TRAINER'S MANUAL

Introduction to Command-Line & Scripting For Bioinformatics

Steve Pederson
Stephen Bent
Dan Kortschak

University of Adelaide 24th March, 2016

TRAINER'S MANUAL

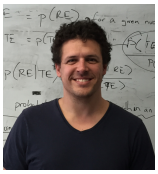
Licensing

Contents

Workshop Information

The Trainers

Mr. Steve Pederson
Co-ordinator
Bioinformatics Hub
The University of Adelaide
South Australia
stephen.pederson@adelaide.edu.au



Dr. Jimmy Breen
Bioinformatician
Bioinformatics Hub & Robinson Research Institute
The University of Adelaide
South Australia
jimmy.breen@adelaide.edu.au

Dr. Hien To
Bioinformatician
Bioinformatics Hub
The University of Adelaide
South Australia
hien.to@adelaide.edu.au

Mr Alastair Ludington
Bioinformatician
Bioinformatics Hub
The University of Adelaide
South Australia
alastair.ludington@adelaide.edu.au

Ms Ramona Rogers
Computing Officer
IT Support
The University of Adelaide
South Australia
ramona.rogers@adelaide.edu.au

Welcome

Thank you for your attendance & welcome to the Introduction to Command Line & Shell Scripting Workshop. This is an offering by the University of Adelaide, Bioinformatics Hub which is a centrally funded initiative from the Department of Vice-Chancellor (Research), with the aim of assisting & enabling researchers in their work. Training workshops & seminars such as this one are an important part of this initiative. The Bioinformatics Hub itself has a web-page at http://www.adelaide.edu.au/bioinformatics-hub/, and **to be kept up to date on upcoming events and workshops, please join the internal Bioinformatics mailing list on** http://list.adelaide.edu.au/mailman/listinfo/bioinfo.

The Bioinformatics Hub has just started a Twitter account, so please follow us on (https://twitter.com/UofABioinfoHub/). We also have an active Slack team for discussing Bioinformatics questions with the local community. Slack teams do require an invitation to join, so please email the Hub on bioinf_hub@adelaide.edu.au to join the community. All are welcome.

Today's workshop has been put together based on previous material and courses prepared by Dr Stephen Bent (*University of Queensland*), with generous technical support & advice provided by Dr Nathan Watson-Haigh (*ACPFG*) and Dr Dan Kortschak (*Adelaide University, Adelson Research Group*). We hope it will be useful in enabling you to continue and to advance your research.

Course Summary

In today's workshop, the morning session will be spent introducing you to the basic tools and concepts required for data handling. In the afternoon session we'll develop these skills to a more advanced level, with progress in both sessions being made at your own pace. Some people may finish early today, but the majority of you probably won't. The VMs you use will be active for the next two weeks should you wish to continue working through the material after the workshop. We will also be using the same VMs for next week's workshop *Introduction to Next Generation Sequencing (NGS) Data*.

The majority of data handling and analysis required in the field of bioinformatics uses the *command line*, alternatively known as the terminal or the *bash shell*. This is a text-based interface in which commands must be typed, as opposed to the Graphical User Interfaces (aka GUIs) that most of us have become accustomed to. Being able to access your computer using these tools enables you to more fully utilise the power & capabilities of your machine, for both Linux & Mac operating systems, and to a lesser extent will even enable you to dig deeper on a Windows system.

Whilst some of the tools we cover today may appear trivial, they are used on a daily basis by those working in the field. These basic tools are essential for writing what are known as *shell scripts*, which we will begin to cover in the afternoon session. These are essentially simple programs that utilise the inbuilt functions of the shell, and are used to automate processes such as de-multiplexing read libraries, or aligning reads to the genome. A knowledge of this simple type of programming and navigation is also essential for accessing the high-performance computing resources such as `phoenix`.

Using the Post-it Notes

For today's session, you will be provided with 3 post-it notes of differing colours. Please use these to signal whether you need help or not by placing them on your monitors. We will interpret these as:

1. **Red** - Help! I can't make something work
2. **Yellow** - I'm working on something, but haven't made it yet
3. **Green** - I've finished the task I was working on

Providing Feedback

While we endeavour to deliver a workshop with quality content and documentation in a venue conducive to an exciting, well run hands-on workshop with a bunch of knowledgeable and likable trainers, we know there are things we could do better.

Whilst we want to know what didn't quite hit the mark for you, what would be most helpful and least depressing, would be for you to provide ways to improve the workshop. i.e. constructive feedback. After all, if we knew something wasn't going to work, we wouldn't have done it or put it into the workshop in the first place!

Clearly, we also want to know what we did well! This gives us that "feel good" factor which will see us through those long days and nights in the lead up to such hands-on workshops!

Document Structure

We have provided you with an electronic copy of the workshop's hands-on tutorial documents. We have done this for two reasons: 1) you will have something to take away with you at the end of the workshop, and 2) you can save time (mis)typing commands on the command line by using copy-and-paste.

We advise you to use Acrobat Reader to view the PDF. This is because it properly supports some features we have implemented to ensure that copy-and-paste of commands works as expected. This includes the appropriate copy-and-paste of special characters like tilde and hyphens as well as skipping line numbers for easy copy-and-past of whole code blocks.

While you could fly through the hands-on sessions doing copy-and-paste, you will learn more if you use the time saved from not having to type all those commands, to understand what each command is doing!

The commands to enter at a terminal look something like this:

```
1 tophat --solexa-quals -g 2 --library-type fr-unstranded -j \
    annotation/Danio_rerio.Zv9.66.spliceSites -o tophat/ZV9_2cells \
    genome/ZV9 data/2cells_1.fastq data/2cells_2.fastq
```

The following styled code is not to be entered at a terminal, it is simply to show you the syntax of the command. You must use your own judgement to substitute in the correct arguments, options, filenames etc

```
tophat [options]* <index_base> <reads_1> <reads_2>
```

The following icons are used in the margin, throughout the documentation to help you navigate around the document more easily:

 Important

 For reference

 Follow these steps

 Questions to answer

 Warning - STOP and read

 Bonus exercise for fast learners

 Advanced exercise for super-fast learners

Computer Setup

We will all be working on our own computers today, and will be accessing Virtual Machines running the Ubuntu operating system on `phoenix`, which is the University of Adelaide's High Performance Computing (HPC) system. The software client `X2GO` which you will have already installed, enables us to access these machines in a familiar Desktop style, even though the majority of our time will be spent within the terminal.

You will have been allocated a VM with an associated IP address. To connect to your VM, **please follow these instructions carefully**. First, we need to create a session with the basic parameters

1. Open X2GO
2. Enter *IntroductionToBash* as the **Session Name**
3. Enter your *IP address* where it say **Host**
4. Enter the word **hub** as the login. **This must be all lower-case**
5. Select **XFCE** from the drop-down menu under **Session Type**
6. Click OK

Now we have created the session, it will appear in your X2Go on the right. To log onto the VM, we simply click on the session, and enter the password **hub**. **Click OK if you receive a message about a security key**. If this process fails, please place the red post-it note on your monitor.

We advise maximising your X2Go window to replicate sitting at the VM as if it is your local machine.

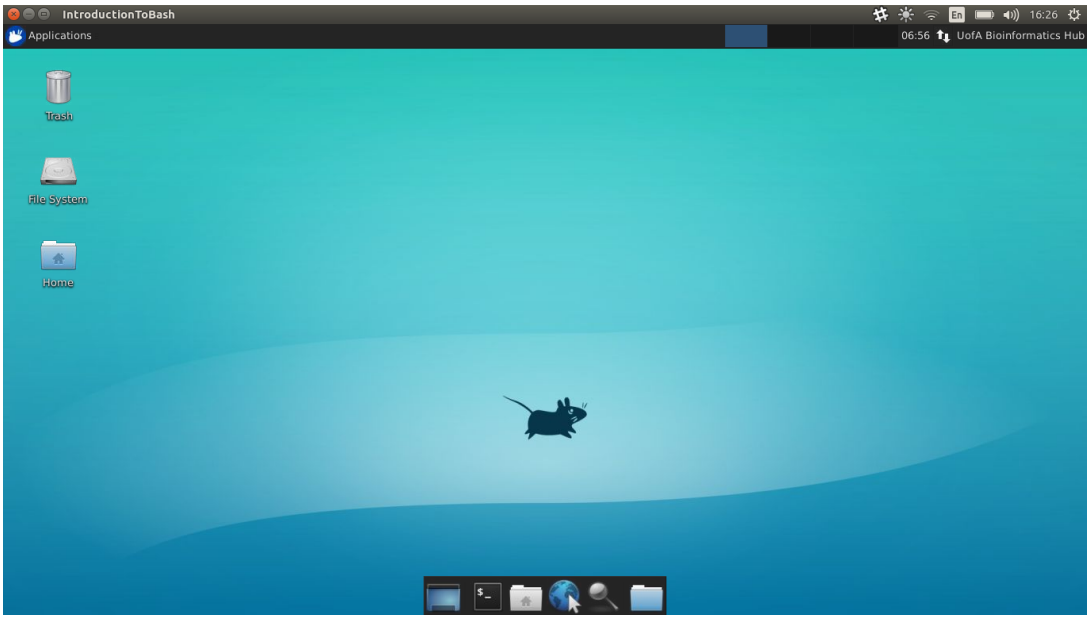


Figure 1: The VM desktop after logon with X2GO

The Ubuntu Desktop

Now that you are connected, you will notice we are in a standard graphical environment. The default Desktop in Ubuntu is Unity, but what we are seeing is one of the many alternatives, known as XFCE. We are using this as it is the most simple for remote connections. As many of us are used to seeing, there are click-able icons on the desktop, and drop-down menus.

Although we won't be using them today, Ubuntu has an built-in Office Suite of programs which you can access from the *Applications* > *Office* menu item. This is where links can be found to open Document Viewer (a .pdf viewer), Libre Office Calc (Excel-like), Libre Office Writer (Word-like) & other standard members of Office Program Suites.

The main interface we will be using today is the `terminal`which can be accessed from the set of icons at the bottom of you screen. **Firefox** can also be accessed from the terminal using the command `firefox &`, by clicking the `Web Browser` icon, or from the drop-down menu in the top left under the group *Internet*.

Introducing The Command Line

Primary Author(s):
Stephen Bent, Robinson Research Institute, University of Adelaide
Steve Pederson, Bioinformatics Hub, University of Adelaide
stephen.pederson@adelaide.edu.au

Contributor(s):
Dan Kortschak, Adelson Research Group, University of Adelaide
dan.kortschak@adelaide.edu.au

Initial Goals

1. Gain familiarity and confidence within the Linux command-line environment
2. Learn how to navigate directories, as well as to copy, move & delete the files within them
3. Look up the name of a command needed to perform a specified task

Background

Command-line tools are the mainstay of analysis of large biological data sets. Good candidate examples for command-line analysis are:

- Manual inspection of fastq, bam & sam files from NGS pipelines
- Automating similar analyses across multiple datasets
- Manipulations of data which are repetitive or laborious to perform manually
- Any analysis that is different from what is available in programs with graphical interfaces (i.e. GUIs).
- Job submission to HPC clusters

Recording all processes as simple scripts also makes an entire analysis more reproducible, as you will have *a record of every procedure you have performed*. It's amazing how often you'll need to revisit something you have performed months ago, and having records of what you've done is immensely useful. From the perspective of an "electronic lab book", this is also very important.

Today we'll explore a few commands to help you gain a little familiarity with some important ones, and to enable you to find help when you're working by yourself. We don't expect you to remember all the commands & options from today. The important thing is to become familiar with the basic syntax for commands, how to put them together, and where to look for help when you're unsure.

Finding your way around

Firstly we need to open a terminal, so **click on the terminal icon** at the bottom of the desktop. There are several variants of the shell, such as the *Bourne-again* shell (i.e. `bash`) and the *C-shell* (please feel free to make all the jokes you can think of). The terminal has a library of commands which are built into it at the time of installing the operating system, and which are part of the Bourne-again Shell, or *bash*. (Historically, it's a replacement for the earlier Bourne Shell, written by Stephen Bourne, so the name is actually a hilarious pun.) We'll explore a few of these commands below, and the words shell or bash will often be used interchangeably with the terminal window. Our apologies to any purists. If you've ever heard of the phrase *shell scripts*, this refers to a series of these commands strung together into a text file which is then able to be executed as a single command.

When you open the terminal on your VM, you'll see the phrase `hub@biohub:~`. This is simply username@computername:, and you may remember you logged on as `hub` during the setup. The tilde (~) represents a shorthand for your home directory, but we'll explore this more later.

Where are we?

When navigating through the folders on any computer, we are all very familiar with clicking through from one folder to another. The folder currently being displayed is usually given in the header of the folder view, so we know where we are currently looking. When using `bash`, we don't click on anything so we need to type commands to change directories, and inspect the contents.

The first thing we need to do is find where we are, when we open `bash`. Type the command `pwd` in the terminal and you will see the output `/home/hub` appear.

```
1 pwd
```

The command `pwd` is what we use for **p**rinting the current (i.e. **w**orking) **d**irectory. Printing in this context means to print some information to the terminal, as opposed to a physical printer. This style of printing harks back to the days when laser printers were not commonplace. Printing information to the terminal itself is what we refer to as printing to the *standard output* or `stdout`.

The directory path that appeared (`/home/hub`) is what will be referred to as your *home directory* for the remainder of the workshop. This is also the information that the tilde (~) represents as a shorthand version, so whenever you see the tilde in a directory path, this is interpreted as meaning `/home/hub`. The *working directory* simply refers to the directory on the computer where you are currently looking, and can be thought of as being the room you are in. We are very familiar with this concept graphically but instead of using a graphical view, we simply have a text-based view.

Looking at the Contents of a Directory

There is another built-in command "`ls`" that we can use to list the contents of a directory. Enter the `ls` command as it is and it will list the contents of the current directory.

```
1 ls
```

Now open your home folder using the icon at the bottom to give the more familiar, graphical view and compare the contents.

The Linux File System

In the above command, the home directory began with a slash, i.e. `/`. On a Linux-based system, this is considered to be the **root directory** of the file system. Windows users would be more familiar with seeing `C:\` as the root of the file system, and this is a very important difference in the two directory structures. Note also that whilst Windows uses the backslash (\) to indicate a new directory, a Linux-based system uses the forward slash (/), or more commonly just referred to simply as "slash", marking another but very important difference between the two.

Although we haven't directly discovered it yet, a Linux-based file system such as Ubuntu or Mac OS-X is also *case-sensitive*, whilst Windows is not. For example, the command `PWD` is completely different to `pwd` and if `PWD` is the name of a command which has been defined in your shell, you will get completely different results than from the intended `pwd` command.

Spaces are also highly important in Linux, so take note of them where-ever they appear in the given commands.

Changing Directories

The command `pwd` is an example of a *command* that is built into the shell. Another built-in command is `cd` which we use to change directory. This changes the directory we are looking in, just like clicking our way through a directory structure in the familiar graphical style we all know well. No matter where we are in a file system, we can move up a directory in the hierarchy by using the command

```
1 cd ..
```

The string "`..`" is the convention for "*one directory above*", whilst a single dot represents the current directory.

Enter the above command and notice that the location immediately to the left of the $ is now given as `/home`. This is also what will be given as the output if we enter the command `pwd`. Note that this given directory is because your personal home folder is within the folder `/home` which contains all of the home folders for all users on the computer. If we now enter `cd ..` one more time we will be in the root directory of the file system. Try this and print the working directory again. As detailed earlier, the output should be the root directory given as `/`.

We can change back to your personal home folder by entering one of either:

```
1 cd /home/hub
```

or

```
1 cd ~
```

or even just

```
1 cd
```

We can also move through multiple directories in one command by separating them with the forward slash "/". For example, we could also get to the root directory from our home directory by typing

```
1 cd ../../
```

Using the above process, return to your home directory `/home/hub`.

Viewing Directories Without Changing Where We Are

Alternatively, we can specify which directory we wish to view the contents of, without having to change into that directory. We simply type the `ls` command, followed by a space, then the directory we wish to view the contents of. To look at the contents of the root directory of the file system (i.e. `/`), we simply add that directory after the command `ls`.

```
1 ls /
```

Here you can see a whole raft of directories which contain the vital information for the computer's operating system. Among them should be the `/home` directory which is one level above your own home directory, and where the home directories for **all users** are

located, as mentioned earlier.

Try to think of two ways we could inspect the contents of the `/home` directory from your own home directory.
`ls /home`
or
`ls ../`

Notice that there are more entries in the `/home` directory. One is your home directory (`/home/hub`), whilst the other is the default home directory that came with the VM (`/home/ubuntu`) before we set up your home directory for today's workshop. Also note that if we ever refer to this higher-level directory, we will write it as `/home`, whereas your personal home directory is always referred to without the preceding forward-slash, and will also be written in normal text font instead of the font we specifically use for code.

Relative vs Absolute Paths

Also note that in your output from `pwd` the resulting path started with the slash, i.e. `/home/hub`. This indicates that it is an **absolute path** as it began with the root of the file system "/". If the slash was missing, it would refer to a sub-directory of the current working directory, and this is what we refer to as a **relative path**. This is an important point which will hopefully become more clear throughout the session.

Another simple example, is that from your home folder (`/home/hub`), the folder `Documents` can be called just using `Documents`. If you are in another folder, you'd probably have to use the full (or absolute) path, which is `/home/hub/Documents`.

A Handy Hint

When working in the terminal, you can scroll through your previous commands by using the up arrow to go backward, and the down arrow to move forward. This can be a big time saver if you've typed a long command with a simple typo, or if you have to do a series of similar commands.

Going Even Deeper

So far, the commands we have used were given either without the use of any subsequent arguments, e.g. `pwd` & `ls`, or with a specific directory as the second argument, e.g. `cd ../` & `ls /home`. Many commands have the additional capacity to specify different options as to how they perform, and these options are often specified between the command name, and the file being operated on. Options are commonly a single letter prefaced with a single dash, or a word prefaced with two dashes. The `ls` command can be given with the option `-l` specified between the command & the directory. This options gives the output in what is known as *long listing* format.

Inspect the contents of your home directory using the long listing format. Please make sure you can tell the difference between the characters `l & 1`.

```
1 ls -l ~
```

The above will give a few lines of output & the first line should be something similar to

```
drwxr-xr-x 2 hub hub 4096 mmm dd hh:mm Desktop
```

where `mmm dd hh:mm` are time and date information.

The important thing to notice is that the word `Desktop` at the end of the line will be coloured **blue**, indicating that it is a directory. The letter '`d`' at the beginning of the initial string of codes `drwxr-xr-x` also indicates this fact. Formally, these letters are known as flags which identify key attributes about each file or directory. We can ignore the fine detail in the rest of these flags until this afternoon (or see the bonus section), but the values `rwx` simply refer to who is able to **r**ead, **w**rite or e**x**ecute the contents of the file or directory. These are very helpful attributes for data security & protection against malicious software.

The entries `hub hub` respectively refer to who is the owner of the directory (or file) & to which group they belong. Again, this information won't be particularly relevant to us today, so we can ignore this until later in our programming careers. In brief, on an Ubuntu system you could have every member of your lab group as an individual user, whilst making every member part of a group. File access permissions can then be applied

to any file based on who created it, or whether they are a member of your lab group. Finally, the value `4096` is the size of the directory structure in bytes, whilst the date & time refer to when the directory was created.

The flags we saw at the beginning of the entries above have a clearly defined structure. The first entry shows the file type and for most common files, this entry will be the "-" seen above. The next entries are three triplets which refer to 1) the file's owner, 2) the group they belong to & 3) all users.

In some other entries you'll come across, the name of the file will be in **green**, indicating that it is a file not a directory. There will also be a '`-`' instead of a '`d`' at the beginning of the initial string of flags. The remainder of the information is essentially the same as for the directories we've already seen.

There are many more options that we could specify to give a slightly different output from the `ls` command. Two particularly helpful ones are the options `-h` and `-R`. We could have specified the previous command as

```
ls -l -h ~
```

This will change the file size to *"human-readable"* format, whilst leaving the remainder of the output unchanged. Try it & you will notice that where we initially saw `4096` bytes, the size is now given as `4.0K`. This can be particularly helpful for larger files, as most NGS files are very large indeed and seeing a file size in gigabytes will be much more informative.

The option `-R` tells the `ls` command to look through each directory recursively. If we enter

```
ls -l -R ~
```

the output will be given in two sections. The first is what we have seen previously, but following that will be the contents of the directory `/home/hub/Desktop`. It should become immediately clear that the output from setting this option can get very large & long depending on which directory you start from. It's probably not a good idea to enter `ls -l -R /` as this will print out the entire contents of your file system.

In the case of the `ls` command we can actually specify all the above options together in the command

```
1 ls -lhR ~
```

This can often save some time, but it is worth noting that not all programmers write their commands in such a way that this convention can be followed. The built-in shell commands are usually fine with this, but many NGS data processing functions do not accept this convention.

Don't Panic!!!

It's easy for things to go wrong when working in the command-line, but if you've accidentally set something running which you need to exit or if you can't see the command prompt, there are some simple options for stopping a process & getting you back on track. Some options to try are:

`Ctrl-c`	kill the current job. This is usually the first port of call when things go wrong. Also see `man kill` or `man killall` for details on how to kill a process.
`Ctrl-d`	end of input. Sometimes `Ctrl-c` doesn't work but this does.

Exploring Commands In More Detail

Most commands we wish to use have a series of options (sometimes called flags) we are able to set. In reality no-one can remember the full suite of available options, so we need to find out how to get this information. In order to help us find what options are able to be specified, every command built-in to the shell has a manual, or a help page which can take some time to get familiar with. These help pages are displayed using the pager known as `less` which essentially turns the terminal window into a text viewer so *we can display text in the terminal window*, but with no capacity for us to edit the text.

To display the help page for `ls` enter the command

```
1 man ls
```

As beforehand, the space between the two is important & in the first word we are invoking the command `man` which then looks for the *manual* associated with the command `ls`. To navigate through the manual page, we need to know a few shortcuts which are part of the `less` pager..

Although we can navigate through the `less` pager using up & down arrows on our keyboards, some helpful shortcuts are:

<enter>	go down one line
<spacebar>	go down one page (i.e. a screenful)
b	go up (i.e. **b**ackwards one page)
<	go to the beginning of the document
>	go to the end of the document
q	exit (i.e. **q**uit the page)

Look through the manual page for the `ls` command. How could we give the directory contents in long listing format, sorted by file size?

```
ls -l -S
```

or

```
ls -lS
```

Many software tools create "*hidden*" files by starting their name with a dot. By default, these files won't be displayed using `ls` How could we make `ls` display these files as well as the main set of visible files.

```
ls -a
```

We can actually find out more about the `less` pager by calling it's own `man` page. Type the command

```
man less
```

and the complete page will appear. This can look a little overwhelming, so try pressing `h` which will take you to a summary of the shortcut keys within `less`. There are a lot of them, so try out a few to jump through the file.

A good one to experiment with would be to search for patterns within the displayed text by prefacing the pattern with a slash. Try searching for a common word like *"the"* or *"to"* to see how the function behaves, then try searching for something a bit more useful, like the word *"move"*.

Accessing Manuals or Help Pages

As well as entering the command `man` before the name of a command you need help with, you can often just enter the name of the command with the options `-h` or `--help` specified. Note the convention of a single hyphen which indicates an individual letter will follow, or a double-hyphen which indicates that a word will follow. Unfortunately the methods can vary a little from command to command, so if one method doesn't get you the manual, just try one of the others.

Sometimes it can take a little bit of looking to find something and it's important to be realise we won't break the computer or accidentally launch a nuclear bomb when we look around. It's very much like picking up a piece of paper to see what's under it. If you don't find something at first, just keep looking and you'll find it eventually.

Try accessing the manual for the command `man` all three ways. Was there a difference in the output depending on how we asked to view the manual?
Yes. Accessing the manual using `man` `man` displayed it a page at a time using `less`. The other two options just printed the entire contents all at once.

Could we access the help page for the command `ls` all three ways?
No. The option `-h` gives the output in human readable format. However, the other two methods work.

Some Useful Commands

So far we have explored the commands `pwd`, `cd`, `ls` & `man` as well as the pager `less`. Inspect the `man` pages for the commands in the following table & fill in the appropriate fields. Have a look at the useful options & try to understand what they will do if specified when invoking the command.

Command	Description of function	Useful options
`man`	Display on-line manual	-k (search for keywords if you don't know the command)
`pwd`	Print working directory, i.e show where you are	none commonly used
`ls`	List contents of a directory	-a, -h, -l, -S, -t, -R
`cd`	Change directory	(scroll down in `man builtins` to find `cd`)
`mv`		-b, -f, -u
`cp`		-b, -f, -u
`rm`		-r (careful...)
`rmdir`		
`mkdir`		-p
`cat`		
`wc`		-l
`head`		-n
`tail`		-n
`echo`		-e
`cut`		-d, -f, -s
`sort`		
`uniq`		-c

Tab auto-complete

A very helpful & time-saving tool in the command line is the ability to automatically complete a command, file or directory name using the <tab > key. Try typing

```
1 ls /home/h <tab>
```

where <tab > represents the tab key.

Notice how the word `hub` is completed automatically! This functionality will automatically fill as far as it can until conflicting options are reached. In this case, there was only one option so it was able to complete all the way to the end of the file path. This enables us to quickly enter long file paths without the risk of typos. Using this trick will save you an enormous amount of time trying to find why something doesn't work. The most common error we'll see today will be mistakes in file paths caused by people not taking advantage of this trick.

Now enter

```
1 ls ~/Do <tab>
```

and it will look like the auto-complete is not working. This is because there are two possibilities & it doesn't know which you want. Hit the tab twice and both will appear in the terminal, then choose one. As well as directory paths, you can use this to auto-complete filenames.

This technique can be used to also find command names. Type in `he` followed by two strike of the <tab > key and it will show you all of the commands that being with the string `he`, such as `head`, `help` or any others that may be installed on your computer. If we'd hit the <tab > key after typing `hea`, then the command `head` would have auto-completed, although clearly this wouldn't have saved you any typing.

Putting It All Together

Now we can use some the above commands to perform something useful.

Creating and Viewing a File

First, let's create a personal directory under `/home/hub` with your first name as the directory name. **Where you see *firstname* below, use your actual firstname.**

```
cd ~
mkdir firstname
```

Now we can change into this directory.

```
cd firstname
```

Create an empty text file. Check the `man` page for `touch` if you're not sure about this line.

```
touch hello.txt
```

We can read it, but it won't have anything in it yet.

```
cat hello.txt
```

Obviously nothing was printed to your terminal in the previous line because the file is empty. Let's write something to the file, then try reading it again. In the following line, the symbol `>>` places the text **at the end** of whatever is already in the file. As the file is empty, this will just write a single line.

```
echo "Hello" >> hello.txt
cat hello.txt
```

We can find a whole lot of information about the file.

```
wc hello.txt
```

In the previous line, what do the three numbers represent? The lines, word & bytes (or characters)

When we added the word `Hello` to our file, we used the symbol >> which actually wrote the word to the end of the file. As the file was completely empty, this placed the word in the first line. This is a VERY handy short-cut for writing information to the end of a file

Now let's add more to the file.

```
echo "It's me" >> hello.txt
```

```
cat hello.txt
wc hello.txt
```

For most of the above commands we could have used the auto-complete feature of the bash terminal. Did you remember this trick?

Copying and Renaming a File

Later today, we're going to look through a file containing a list of words. It's currently on your VM in the folder `/usr/share/dict` and has the name `cracklib-small`. Let's copy this to your `firstname` folder, in your home directory.

```
cp /usr/share/dict/cracklib-small ~/firstname
```

Next we will rename the file. Note, that in bash there is no "rename" tool. Instead we "move" from it's current location to any location we choose, with whatever name we choose. In the following, the location is the same so we are effectively just giving the file a new name.

```
mv ~/firstname/cracklib-small ~/firstname/words
```

We can look at the first 5 lines of the file using

```
head -n5 ~/firstname/words
```

Or we can look at the last 10 lines of the file using

```
tail -n10 ~/firstname/words
```

We could even page through the file using `less`. (Remember to hit `q` to exit the pager.)

```
less ~/firstname/words
```

We can even find how many lines there are in the file by using

```
wc -l ~/firstname/words
```

Did we need to enter the full file path in the above commands, or could we have saved ourselves some effort by changing into the `~/firstname` directory?
Yes, we could have changed into the `~/firstname` directory & just entered the file name.

Summary

Now we have experience in several key tasks which are some of the most common tasks needed for any bioinformatics analysis, or for any general HPC usage.

1. How to access a manual
2. How to create and navigate through directories
3. How to create and view files
4. How to copy, move and rename files

More Tips And Tricks

Primary Author(s):
Stephen Bent, Robinson Research Institute, University of Adelaide
Steve Pederson, Bioinformatics Hub, University of Adelaide
stephen.pederson@adelaide.edu.au

Contributor(s):
Dan Kortschak, Adelson Research Group, University of Adelaide
dan.kortschak@adelaide.edu.au

Text In the Terminal

We can display a line of text in `stdout` by using the command `echo`. The most simple function that people learn to write in most languages is called `Hello World` and we'll do the same thing today.

```
1 echo 'Hello World'
```

That's pretty amazing isn't it & you can make the terminal window say anything you want without meaning it.

```
1 echo 'This computer will self destruct in 10 seconds!'
```

There are a few subtleties about text which are worth noting. Inspect the `man echo` page & note the effects of the `-e` option. This allows you to specify tabs, new lines & other special characters by using the backslash to signify these characters. This is an important concept & the use of a backslash to "escape" normal meaning of a character is very common. In the following, we are using the backslash to escape the normal meanings of the `t` and `n` characters, and they can take on their "special" meaning, such as a `tab` delimiter, or a newline. Try the following three commands & see what effects these special characters have.

```
1 echo 'Hello\tWorld'
2 echo -e 'Hello\tWorld'
3 echo -e 'Hello\nWorld'
```

Redirection Using The Pipe Symbol

A very common process in the command line is to take the output of one process and send it to another. For example, we might want to cut a column from a csv file with categorical information, and then send that field to have the different categories counted by another tool. This is where the concept of `stdout` becomes more important.

By default, most command-line tools print their output to `stdout` but instead, we can send this to another tool using the pipe (`|`) symbol. This is exactly like putting a pipe on the output of one process, and routing it to the *standard input* or `stdin` of another. We can use this to build up a series or chain of processes in a single line. A slightly ridiculous example might be to print lines 96 to 100 of the file `words` by combining `head` and `tail`

```
1 head -n100 ~/firstname/words | tail -n5
```

Note that we didn't give a file to `tail` to work with in the above command. It took the `stdout` from the command `head` as it's input via `stdin`. This is how the pipe will work on virtually all occasions.

As another simple example, we could take some long output from the `ls` command & send it to the pager `less`.

```
1 ls -lh /usr/bin | less
```

Page through the output until you get bored, then hit `q` to quit.

Sending Output To A File

So far, the only output we have seen has been in the terminal, i.e `stdout`. Similar to the pipe command, we can redirect the output of a command **to a file** instead of `stdout`, and we do this using the greater than symbol (>), which we can almost envisage as an arrow.

To see this in action, we'll collect all the words matching our key pattern, & write these to a file.

```
1 cd ~/firstname
2 egrep 'ample$' words > ampleEndings.txt
3 less ampleEndings.txt
```

This is similar to the >> trick we used earlier, but using only a single > symbol will directly overwrite any existing file with the new information. **When using the command line, there is no `Undo` command.** Once we do something, it cannot be undone!

Once you've had a quick look at the file, exit the less pager and delete the file using the `rm` command.

```
1 rm ampleEndings.txt
```

Downloading Data

The terminal makes obtaining data from any publicly available sources very easy using the command `wget`, once you know the location of a file. Let's find a genome to download by opening Firefox & heading to http://www.ncbi.nlm.nih.gov/genome/1099. This is the page for everyone's favourite probiotic `Lactobacillus_acidophilus`. We'll download the *General Feature Format* file with the suffix `.gff` file which contains information about any identified genomic features, and the link to this is in the box at the top of the page. Right-click on the link to the GFF & select *Copy Link Location*. Now head back to the terminal , and if you're not already in your personal directory, i.e. `~/firstname` then change into this directory and type the command `wget`. Once you've typed this command,

right-click and paste the contents of the clipboard to give the following (apologies for the small font):

```
wget <<paste address here>>
```

This will download the selected file into your current working directory. Don't forget to have a look at the `man` page to make sure you understand how the command `wget` works.

The file we have downloaded using `wget` is in simple text format, but it has been compressed for easier transfer. The suffix `.gff` really just denoted a specific structure within the text file so that we know where to look within the file for certain information. Most software that uses a file in this format will manually check for the correct structure first. However, when working in the bioinformatics field, you will often see or require compressed files . This is a common enough concept for even Windows and MAC users, but a few specific compression types are commonly used in bioinformatics. The main differences are in the compression algorithms used, and that is far beyond the scope of what most of us need to worry about.

Working With Compressed Files

The most common types of compression you will see are:

Suffix	Compress	Extract	Useful Arguments
.zip	zip	unzip	-d, -c, -f
.gz	gzip	gunzip zcat	-d, -c, -f
.tar.gz	tar	tar	-x, -v, -f, -z
.bz2	bzip2	bunzip2	

Often, files you download this way will be compressed (tar: tape archive) and archived (zipped). If you see file name suffixes like .tar, .zip, .gz, and/or .bz2, among others, that is what these are. To explore what these command-line options do, please check the `man` pages.

Just to demonstrate how we do this on the command line, we could compress one of our fastq files

```
cd ~/firstname
gzip words
ls
```

How would we now decompress (or extract) this file?
`gunzip words.gz`

How could we have kept our compressed file, along with the decompressed output?
`gunzip -k words.gz`

Why would we use `zcat` instead of `gunzip -c`?
It's less typing, and it's easier to read back

The `GFF` file has been downloaded as a compressed file using the `gzip` program so the first thing we need to do is uncompress it.

```
1 gunzip GCF_000011985.1_ASM1198v1_genomic.gff.gz
```

Don't forget to use tab auto-complete!

The first 7 lines of this file is what we refer to as a *header*, which contains important information about how the file was generated in a standardised format. Many file formats have these structures at the beginning but for our purposes today we don't need to use any of this information so we can move on. Have a look at the beginning of the file to see what it looks like.

```
1 head GCF_000011985.1_ASM1198v1_genomic.gff
```

This file begins with a few lines of header information which begin with one or two hash symbols. These lines contain helpful information about how the file was generated. The remainder of the file contains information about the genomic features themselves, in tab-separated format. Each line represents a genomic feature, and the tab-separated values correspond to columns in a spreadsheet. The first feature is annotated as a *region* in the third field, whilst the second feature is annotated as a *gene*. We'll come back to this file in the next section.

Regular Expressions

Primary Author(s):
Stephen Bent, Robinson Research Institute, University of Adelaide
Steve Pederson, Bioinformatics Hub, University of Adelaide
stephen.pederson@adelaide.edu.au

Contributor(s):
Dan Kortschak, Adelson Research Group, University of Adelaide
dan.kortschak@adelaide.edu.au

Introduction

Regular expressions are a powerful & flexible way of searching for text strings amongst a large document or file. Most of us are familiar with searching for a word within a file using software such as MS Word or Excel, but regular expressions allow us to search for these with more power and flexibility, particularly in the context of genomics. Instead of searching strictly for a word or text string, we can search using less strict matching criteria. For example, we could search for a sequence that is either `AGT` or `ACT` by using the patterns `A[GC]T` or `A(G|C)T`. These two patterns will search for an `A`, followed by either a `G` or `C`, then followed strictly by a `T`. Similarly a match to `ANNT` can be found by using the patterns `A[AGCT][AGCT]T` or `A[AGCT]{2}T`.

Whilst the bash shell has a great capacity for searching a file to matches to regular expressions, this is where languages like *perl* and *python* offer a great degree more power. The commands `awk` & `sed` which we will look at later today also use regular expressions to great effect.

This type of searching is also very common for matching sample names, or extracting key pieces of information from master sample sheets.

The command `grep`

The built-in command which searches using regular expressions in the terminal is `grep`. This function searches a file or input on a line-by-line basis, making patterns split across lines more difficult to find, which is one place that a programming language like Python or Perl would become preferable.. The `man grep` page contains more detail on regular expressions under the `REGULAR EXPRESSIONS` header (scroll down a few pages). As can be seen in the `man` page, the command follows the form

```
grep [OPTIONS] 'pattern' filename
```

The option `-E` is preferable as it it stand for Extended, which we can think of as "Easier". As well as the series of conventional numbers and characters that we are familiar with, we can match to characters with special meaning, as we saw above where enclosing the two letters in brackets gave the option of matching either. The `-E` option opens up the full set of wild-card characters, and can also be called simply by using `egrep` instead of `grep -E`. This is the default version that many of us use.

Special Character	Meaning
\w	match any letter or digit, i.e. a word character
\s	match any white space character, includes spaces, tabs & end-of-line marks
\d	match any digit from 0 to 9
.	matches any single character
+	matches one or more of the preceding character (or pattern)
*	matches zero or more of the preceding character (or pattern)
?	matches zero or one of the preceding character (or pattern)
{x} or {x,y}	matches x or between x and y instances of the preceding character
^	matches the beginning of a line (when not inside square brackets)
$	matches the end of a line
()	contents of the parentheses treated as a single pattern
[]	matches any one of the characters inside the brackets
[^]	matches anything other than any of the characters in the brackets
|	either the string before or the string after the ”pipe” (use parentheses)
\	don’t treat the following character in the way you normally would. This is why the first three entries in this table started with a backslash, as this gives them their “special” properties, whereas placing a backslash before a ‘.’ symbol will enable it to function as an actual dot/full-stop.

Pattern Searching

In this section we’ll learn the basics of using the `egrep` command & what forms the output can take. In your `~/firstname/` directory you can find the file `words`, which we placed there earlier today. This is simply a text file with a different word on every line.

Change into this directory or else `egrep` won’t be able to find the file `words`.

Now let's try a few searches to get a feel for the basic syntax of the command. Using the previous table of special characters, try to describe what you're searching for on your notes **BEFORE** you enter the command. Do the results correspond with what you expected to see?

```
1 egrep 'fr..ol' words
```

```
1 egrep 'fr.[jsm]ol' words
```

```
1 egrep 'fr.[^jsm]ol' words
```

```
1 egrep 'fr..ol$' words
```

```
1 egrep 'fr.+ol$' words
```

```
1 egrep 'cat|dog' words
```

```
1 egrep '^w.+(cat|dog)' words
```

In the above, we were changing the pattern to extract different results from the files. Now we'll try a few different options to change the output, whilst leaving the pattern unchanged. If you're unsure about some of the options, don't forget to consult the `man` page.

```
1 egrep 'louse' words
```

```
1 egrep -w 'louse' words
```

```
1 egrep -wn 'louse' words
```

```
1 egrep -wC2 'louse' words
```

```
1 egrep -c 'louse' words
```

A More Biological Context

Let's return to the `GFF` file we downloaded in the previous section.

Before we use regular expressions, use a previous tool to find how many features you think are contained in this file?

```
wc -l GCF_000011985.1_ASM1198v1_genomic.gff
```

As we know, this file contains 7 header lines beginning with `#`, and there may be more strange lines later in the file. Can you think of a way to count features by telling `egrep` to ignore these lines? You may need to check the manual.

This will give 4355, but the first 7 lines are header lines. To count the non-header lines you could try several things:

```
egrep -vc '^#' GCF_000011985.1_ASM1198v1_genomic.gff
```

or

```
egrep -c '^[^#]' GCF_000011985.1_ASM1198v1_genomic.gff
```

As mentioned above, this file contains multiple features such as *regions*, *genes*, *CDSs*, *exons* or *tRNAs*. If we wanted to find how many regions are annotated in this file we could use the processes we've learned above:

```
1 egrep -c `region' GCF\_000011985.1\_ASM1198v1\_genomic.gff
```

If we wanted to count how many genes are annotated, the first idea we might have would be to do something similar using a search for the pattern `'gene'`.

Do you think this is the number of genes? Try searching for the number of coding DNA sequences using the same approach (i.e. CDS) & then add the two numbers? Is this more than the total number of features we found earlier? Can you think of a way around this using regular expressions? Note that some of the occurrences of the word *gene* appears in many lines which are not genes. We need to restrict the search to one of the tab-separated fields by including a white-space character in the search. The command:
`egrep -n '.+\s.+\sgene\s' GCF_000011985.1_ASM1198v1_genomic.gff | wc -l`
will give a much different result as now we are searching for the word gene surrounded by white-space, after at least two tab delimiters.

Alternatively, there is a command `cut` available. Call the manual page (`man cut`) and inspect the option `-f`. The information about the types of features annotated is in the third field. Try to think of a way of searching this field alone. `cut -f3 GCF_000011985.1_ASM1198v1_genomic.gff | grep -c 'gene'`
Or even more accurately `cut -f3 -s GCF_000011985.1_ASM1198v1_genomic.gff | egrep -c 'gene'`
although it gives the same results in this instance

A similar question may be how many *types* of features are in this file? The commands `cut`, along with `sort` and `uniq` may prove to be useful when answering this. `cut -f3 -s GCF_000011985.1_ASM1198v1_genomic.gff | sort | uniq | wc -l`

Or we could ask how many of each type of feature are there in this file. `cut -f3 -s GCF_000011985.1_ASM1198v1_genomic.gff | sort | uniq -c`

Pattern Matching in DNA sequences

Let's try searching for some DNA patterns using a fasta file. We can download the genomic sequence for the same organism, so once again head to the same page as last time, but this time copy the link to the **genome**.

```
wget <<paste address here>>
gunzip GCF_000011985.1_ASM1198v1_genomic.fna.gz
```

Let's try using some of the tricks we've learned above.

How many times does the sequence `GAAAGGATTA` appear in the genome?
`egrep -c 'GAAAGGATTA' GCF_000011985.1_ASM1198v1_genomic.fna`
Gives 3
How many times does this appear, but with two **or more** Ts before the final A?
`egrep -c 'GAAAGGATT+A' GCF_000011985.1_ASM1198v1_genomic.fna`
Gives 4
How many times does this appear, but with two **or more** Ts but with a G or C following the final T?
`egrep -c 'GAAAGGATT+[Â]' ...` Gives 7. Clearly, they could also specify (GC) instead of [Â].

The Tools sed & awk

Primary Author(s):
Stephen Bent, Robinson Research Institute, University of Adelaide
Steve Pederson, Bioinformatics Hub, University of Adelaide
stephen.pederson@adelaide.edu.au

Contributor(s):
Dan Kortschak, Adelson Research Group, University of Adelaide
dan.kortschak@adelaide.edu.au

sed: The Stream Editor

One very useful command in the terminal is `sed`, which is short for *stream editor*. Instead of the `man` page for `sed` the `info sed` page is larger but a little easier to digest. This is a very powerful command which can be a little overwhelming at first. If using this for your own scripts & you can't figure something out, remember "Google is your friend" & sites like www.stackoverflow.com are full of people wrestling with similar problems to you. You can be certain you're not the first person to be stumped by a problem & these are great places to start looking for help. Even advanced programmers use Google & Stack Overflow to find solutions.

For today, there are two key `sed` functionalities that we want to introduce.

1. Using `sed` to alter the contents of a file/input;
2. Using `sed` to print regions of a file

Altering a file or other input

`sed` uses *regular expressions* that we have come across under the `grep` section, and we can use these to replace strings or characters within a text string. The command works in the form

```
sed SCRIPT INPUT
```

and the script section is where all the action happens. Input can be given to `sed` as either a file, or just as a text stream via `stdin` using the *pipe* symbol that we have already introduced.

In the following example the script section begins with an 's' to indicate that we are going to make a substitution. The beginning of the first pattern (i.e. the *regexp* we are searching for) is denoted with the backslash, with the identical delimiter indicating the replacement pattern, and this is in turn completed with the same delimiter. Try this simple example from the link http://www.grymoire.com/Unix/Sed.html which is a very detailed & helpful resource about the usage `sed`. Here we are sending the input to the command via the pipe, so no 'INPUT' section is required:

```
1 echo Sunday | sed 's/day/night/'
```

Here you are passing `sed` the string Sunday, and `sed` takes day and turns it into night. `sed` will only replace the first instance of the string on any line, so try:

```
1 echo Sundayday | sed 's/day/night/'
```

Note that it only replaced the first instance of day and left the second. However, you can make it 'global', where it switches every instance by using the 'g' option at the end of the pattern like this:

```
1 echo Sundayday | sed 's/day/night/g'
```

You can also 'capture' parts of the pattern in parentheses and access that in the second part of the regular expression (what you are switching to) using \1, \2, etc., to denoted the number of the captured string, in the order they were captured. If you want to match 'ATGNNNTGA', where N is any base, and just output these three bases you could try the following:

```
1 echo 'ATGCCAGTA' | sed -r 's/ATG(.{3})GTA/\1/g'
```

Clearly, we have just given this command a sequence so we know exactly what to expect. However, hopefully this demonstrates the concept of extracting a subset of the sequence.

Or if we needed to replace those three bases with an expanded repeat of them, you could do the following where we capture the undefined string between `ATG` & `GTA`, and expand it:

```
1 echo 'ATGCCAGTA' | sed -r 's/ATG(.{3})GTA/ATG\1\1\1GTA/g'
```

The \1 take the contents of the first parenthesis and uses it in the substitution, even though you don't know what the bases are. Note that the '-r' option was set for these operations, which turns on extended regular expression capabilities. This can be a powerful tool & multiple parentheses can also be used:

```
1 echo 'ATGCCAGTA' | sed -r 's/(ATG)(.{3})(GTA)/\3\2\2\1/g'
```

In this last command we switched the order of the first & last triplet, and expanded the middle unknown string twice. Note how quickly this starts to look confusing though! Taking care to be clear when writing these types of procedures can be an important idea when you have to go back & re-read your code a year or two later. (Yes this will happen a lot!!!)

The use of backslashes to delineate each section of the script used by `sed` is the most common convention, but we are not restricted to it. We could have uses any 'wild-card' type character to follow the 's', such as '*' ':' or '%' although this must be consistent across all sections of the script, and should only be used if the backslash itself is part of the search or replacement string.

Displaying a region from a file

The command `sed` can also be used to replicate some of the functionality of the `head` & grep commands, but with a little more power at your fingertips. By default `sed` will print the entire input stream it receives, but setting the option '-n' will turn this off. Try this by adding an 'n' immediately after the '-r' in one of the above lines & you will notice you receive no output. This is useful if we wish to restrict our output to a subset of lines within a file, and by including a 'p' at the end of the script section, only the section matching the results of the script will be printed.

Make sure you are in the correct directory & we can look through the `GFF` file again.

```
sed -n '1,10p' GCF_000011985.1_ASM1198v1_genomic.gff
```

This will print the first 10 lines, like the `head` command will by default. However, we could now print any range of lines we choose. Try this by changing the script to something interesting like '101,112p'.

We could also restrict the range to specific lines by using the `sed` increment operator '~'.

```
sed -n '1~10p' GCF_000011985.1_ASM1198v1_genomic.gff
```

This will print every 10th line, beginning at the first.

We can also make sed operate like `grep` by making it only print the lines which match a pattern.

```
sed -rn '/TGCAGGCTCT.+(GA){2}.+/ p' pair1.fq
```

Note however, that the line numbers are not present in this output.

How would we use `sed` to extract the lines from the genome sequence which match the pattern `ATGC.+ACAA.*`? What text does this pattern represent?

```
sed -rn '/ATGC.+ACAA.*/p' GCF_000011985.1_ASM1198v1_genomic.fna
```

How would we use the pipe to extract the unspecified sequence between the ATGC and the ACAA?

```
sed -rn '/ATGC.+ACAA.*/p' GCF_000011985.1_ASM1198v1_genomic.fna | sed -r 's/ATGC(.+)ACAA.*/\1/g'
```

Some Important Programming Concepts

Before moving on to awk, we need to quickly recap two of the most widely used techniques in programming:

1. The `for` loop
2. Logical tests using an `if` statement

For Loops

A `for` loop is what we use to cycle through an input one item at a time. As a simple example, we could print each number from a set of numbers.

```
1 for i in 1 2 3; do (echo -e $i); done
```

In the above code, the fragment before the semi-colon asked the program to cycle through the values 1, 2 & 3, letting the variable 'i' take each value in order of appearance. First i = 1, then i = 2 & finally i = 3. If you're wondering why we chose 'i', it just seemed like a sensible choice for an integer. We simply needed to choose a name for a *variable* which we would pass the values to.

After assigning each value to `i`, was the instruction on what to do for each value. Note that the value of the variable 'i' was *prefaced by the dollar sign ($). This is how the bash shell knows it is a variable, not the letter 'i'.* The command `done` then finished the `do` command. All commands like `do`, `if` or `case` have completing statements, which respectively are `done`, `fi` & `esac`.

Another important concept which was glossed over in the previous paragraph is that of a *variable*. These are essentially just '*placeholders*' which have a value that can change (hence the name). In the above loop, the same operation was performed on the variable `i`, but the value changed from 1 to 2 to 3. Variables in shell scripts can hold numbers or text strings and don't have to be formally defined as in some other languages.

An alternate approach could be to make a breathtaking claim about some files. Here we'll use the variable called 'f', which seems sensible for a filename.

```
1 cd ~/firstname
2 for f in $(ls); do (echo -e "I can see the file $f"); done
```

Note, that we've also assigned the output from the command `ls` to this variable, by using the $() syntax. The use of double quotes for the `echo` command also allows us to refer to the values held by f. Single quotes at this point would only return the characters '$f'.

If Statements

If statements are those which only have a binary 'yes' or 'no' response, or more correctly a `TRUE/FALSE` response. For example, we could specify things like:

- `if` (i> 1) then `do` something, or
- `if` (fileName==bob.txt) then `do` something else

Notice that in the second if statement, there was a double equals sign. This is the programmers way of saying *compare* the first argument with the second argument. A single equals sign is generally interpreted by a program as *assign* the value of the first argument to be the second argument. This use of 'double operators' is very common, notably you will see && to represent the command '*and*', and || to represent '*or*.' A final useful trick to be aware of is the use of an exclamation mark to reverse a command. A good example of this is the use of the command '!=' as the representation of *not equal to* in a logical test.

awk: A command and a language

Moving on to `awk`, this is a very useful tool which can be used either as a command, as well as functioning as it's own language. We'll just use it as a command today, and it is extremely useful for dealing with tab- or comma-separated files, such as we often see in biological data.

The basic structure of an `awk` command is:

```
awk `/<pattern>/' file
```

`awk` will then search the file and output any line containing the regular expression pattern (kind of like `grep`). With `awk`, you can also do:

```
awk `\{<code>\}' file
```

where you can put a program, or set of instructions in the curly braces. In the code, you can specify values from different columns of the file by using the numbers $1, $2, etc., (or you can use $0 for the whole line). Values can also be returned in the output by using the command `print` followed by the field number.

We have that *.gff file and we've already looked at it at little, so let's pull out some particular features! Make your terminal as wide as the screen, then change into the appropriate directory & enter

```
1 awk '{if ($3=="rRNA") print $0}' GCF_000011985.1_ASM1198v1_genomic.gff
```

Here we've specified that the third field must be an rRNA, so this will give us all of the

ribosomal RNAs annotated in the file.

We could make it a little more complex and just look for genes in a given region. **In the following line, the symbol '\' has been placed here to indicate it is a single line, extending beyond the width of the page. Do not enter this character!**

```
1 awk '{if ( ($3=="gene") && ($4 > 10000) && ($4 < 20000) ) print $0}' \
    GCF_000011985.1_ASM1198v1_genomic.gff
```

In the above code, `$3==''gene''` asks for the entry in the third field to be "gene." The next two fragments request for the values in the fourth field (i.e. $4) to be between 10000 & 20000. Thus we have found all the features annotated as genes in a 10,000bp region of this genome. Notice that each these three commands were enclosed in a pair of brackets within an outer pair of brackets. This gave a command of the form:
`( (Condition1) && (Condition2) && (Condition3) )`
After this came the fragment `print $0` which asked `awk` to print the entire line if the 3 conditions are true. You've just written (& hopefully understood) a computer program!

Another example (what does this do?):

```
1 awk '{if (($5 - $4 > 1000) && ($3 == "gene")) print $0}' \
    GCF_000011985.1_ASM1198v1_genomic.gff
```

If you don't want to output all of the columns, you can specify which ones to output. While we're at it, let's save the output as a file:

```
1 awk '{if (($5 - $4 > 1000) && ($3 == "gene")) print $1, $2, $4, $5, \
    $9}' GCF_000011985.1_ASM1198v1_genomic.gff > awkout.txt
```

The command `awk` has a pretty serious set of in-built commands which can be used in the code sections as above. Although it looks a little overwhelming, there is a detailed page http://www.grymoire.com/Unix/Awk.html which gives a rundown on the full capabilities of the language. One command that we may find helpful is `length`, which counts the number of characters in a line.

In your `trainingData` directory you'll find another file `seqData.fastq`. This is a small set of reads from some mRNA pulled down by Immunoprecipitation, and they've been processed to have varying lengths.

Can you think of a way to combine `sed` and `awk` to create a new file `lengths.txt` which contains the length of every read in the file `seqData.fastq`?

```
sed -n '2~4p' seqData.fastq | awk '{print length($1)}' > lengths.txt
```

Writing Scripts

Primary Author(s):
Stephen Bent, Robinson Research Institute, University of Adelaide
Steve Pederson, Bioinformatics Hub, University of Adelaide
stephen.pederson@adelaide.edu.au

Contributor(s):
Dan Kortschak, Adelson Research Group, University of Adelaide
dan.kortschak@adelaide.edu.au
Jimmy Breen, Robinson Research Institute & Bioinformatics Hub, University of Adelaide
jimmy.breen@adelaide.edu.au

Sometimes we need to perform repetitive tasks on multiple files, or need to perform complex series of tasks and writing the set of instructions as a script is a very powerful way of performing these tasks. They are also an excellent way of ensuring the commands you have used in your research are retained for future reference. Keeping copies of all electronic processes to ensure reproducibility is a very important component of any research. Writing scripts requires an understanding of several key concepts which form the foundation of much computer programming, so let's walk our way through a few of them.

Shell Scripts

Now that we've been through just some of the concepts & tools we can use when writing scripts, it's time to tackle one of our own where we can bring it all together.

Every bash script begins with what is known as a *shebang*, which we would commonly recognise as a hash sign followed by an exclamation mark, i.e `#!`. This is immediately followed by `/bin/bash`, which tells the interpreter to run the command `bash` in the directory `/bin`. (This is actually where the program `bash` lives on an ubuntu system.) This opening sequence is vital & tells the computer how to respond to all of the following commands. As a string this looks like:

```
#!/bin/bash
```

The hash symbol generally functions as a comment character in scripts. Sometimes we can include lines in a script to remind ourselves what we're trying to do, and we can preface these with the hash to ensure the interpreter doesn't try to run them. It's presence as a comment here, followed by the exclamation mark, is specifically looked for by the interpreter but beyond this specific occurrence, comment lines are generally ignored by scripts & programs.

Some Example Scripts

Let's now look at some simple scripts. These are really just examples of some useful things you can do & may not really be the best scripts from a technical perspective. Hopefully they give you some pointers so you can get going

A Simple Example to Start

Don't try to enter these commands directly in the terminal!!! They are designed to be placed in a script which we will do after we've inspected the contents of the script. First, just have a look through the script & make sure you understand what the script is doing.

Also remember that any long lines of code may be automatically broken into new lines on your page by the '\' character. We don't want to enter this character when we create our script. Note that the line numbers on the left of the code don't change when this happens, e.g. line 9.

Before we go any further, have a look at the following script.

```
#!/bin/bash

# First we'll declare some variables with some text strings
ME='Put your name here'
MESSAGE='This is your first script'

# Now well place these variables into a command to get some output
echo -e "Hello ${ME}\n${MESSAGE}\nWell Done!"
```

Firstly, you may notice some lines that begin with the # character. These are *comments* which have no impact on the execution of the script, but are written so you can understand what you were thinking when you wrote it. If you look at your code 6 months from now, there is a very strong chance that you won't recall exactly what you were thinking, so these comments can be a good place just to explain something to the future version of yourself. There is a school of thought which says that you write code primarily for humans to read, not for the computer to understand.

In the above script, there are two variables. Although we have initially set them to be one value, they are still variables. What are their names? ME & MESSAGE

First we'll create an empty file which will become our script. We'll give it the suffix `.sh` as that is the common convention for bash scripts.

```
1 cd ~/firstname
2 touch wellDone.sh
```

Now using the text editor *nano*, enter the above code into this file *setting your actual name as the ME variable*, and save it by using `Ctrl+o`, which is indicated as `^O` in the nano screen.

```
nano wellDone.sh
```

Once you're finished, you can exit the `nano` editor by hitting `Ctrl+x`.

Another coding style which can be helpful is the enclosing of each variable name in curly braces every time the value is called. Whilst not being strictly required, this can make it easy for you to follow in the future when you're looking back. Variables have also been names using strictly upper-case letters. This is another optional coding style, but can also make things clear for you as you look back through your work. Most command line tools use strictly lower-case names, so this is another reason the upper-case variable names can be helpful.

Unfortunately, this script cannot be executed yet but we can easily enable execution of the code inside the script. If you recall the flags from earlier which denoted the read/write/execute permissions of a file, all we need to do is set the execute permission for this file. First we'll look at the files in the folder using `ls -l` and note these triplets should be `rw-` for the user & the group you belong to. To make this script executable, enter the following in your terminal.

```
cd ~/firstname
chmod +x wellDone.sh
ls -l
```

Notice that the third flag in the triplet has now become an `x`. This indicates that we can now execute the file in the terminal. As a security measure, Linux doesn't allow you to execute a script from within the same directory so to execute it enter the following:

```
./wellDone.sh
```

Making a Small Change

Now let's change the variable `ME` in the script to read as

```
ME=$1
```

and save this as `wellDone2.sh`. (You may like to create this first using `cp`) You'll now need to set the execute permission again.

```
chmod +x wellDone2.sh
```

This time we have set the script to *receive input from stdin* (i.e. the terminal), and we will need to supply a value, which will then be placed in the variable `ME`. Choose whichever random name you want and enter the following

```
./wellDone2.sh Boris
```

As you can imagine, this style of scripting can be useful for iterating over multiple objects. A trivial example, which builds on a now familiar concept would be to try the following.

```
for n in Boris Fred; do (./wellDone2.sh $n); done
```

A more complicated script

Here's a more complicated script with some more formal procedures. This is a script which will extract only the CDS features from the .gff file we have been working with, and export them to a separate file. Look through each line carefully & write down your understanding of what each line is asking the program to do.

```
#!/bin/bash

# Declare some helpful variables
FILEDIR=~/firstname
FILENAME=GCF_000011985.1_ASM1198v1_genomic.gff
OUTFILE=GCF_000011985.1_ASM1198v1_CDS.txt

# Make sure the directory exists
if [ -d ${FILEDIR} ];
  then
    echo Changing to ${FILEDIR}
    cd ${FILEDIR}
  else
    echo Cannot find directory ${FILEDIR}
    exit 1
fi

# If the file exists, extract the important CDS data
if [ -a ${FILENAME} ];
  then
    echo Extracting CDS data from ${FILEDIR}/${FILENAME}
    echo "SeqID Source Start Stop Strand Tags" > ${OUTFILE}
    awk '{if (($3=="CDS")) print $1, $2, $4, $5, $7, $9}' ${FILENAME} >> \
       ${OUTFILE}
  else
    echo Cannot find ${FILENAME}
    exit 1
fi
```

Notice that this time we didn't require a file to be given to the script. We defined it within the script, as we did for the output file.

The directory & file checking stages were of the form if [...]. This is a curious command that checks for the presence of something. The options -d & -a specify a directory or file respectively.

Will the above script generate a tab, comma or space delimited text file?
It will be space delimited. We could have specified tab delimited by inserting "\t" between each field.

Open the `gedit` text editor & save the blank file in your directory as *extract_CDS.sh*. Now write this above script into the editor, but *taking care to use the directory where you have the .gff file stored in the appropriate place.* Once you have written the script, save it & close it. Now make it executable and run it.

Moving towards High Performance Computing

High Performance Computing

In current genomics era, where we regularly work with large datasets, the amount of resources available on desktop computers are often insufficient to enable your script finish in a reasonable time. For these datasets, it maybe useful to work on a high performance computing system, which will enable your data or command to be run simultaneously on $>$ 8 threads, i.e. in parallel. It is possible to gain access to large computing resources through the University's phoenix HPC https://www.phoenix.adelaide.edu.au, enabling analysis of large datasets efficiently.

Having many users on one machine at one time also means that there needs to be a system which determines who runs what and when. Phoenix uses a "queuing" system, whereby users submit jobs to a queue, and then executed when the appropriate resources on the machine become available. The command is executed using a slurm script (shown below). This script contains extra parameters such as:

- `-n`: The number of threads to allow
- `--time`: The maximum time it is allowed to takes to completion
- `--mem`: The amount of memory to allocate to the job

```
#!/bin/bash
#SBATCH -p batch
#SBATCH -N 1
#SBATCH -n 8
#SBATCH --time=20:00:00
#SBATCH --mem=20GB

NAME=$1
HISAT2_HOME=~/hisat2-2.0.4
REF=~/grcm38_snp_tran/genome_snp_tran
SAM=${NAME}.sam

# NB: For single end reads only
${HISAT2_HOME}/hisat2 --dta -p 4 -q -x ${REF} -U ${NAME}.fastq -S ${SAM}
```

We don't need to write this script. It is included here as a simple example of a real world script as used by researchers in bioinformatics. This script may look a little intimidating at first, but slowly work your way through each line & try to understand what each line is specifying. This particular script will allow users to submit a single `fastq` file to be aligned to a reference genome using the alignment too `hista2`.

Space for Personal Notes or Feedback

Space for Personal Notes or Feedback

Space for Personal Notes or Feedback