTTCATATGGATATTCCATGCAGGATCTGC
AATAACTACCGAAAACTTTCC
+
:<?AAFF1FGGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGDG>FGGGGEG
GGGGGGGGGEDG<CC0FFGGG
@D00733:181:CAH6EANXX:8:2210:2126:2186 1:N:0:ATCTCA
GATTTCGAGCTAACGCTGCAAGAATCCTCATCGCTACAGATGTAGCATCCAGAGGTTTGGATATCCCAACTGTTGAGCTTGTAGTGAATTACGATATACCTTCA
GACCCAGATGTATTCATCCAT
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$
```



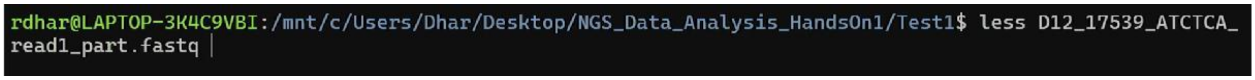
```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ tail D12_17539_ATCTCA_
read1_part.fastq
+
BB?<BGFDGDGGGGGGGGGGGGGGGGBF@FGGGGGGGGGGGGGGGGGGGGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG88FFGG
GGGEGGGGGGFGGGGFGGGGG
@D00733:181:CAH6EANXX:8:2210:14413:88093 1:N:0:ATCTCA
GTCAAAGGCTTCAGAAGTGGACATTTCGTAGTAGAAACCACCCAAAACACAGAAGATACCGATAGACCAGATCCAGACACGGACGACAGTAACAATATCAGTCC
AGTTTTCAGACCACCAACCGA
+
A<:A@EGEFCFGG@FG@CCFGGFEGFGGGGGGGGGGGGGBEGGGGGGGGGGG1EFGGGGGGGGGGG@GGGFGGGGGGGGGGGGGGGGG.>;FGCGG=FFGGFGE
GE;CGEGGEGGDGGGGGGGGG
@D00733:181:CAH6EANXX:8:2210:14373:88128 1:N:0:ATCTCA
GAGCTCCAACTTGTCCTTGTACCATTGGTTCCAGAGGCAGTAGAGCGCATACAAGTGACGGTCCTGGCCGAGCCCCTGGGAGCATTCCCTTGTGATTTGCGAGT
GCGCCGTACAGGCGTCGTGCC
+
AABB0CGGGGGGFGGGGGGGGGGGGGGGGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGCGGGGFGGGGGGGGGGGGGGEGGGGGGGGGCCGGDGGG
GGGGGGGGGGGGGDCGGDGGB
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$
```



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ more D12_17539_ATCTCA_
read1_part.fastq
```



```
@D00733:181:CAH6EANXX:8:2210:1455:2106 1:N:0:ATCTCA
CACCTTATGCATGTTATTTTCAGGATTCAGCAACACCAAATCAAGAAGGTATTTTAGAATTACATCTGTCTCTTATACACATCTCCGAGCCCACGAGACATCTC
AGGATCTCGTATGCCGTCTTC
+
3<3?ACBCGGGGGGCG>@GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGDGGGGGGGGGGGGGGGGG
GGFG;;@FGGGGGGGGGGGF@
@D00733:181:CAH6EANXX:8:2210:1812:2147 1:N:0:ATCTCA
ATAATACCTTCATCTTGTAGTACGTGCAAAGGCAAACCTAATAGTTCAATATCGTTGCAAGTACCGTATGGTAGGTTCATATGGATATTCCATGCAGGATCTGC
AATAACTACCGAAAACTTTCC
+
:<?AAFF1FGGGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGDG>FGGGGEG
GGGGGGGGGEDG<CC0FFGGG
@D00733:181:CAH6EANXX:8:2210:2126:2186 1:N:0:ATCTCA
GATTTCGAGCTAACGCTGCAAGAATCCTCATCGCTACAGATGTAGCATCCAGAGGTTTGGATATCCCAACTGTTGAGCTTGTAGTGAATTACGATATACCTTCA
GACCCAGATGTATTCATCCAT
+
AA3ABGGGGGGGGGCGGGGGGGGGGGGGEGGGGGGGGGGG1FGGFGGGGGGGGGFFGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGFGGGFGGGGGGGGG
GGGGG>DGGGGGEGGGGGG=F
@D00733:181:CAH6EANXX:8:2210:3272:2114 1:N:0:ATCTCA
GGTCAATGGCTGTACGCGGTTCAAGAGTAGTTTGCATTCAGTGGAAAGCGTGGGTAACTGACGGTAATCGTAATCTTGCGGCAACAGCATATTTTCGTCTGCCT
GAAATGCCTTCACAAACTGGT
+
A?<AAG1EFGGGGGGGFGGGGGGGGGGGGGGD>GGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGFGDGGGGGGGGGGGGGGGGG[illegible]GGGGGGGGGGGG
GGGGGGGGGGGGGEGGGGGG@
@D00733:181:CAH6EANXX:8:2210:3374:2154 1:N:0:ATCTCA
GTGTTTGAGGTTAATAATTTGAATGAACAGAGGATCAAATCCAGCATTGGCTTTAATAGCTGATTCAGTAATAATAACG[illegible]AATATC
TTTTGTCAAATTATCCTGAAT
```



```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ less D12_17539_ATCTCA_
read1_part.fastq
```

You just want to see part of this and see whether they look good or alright, right? So, for that, you can use this head and tail command. You can also use this less command, as I mentioned more or less right. This will also give you some idea about this dataset, right? And if you so, for less, you see like you're kind of moving along, right, and you can actually search certain strings if you want. For example, if you want to search this number and see if it is there, you can see it here. So, less is slightly more functional than this head and tail, ok? If you want to come out of this file, just type q, and it will quit ok. So, these are some of the commands, and finally, I will talk about this data compression. So, the compression that we use in Windows is zip compression, right? So, here you can do this with terminal zip file 1 dot zip file 1 right. So, remember, we need to specify the zip file first, followed by the file that we want to compress, or we can use zip minus r folder 1 dot zip folder 1 if you want to compress the folder of folder 1. And finally, we can also unzip using the terminal, which is unzip command unzip file 1 dot zip ok. So, this will actually illustrate a bit more because there are other compressions that are specific to Linux. So, these are the references that we have used, and we have introduced the command-line-based system right or shell for this NGS data analysis, and these are required for navigating to the unique system right. This will be used all the time. Thank you. .