

# Module 3

# Introduction to Computer Programming

## Introduction

Recent advances in high-throughput biology have transformed modern biology into an area overflowing with large datasets. These datasets are pushing the limits of desktop applications, and we are now finding that using an Excel spreadsheet is simply not enough. In addition, we often find ourselves in situations where we have to repeat the same task or procedure for every item in our dataset. This repetitive process can often be tedious and time consuming. By learning basic programming we can write small programs, called scripts that allow us to easily manage these large datasets and to automate repetitive tasks. Learning to program can be a daunting task, but it is also extremely worthwhile. Not only will you improve your research, you will also learn new concepts and have a lot of fun!

## Aims

The aim of this module is to present you with an introduction to programming, guiding you through useful Linux commands and essential programming concepts. The first part of this module introduces you to more advanced Linux concepts, illustrating these concepts with meaningful examples and exercises. The second part of the module introduces the notion of shell scripting and demonstrates how to save useful Linux commands for future use so that they can be used over and over again. Finally, the third part of the module introduces the basic concepts of the Perl programming language which can be used for more advanced analysis when shell scripting is not quite sufficient.

Many of the examples and exercises throughout this module are designed with sequence manipulation tasks in mind, and at the end of this module we hope you will be able to write some small scripts to help you with your research. Like all modules, 'if you don't understand, please ask'. No-one is going to become a programming expert in a few hours; the overall purpose of the module is to give you a taste of what writing small scripts can do to automate and accelerate your analysis.

If at the end of this module you would like to learn more about programming, we have provided a list of useful resources for further reading.

## Advanced Linux

Increasingly, the output of biological research exists as *in silico* data, usually in the form of large text files. Linux is particularly suitable for working with such files and has several powerful and flexible commands that can be used to process and analyse this data. One advantage of learning how to use Linux is that many of the commands can be combined in an almost unlimited fashion. So if you can learn just six Linux commands, you will be able to do a lot more than just six things.

You have already been introduced to some basic Linux commands including `ls`, `pwd` and `cp`. Linux contains hundreds of commands, but to conduct your analysis you will probably only need 10 or so to achieve most of what you want to do. In the following exercises we will introduce you to more of these Linux commands and provide examples of how they can be used in bioinformatics analysis.

### Exercises : More Linux commands

To begin: open a terminal window, move into the `Module_3_Programming` directory, then the `Linux` directory ( `cd Module_3_Programming/Linux` ) and follow the instructions below.

#### BED files

We will be using a BED file in the examples that follow. A BED file (.bed) is a tab-delimited text file that defines a set of features. To see the format of a BED file you can view it by running:

```
cd Module_3_Programming/Linux
```

```
less Pfalciparum.bed
```

BED lines have three required fields and nine additional optional fields. The first three required BED fields are:

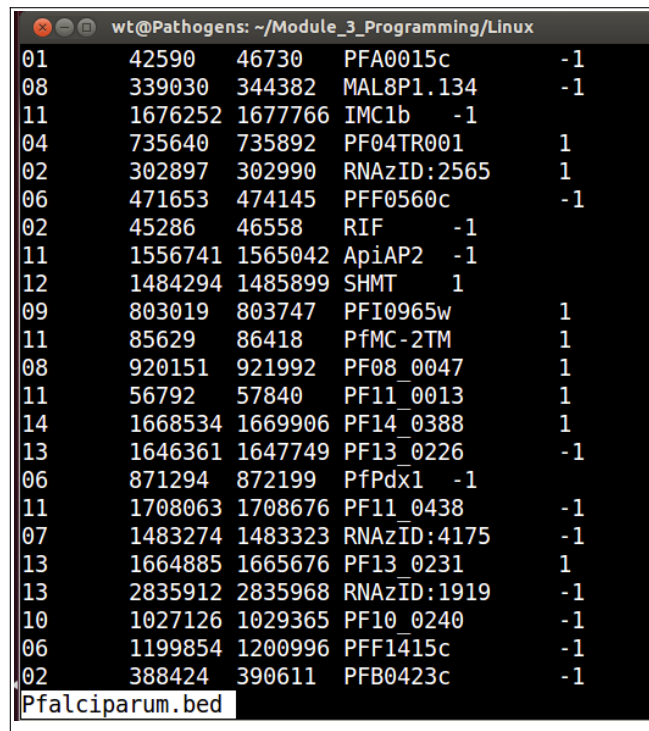
- **chrom** - The name of the chromosome or scaffold
- **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
- **chromEnd** - The ending position of the feature in the chromosome or scaffold.

Other additional optional BED fields include name, score, and strand. For more information on BED files see: <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>

#### Getting help man

To obtain further information on any of the Linux commands listed below you can use the `man` command. For example, to get a full description and examples of how to use the `sort` command type the following in a terminal window.

```
man sort
```



Chromosome	Start	End	Gene Name	Strand
01	42590	46730	PFA0015c	-1
08	339030	344382	MAL8P1.134	-1
11	1676252	1677766	IMC1b	-1
04	735640	735892	PF04TR001	1
02	302897	302990	RNAzID:2565	1
06	471653	474145	PFF0560c	-1
02	45286	46558	RIF	-1
11	1556741	1565042	ApiAP2	-1
12	1484294	1485899	SHMT	1
09	803019	803747	PFI0965w	1
11	85629	86418	PfMC-2TM	1
08	920151	921992	PF08_0047	1
11	56792	57840	PF11_0013	1
14	1668534	1669906	PF14_0388	1
13	1646361	1647749	PF13_0226	-1
06	871294	872199	PfPdx1	-1
11	1708063	1708676	PF11_0438	-1
07	1483274	1483323	RNAzID:4175	-1
13	1664885	1665676	PF13_0231	1
13	2835912	2835968	RNAzID:1919	-1
10	1027126	1029365	PF10_0240	-1
06	1199854	1200996	PFF1415c	-1
02	388424	390611	PFB0423c	-1
Pfalciparum.bed				

**Figure 1** The first part of the Pfalciparum.bed file

**sort** - Sort values in a file or piped input

This command lets you sort the contents of the input. When you sort the input, lines with identical content end up next to each other in the output, which can then be fed to `uniq` (see below) to count the number of unique lines in the input.

To sort the contents of the BED file type the following command:

```
sort Pfalciparum.bed [ENTER]
```

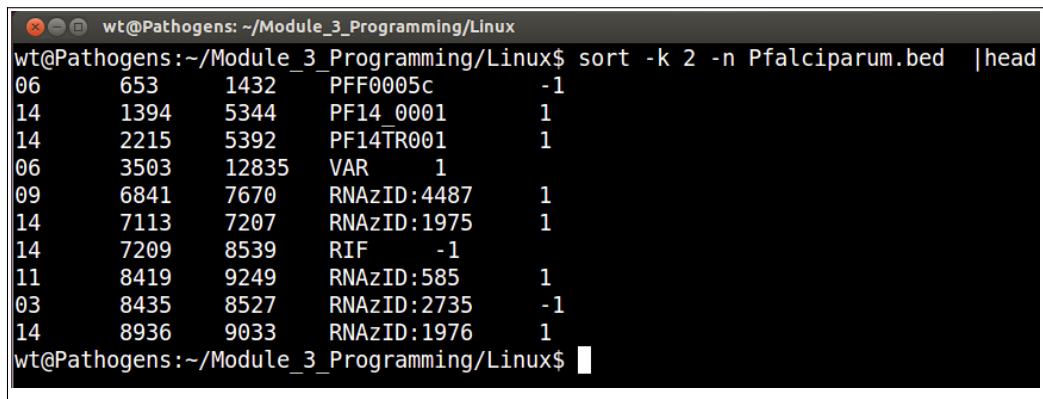
To sort the contents of the BED file on position type the following command:

```
sort -k 2 -n Pfalciparum.bed [ENTER]
```

The sort command can sort by multiple columns e.g. 1st column and then 2nd column by specifying successive `-k` parameters in the command. Modify the previous command to sort the BED file on chromosome and then gene position.

What does the `-n` option do?

**Hint:** use the command `man sort`.



```

wt@Pathogens: ~/Module_3_Programming/Linux
wt@Pathogens:~/Module_3_Programming/Linux$ sort -k 2 -n Pfalciparum.bed | head
06      653      1432    PFF0005c      -1
14     1394     5344    PF14_0001       1
14     2215     5392    PF14TR001       1
06     3503     12835    VAR           1
09     6841     7670    RNAzID:4487     1
14     7113     7207    RNAzID:1975     1
14     7209     8539    RIF           -1
11     8419     9249    RNAzID:585      1
03     8435     8527    RNAzID:2735    -1
14     8936     9033    RNAzID:1976     1
wt@Pathogens:~/Module_3_Programming/Linux$

```

**Figure 2** Sorting on a column. Since there is a lot of output head is used to return the 1st 10 lines.

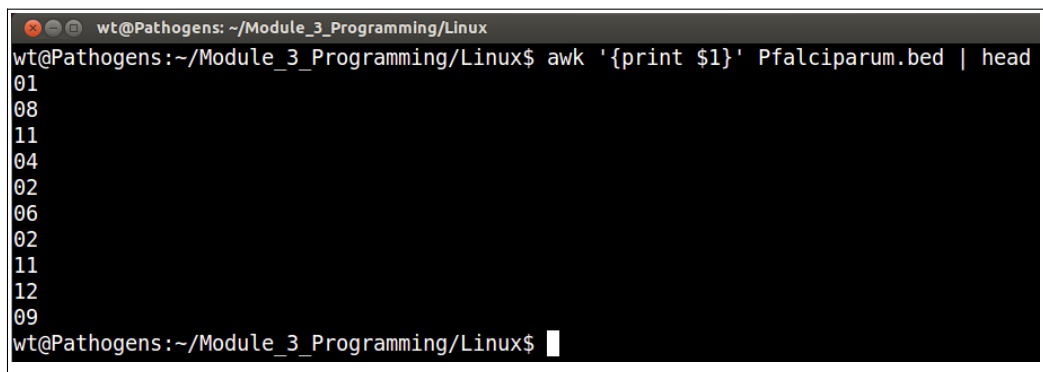
### **awk** - Pattern scanning and processing language

This command lets you manipulate the input text, making it very useful for chopping out the bits of a file that your interested in and outputting them to another file or command.

To print out the first column of the BED file, enter the following command:

```
awk '{print $1}' Pfalciparum.bed [ENTER]
```

This is a very powerful and complex command, and is often used in conjunction with sed which will be talked about later.



```

wt@Pathogens: ~/Module_3_Programming/Linux
wt@Pathogens:~/Module_3_Programming/Linux$ awk '{print $1}' Pfalciparum.bed | head
01
08
11
04
02
06
02
11
12
09
wt@Pathogens:~/Module_3_Programming/Linux$

```

**Figure 3** Extracting the first column with awk. Since there is a lot of output head is used to return the 1st 10 lines.

**uniq** - extract unique lines from a file or piped input

The **uniq** command is usually used in combination with **sort** to count unique values in the input.

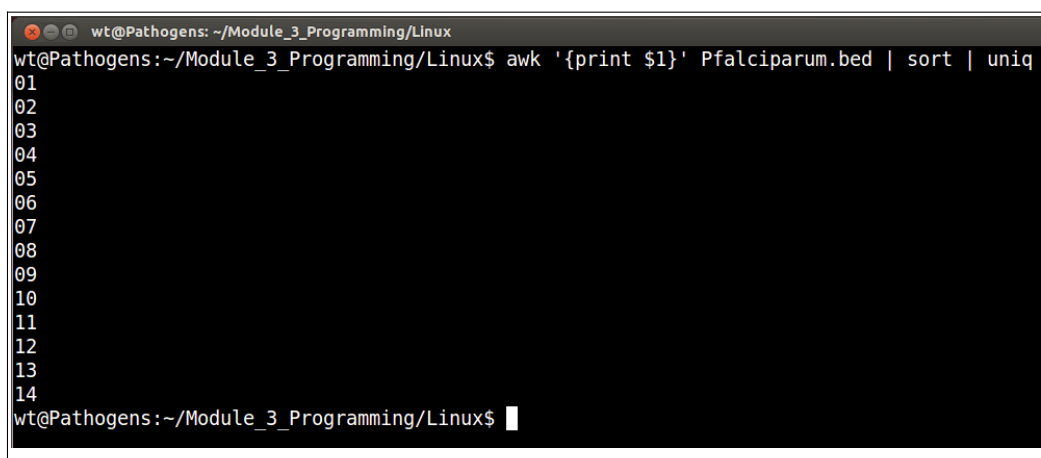
To get the list of chromosomes in the BED file type the following command.

```
awk '{print $1}' Pfalciparum.bed | sort | uniq [ENTER]
```

How many chromosomes are there?

Now modify the previous command to count the number of features per chromosome.

**Hint:** use the **man** command to look at what are the options for the **uniq** command.



```
wt@Pathogens: ~/Module_3_Programming/Linux
wt@Pathogens:~/Module_3_Programming/Linux$ awk '{print $1}' Pfalciparum.bed | sort | uniq
01
02
03
04
05
06
07
08
09
10
11
12
13
14
wt@Pathogens:~/Module_3_Programming/Linux$
```

**Figure 4** Counting the chromosomes in the BED file

**find** - Finds files matching an expression

The **find** command will search the directory and all sub directories and return a list of files. It can filter the files for you if you tell it the name of the file your searching for. A **\*** denotes a wildcard which can stand for any character(s).

To find all files in the current directory and all sub directories:

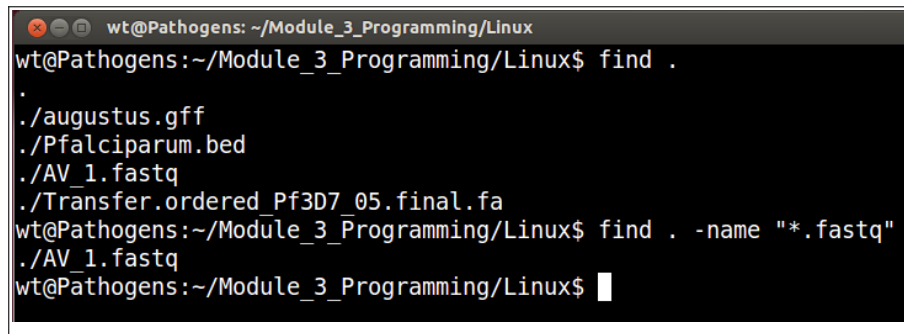
```
find . [ENTER]
```

To find all fastq files in the current directory and all sub directories:

```
find . -name "*.fastq" [ENTER]
```

Modify the command above to find all bed files in the current directory.

Use the **find** command to find all fastq files in the **Module\_3\_Programming** directory.



```

wt@Pathogens: ~/Module_3_Programming/Linux
wt@Pathogens:~/Module_3_Programming/Linux$ find .
.
./augustus.gff
./Pfalci-parum.bed
./AV_1.fastq
./Transfer.ordered_Pf3D7_05.final.fa
wt@Pathogens:~/Module_3_Programming/Linux$ find . -name "*.fastq"
./AV_1.fastq
wt@Pathogens:~/Module_3_Programming/Linux$

```

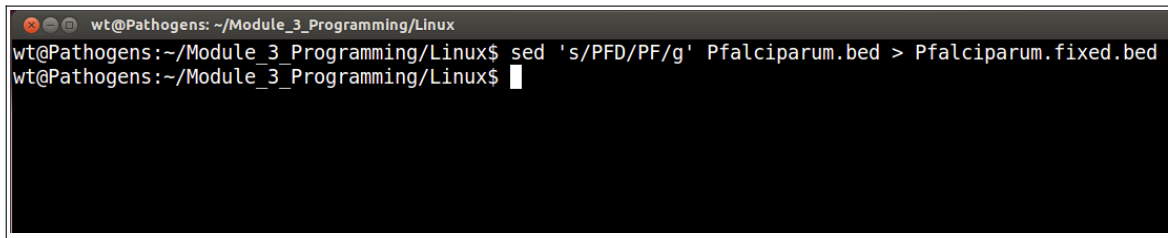
**Figure 5** Finding files in the current directory

**sed** - For filtering and transforming text

sed is a powerful command line tool for editing the contents of a file and outputting the modified contents to another file. For example, you can find all occurrences of a particular string of text on each line of a file and replace them with another string of text. For example, we have found a mistake in the feature names in our BED file where we need to replace all feature names that begin with PFD with PF. To do this use the command:

```
sed 's/PFD/PF/g' Pfalci-parum.bed > Pfalci-parum.fixed.bed [ENTER]
```

Now modify the command to change all 'VAR' features to be called 'VAR\_gene' and all 'RIF' features to be called 'RIFIN\_gene'. Write the final output to a file called Pfalci-parum.modified.bed. How many features have changed?



```

wt@Pathogens: ~/Module_3_Programming/Linux
wt@Pathogens:~/Module_3_Programming/Linux$ sed 's/PFD/PF/g' Pfalci-parum.bed > Pfalci-parum.fixed.bed
wt@Pathogens:~/Module_3_Programming/Linux$

```

**Figure 6** Using sed to find and replace text in a file

## Conclusion

The commands we have just seen can process vast amounts of information in a very short amount of time. They can be joined together to manipulate data and calculate results. Bioinformatics software often produces vast quantities of output results and these commands will help you filter things down to a more manageable level so that you can get meaningful findings out the other end. Learning how to use this small set of commands will save you a substantial amount of time.

# What is a computer program?

A computer program is a sequence of written statements or instructions that can be understood by a computer in order to perform an overall task. A program must be written in a specific language (called a programming language) that is understood by the computer.

Several programming languages exist e.g. Perl, Python, Java, C++. However, it is not important which language you learn first because once you are familiar with one programming language, it is much easier to learn others. There is often a distinction between interpreted (e.g. Perl and Python) and compiled (e.g. C++ and Java) languages. People often call programs written in an interpreted language ‘scripts’. All you need to know is that a script is just a program and scripting is just programming. In the remainder of this module we will introduce you to the notion of scripting using shell scripting and Perl.

## Shell scripting

If you have a set of useful Linux commands that you want to use over and over again on different datasets, how do you do this without having to type the same commands over and over again? The command line has its own built-in programming language and can be used to create a “shell script”.

To create a shell script you just create a plain text file and add a series of Linux commands to the file and then treat that file as if it was any other Linux command or program. When you want to repeatedly execute the series of commands for multiple datasets, the shell script can be used to automate the task and save you lots of time.

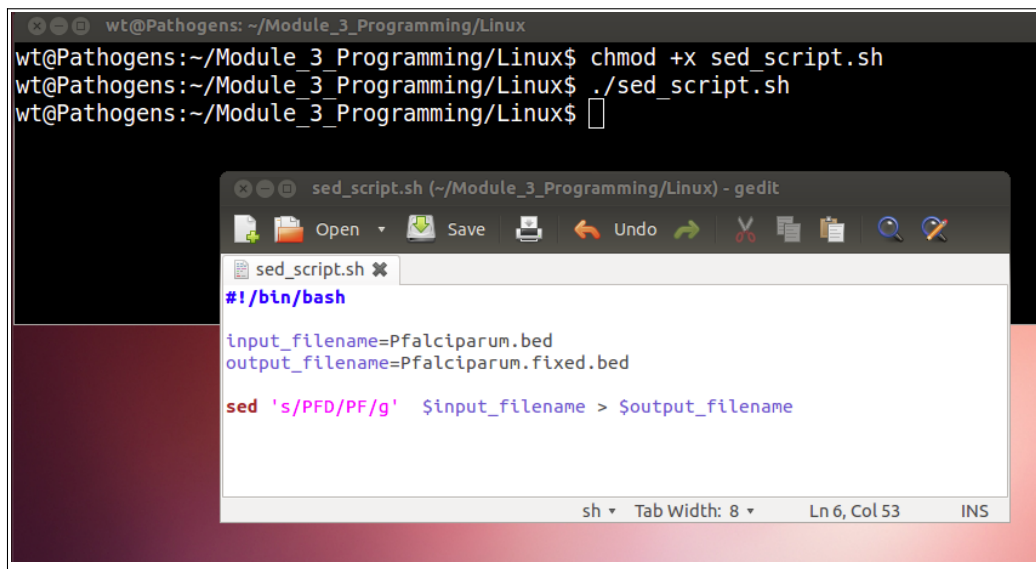
Typically there are many shell programs on a Linux system that can be used to execute your shell scripts, these include ksh, tcsh and bash. You do not need to worry about this for now, you just need to know that for the remainder of this module you will use bash (Bourne Again SHell) and the shell scripts that you write will be called bash scripts.

### Writing shell scripts

Any text editor can be used to write your script, but just remember that a word processor application (like Microsoft Word or LibreOffice) is **not** a text editor. A text editor doesn’t allow you to format text (e.g. bold, italics, font sizes) and produces files in a plain (non-proprietary) format (.txt) that is readable in any computer.

#### Some freely available text editors

- Linux: vi, vim, nano, gedit, emacs
- Windows: notepad, Textpad, PSPad, Notepad++
- Mac OSX: vi, vim, nano, gedit, emacs, TextWrangler, TextEdit



```
wt@Pathogens: ~/Module_3_Programming/Linux
wt@Pathogens:~/Module_3_Programming/Linux$ chmod +x sed_script.sh
wt@Pathogens:~/Module_3_Programming/Linux$ ./sed_script.sh
wt@Pathogens:~/Module_3_Programming/Linux$
```

```
sed_script.sh (~/.Module_3_Programming/Linux) - gedit
#!/bin/bash
input_filename=Pfalcciparum.bed
output_filename=Pfalcciparum.fixed.bed
sed 's/PFD/PF/g' $input_filename > $output_filename
```

**Figure 7** A basic shell script which does the same thing as the previous example. The colours are added by the text editor to make it easier for a person to read and understand the script. The colours aren't used by the computer.

The basic structure of a shell script is shown in Figure 7. It is essentially just a list of Linux commands which are the individual instructions for the computer to follow. When creating a shell script it is standard practice to save it as a .sh file instead of a .txt file.

### Running shell scripts

In order to get the computer to follow the instructions in a script you must execute (i.e. run) the script. In Linux, text files can be executed (i.e. run) as programs, provided they contain instructions in some language **and** the very first line of the text file starts with **#!** followed by the path to a program that can understand (interpret) the instructions.

To run (execute) a text file you must give it execution privileges using `chmod`:

```
chmod +x sed_script.sh
```

and then execute it from the command line:

```
./sed_script.sh
```



## Exercise : Hello World!

First let us look at a basic shell script. Navigate to the `Module_3_Programming` directory and then to the `BASH` directory and using your preferred text editor open the file `hello.sh`. You should see the shell script shown in Figure 8.

```
1  #!/bin/bash
3  #print to the screen
4  echo "Hello world!"
```

**Figure 8** Hello world script

This is a simple shell script which prints the text "Hello world!" to the screen.

- **Line 1** tells the computer which program to use to execute this file, in this case it is the **bash** program. Every bash shell script that you write will always begin with the text `#!/bin/bash`.
- **Line 3** is a comment and acts as a note to yourself about what a line of code does. It is always good practice to add explanatory comments. In shell scripts, comments start with a `#` which tells the computer to ignore everything on this line.
- **Line 4** contains the Linux command `echo` which just prints text to the screen.

In a terminal window, type the following commands to give the `hello.sh` script execution privileges and run the script.

```
chmod +x hello.sh
```

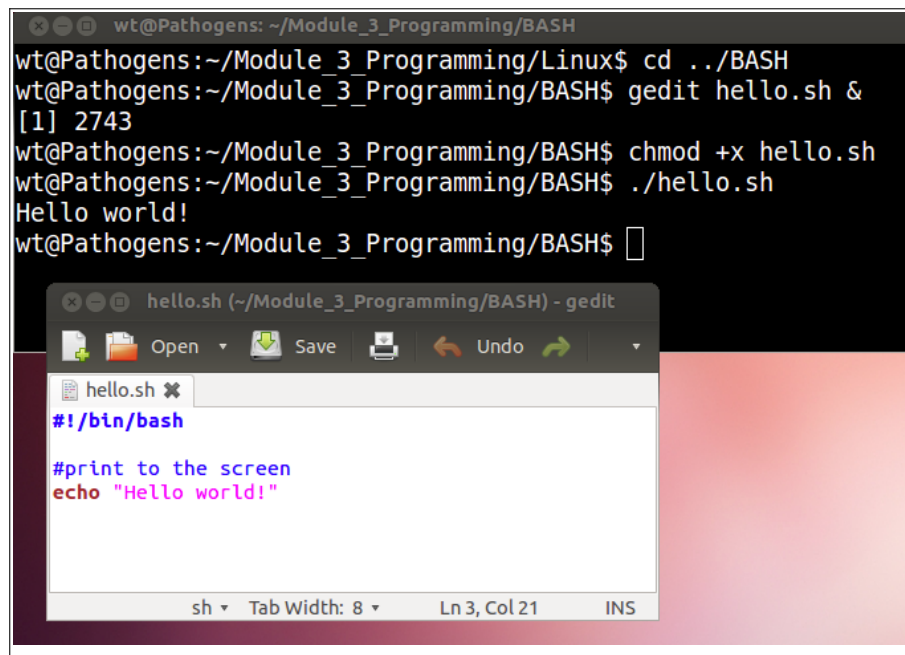
```
./hello.sh
```

```
Hello world!
```

Modify the script `hello.sh` so that it prints your name to the screen. Save this script as `my-name.sh`, give it execution privileges and then run it.

Congratulations, you have just written your first shell script!

Now it is time to make some Linux shell scripts that might actually be useful.



The image shows a terminal window and a gedit editor. The terminal window, titled 'wt@Pathogens: ~/Module\_3\_Programming/BASH', displays the following commands and output:

```
wt@Pathogens:~/Module_3_Programming/Linux$ cd ../BASH
wt@Pathogens:~/Module_3_Programming/BASH$ gedit hello.sh &
[1] 2743
wt@Pathogens:~/Module_3_Programming/BASH$ chmod +x hello.sh
wt@Pathogens:~/Module_3_Programming/BASH$ ./hello.sh
Hello world!
wt@Pathogens:~/Module_3_Programming/BASH$
```

The gedit editor, titled 'hello.sh (~/.Module\_3\_Programming/BASH) - gedit', shows the contents of the script:

```
#!/bin/bash

#print to the screen
echo "Hello world!"
```

The editor's status bar at the bottom indicates 'sh', 'Tab Width: 8', 'Ln 3, Col 21', and 'INS'.

**Figure 9** Hello world script

## Exercise : BWA mapping script

One common task in bioinformatics is to take raw reads from a sequencing machine and align them to a reference sequence (called mapping). This task requires a number of different commands to be run, with the `bwa` command performing the alignment of the sequences. If you have several different files (e.g. from different samples) of sequence data to analyse this can be quite time consuming. Therefore in this exercise we will create a shell script that can be used over and over again to map different samples (also known as lanes) of sequence data. Essentially this involves taking in 3 files and producing a single BAM file as output.

Navigate to the `Module_3_Programming` directory and then to the `BASH` directory and using your preferred text editor open the file `map_lanes.sh`. You should see the shell script shown in Figure 10.

```

1  #!/bin/bash

3  #read in values from command line
4  fastq1=$1
5  fastq2=$2
6  ref=$3
7  output=$4

9  #index the reference file
10 bwa index $ref

12 #map the sequence data
13 bwa aln $ref $fastq1 > F.sai
14 bwa aln $ref $fastq2 > R.sai
15 bwa sampe -a 700 $ref F.sai R.sai $fastq1 $fastq2 > $output.sam

17 #create a sorted and indexed bam file
18 samtools view -b -S $output.sam > $output.tmp.bam
19 samtools sort $output.tmp.bam $output
20 samtools index $output.bam

```

**Figure 10** A shell script to map sequence data with bwa

This script performs the following set of standard mapping tasks:

- **Line 1** tells the computer which program to use to execute this file, in this case it is the **bash** program.
- **Lines 4-7** reads in the values passed to the script from the command line. These values are called command line arguments and we will discuss these in more detail later.
- **Line 10** indexes the reference file.
- **Lines 13-14** aligns the fastq files to the reference genome.
- **Line 15** extracts alignments from bwa's proprietary binary .sai file to a .sam file.
- **Line 18** converts the .sam file into a .bam file using samtools.
- **Lines 19-20** sorts and indexes the .bam file so that it can be viewed with Artemis.

In a terminal window, make the `map_lanes.sh` script executable and run it using the following commands:

```
chmod +x map_lanes.sh
```

```
./map_lanes.sh NV_1.fastq NV_2.fastq L2_cat.fasta NV
```

Please note that this script will run for several minutes, so please be patient. Lots of information about the progress of the mapping will be printed to the screen, but its rare you'd ever need to look at it. The data is from a *Chlamydia trachomatis* sample. While we wait let us learn more about variables and command line arguments.

### Variables

In bash scripting, as in any scripting language, you use containers called variables to store data, change it, and access it later. New variables can be created like this:

```
name=value
```

In a bash script, you must do it exactly like this, with no spaces on either side of the equals sign, the variable name must contain only alphanumeric characters and underscores and it cannot start with a numeric character. Accessing the values stored in a variable can be done like this:

```
$name
```

In the `map_lanes.sh` script we create four different variables and use them to store the values that are passed to the script from the command line.

```
fastq1=$1
```

```
fastq2=$2
```

```
ref=$3
```

```
output=$4
```

Later in the script we access the values stored in these variables. For example, we index the reference genome by passing the value that is stored in the `ref` variable to the `bwa index` command.

```
bwa index $ref
```

### Command Line Arguments

Since we want to use the `map_lanes.sh` script on different datasets, it takes some arguments on the command line telling it what to work on. These arguments are:

- Name of the input fastq files
- Name of the reference file to use
- A prefix to use when writing output files (e.g. `<prefix>.bam`).

Remember we have run the `map_lanes.sh` script with the following command line arguments

```
$ ./map_lanes.sh NV_1.fastq NV_2.fastq L2_cat.fasta NV
```

A shell script can have any number of command line arguments which can be accessed in the script using the variables `$0`, `$1`, `$2`, `$3`, `$4`, `$5` etc.

- The variable `$0` is the script's name, when run with the command above this variable will contain the value `./map_lanes.sh`
- The variable `$1` is the first argument passed to the script, when run with the command above this variable will contain the value `"NV_1.fastq"`
- The variable `$2` is the second argument passed to the script, when run with the command above this variable will contain the value `"NV_2.fastq"`
- The variable `$3` is the third argument passed to the script, when run with the command above this variable will contain the value `"L2_cat.fasta"`
- The variable `$4` is the fourth argument passed to the script, when run with the command above this variable will contain the value `"NV"`
- The total number of arguments is stored in `$#`.

When the `map_lanes.sh` script is finished running, type `ls` to see the contents of the directory. You should see a new file called `NV.bam` which contains the results of mapping the files `NV_1.fastq` and `NV_2.fastq` to the `L2_cat.fasta` reference sequence.

Why is the file called `NV.bam`?

Now use the `map_lanes.sh` script to map the files `AV_1.fastq` and `AV_2.fastq` to the `L2_cat.fasta` reference sequence.

## Exercise : BWA mapping script - what could go wrong?

Try running the `map_lanes.sh` script with the following command line arguments:

```
$ ./map_lanes.sh NV_1.fastq NV_2.fastq L2.fasta NV2
```

Did the script run successfully? If not, why not?

Often, the difference between a good script and a poor script is assessed in terms of the robustness of the script. That is, the ability of the script to handle situations in which something goes wrong. In this case, does the `map_lanes.sh` script handle the situation where a file supplied by the user does not exist?

In this example we will look at improving the robustness of the `map_lanes.sh` script by adding some argument and error checking to the script. Navigate to the `Module_3_Programming` directory and then to the `BASH` directory and using your preferred text editor open the file `map_lanes_validate_inputs.sh`. You should see the shell script shown in Figure 11.

```

1  #!/bin/bash

3  #read in values from command line
4  fastq1=$1
5  fastq2=$2
6  ref=$3
7  output=$4

9  #check the correct number of arguments are passed to the script
10 if [ $# != 4 ]; then
11     echo "Usage: $0 fastq1 fastq2 reference out_prefix"
12     exit
13 fi

15 #check the fastq and reference files passed to the script exist
16 if [ ! -f $fastq1 ] || [ ! -f $fastq2 ] || [ ! -f $ref ]; then
17     echo "Error: One of the input files does not exist"
18     exit
19 fi

21 #index the reference file
22 bwa index $ref

24 #map the sequence data
25 bwa aln $ref $fastq1 > F.sai
26 bwa aln $ref $fastq2 > R.sai
27 bwa sampe -a 700 $ref F.sai R.sai $fastq1 $fastq2 > $output.sam

29 #create a sorted and indexed bam file
30 samtools view -b -S $output.sam > $output.tmp.bam
31 samtools sort $output.tmp.bam $output
32 samtools index $output.bam

```

**Figure 11** Error checking in a shell script

This script performs some checks on the values passed to it from the command line and then performs a set of standard mapping tasks:

- **Lines 1-7** tell the computer which program to use to execute this file and reads in the values passed to the script from the command line.
- **Lines 10-13** checks that the correct number of command line arguments have been passed to the script. The lines say if the number of command line arguments passed to the script is NOT EQUAL TO 4, print a message to the screen telling the user what the correct usage is and exit the script.
- **Lines 16-19** checks that all the files passed to the script exist. The lines say if the first fastq file does not exist OR the second fastq file does not exist OR the reference file does not exist, print an error message to the screen and exit the script.
- **Lines 22-32** perform a set of standard mapping tasks.

**Please note:**

`$#` is the number of command line arguments

`!=` means NOT EQUAL TO

`$0` is the name of the script

`!` is the NOT operator

`-f` checks if a file exists

`||` means OR

Modify the script to check to see if the output file already exists. If it does exist, print a warning message and exit from the script.

### Decision Statements

Sometimes you will want to perform different tasks depending on whether a condition is true or false. In bash this can be achieved with the keyword, `if`. The `if` statement consists of a condition that is evaluated, and a block of code that is run if the condition evaluates to true as shown in Figure 12.

```
1 if [ CONDITION ]; then
2   # instructions to follow if condition is true
3 fi
```

**Figure 12** BASH if statement

## Loops

In bioinformatics we often have to perform the same action/analysis multiple times. For example, its quite common to multiplex a 96 well plate of samples into a single Illumina lane, so to analyse your data you'll need to run the same commands on all 96 sets of sequencing data. Rather than running a script over and over again, you can use a loop. It will keep running a set of commands until a condition is met, for example, loop over all files in a directory and run the commands on each file.

In bash this can be achieved with the keyword `FOR`. The `FOR` statement consists of a list and a variable name, then a block of commands to run. In the example in Figure 13, the `ls` command is run to get a list of files in the current directory. Each file is then taken in turn and is assigned to the variable `i`. The block of code is then run, and `$i` contains the name of the file. Here we just print out the filename, but you can use any command.

```
1  FOR i in $( ls ); DO
2      echo $i
3  DONE
```

**Figure 13** Bash FOR statement

## Exercise : Mapping to Multiple references

Modify the script `map_lanes_validate_inputs.sh` from the previous exercise so that it takes in 2 fastq files and a directory containing references as input, and maps the fastq files to each reference. The references are contained in the sub directory called `references`. It is sometimes necessary to map to multiple different references, because you may not know in advance which reference best represents the underlying genome of the sample you've sequenced.

Save this script as `multiple_mappings.sh`, make it executable and run it with the following command line arguments:

```
./multiple_mappings.sh NV_1.fastq NV_2.fastq references
```

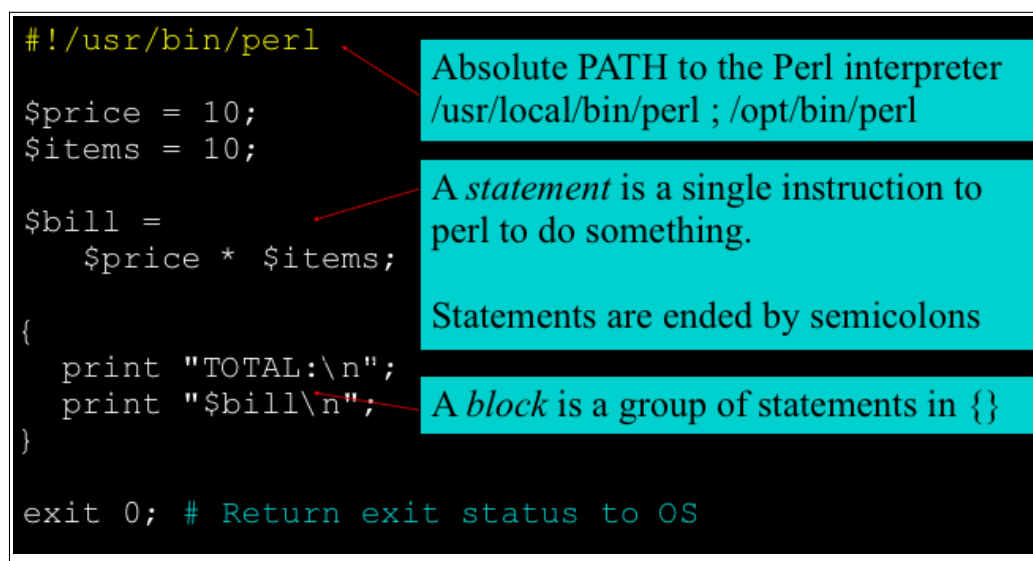
**Hint:** There are 3 references in the directory, so 3 BAM files should be produced as output.



## Introducing Perl

Perl stands for "Practical Extraction and Report Language"<sup>1</sup> which gives you an idea of what the aims were of its creator, Larry Wall. It has been widely used in the UNIX/Linux world for years for the automation of routine tasks, and for summarising large volumes of textual data.

In particular Perl excels at manipulating text, and at dealing with text in different formats. Perl, at its core, provides a wide collection of built-in functions for finding patterns in text, dividing the content of files into meaningful chunks for further processing, etc. Whether you are interested in processing DNA or protein sequences, mapping and processing features onto these sequences, manipulating multiple sequence alignments, etc. rest assured Perl can come to the rescue.



**Figure 14** A basic Perl script

What is a Perl script? A Perl script is a plain text file, containing statements or instructions written in the Perl programming language. Statements can be grouped into *blocks* which have curly brackets around them. Blocks can contain other blocks. An example Perl script is shown in Figure 14. When creating a Perl script it is standard practice to save it as a .pl file instead of a .txt file.

Just like a shell script, to run (execute) a Perl script you must first give it execution privileges:

```
chmod +x program.pl
```

and then execute it from the command line:

```
./program.pl
```

<sup>1</sup> Less charitable people have said it stands for "Pathologically Eclectic Rubbish Lister"

## Exercise : Hello World!

First let us look at a basic Perl script. Navigate to the `Module_3_Programming` directory and then to the `PERL` directory and using your preferred text editor open the file `hello.pl`. You should see the Perl script shown in Figure 15.

```
1  #!/usr/bin/perl
3  use strict;
5  #print to the screen
6  print "Hello world!\n";
```

**Figure 15** Hello world

This is a simple Perl script which prints the text "Hello world!" to the screen.

- **Line 1** tells the computer which program to use to execute this file, in this case it is the **perl** program. Every Perl script that you write will always begin with Lines 1-3.
- **Line 5** is a comment and acts as a note to yourself about what a line of code does. It is always good practice to add explanatory comments. In Perl, comments start with a **#** which tells the computer to ignore everything on this line.
- **Line 6** is an instruction that tells the computer to print "Hello world!" to the screen. In Perl, each instruction ends with a semi-colon.

In a terminal window, give the `hello.pl` script execution privileges and then run it.

```
chmod +x hello.pl
```

```
./hello.pl
```

```
Hello world!
```

Modify the script `hello.pl` using a text editor so that it prints your name to the screen. Save this script as `myname.pl`, give it execution privileges and then run it.

Congratulations, you have just written your first Perl script!

**Introduction to Regular Expressions.** A regular expression often called a pattern, is an expression that matches strings of text, such as particular characters, words, or patterns of characters. Common abbreviations for "regular expression" include regex and regexp.

Some matching operators:

- . any character
- \s a space
- \t a tab
- \n a newline
- \d a digit (0-9)
- \S a non-space character (anything that is not a space or a tab)
- \w an alphanumeric character

Some quantifiers:

- + one or more times
- \* zero or more times
- ? zero or one time

Here are some examples:

colou?r matches both "color" and "colour".

ab\*c matches "ac", "abc", "abbc", "abbbc", and so on.

ab+c matches "abc", "abbc", "abbbc", and so on, but not "ac".

Additional reading: [http://en.wikipedia.org/wiki/Regular\\_expression](http://en.wikipedia.org/wiki/Regular_expression)

**Exercise : Check the characters in a FASTA file are valid**

```

1  #!/usr/bin/perl
2  use strict;

4  my $input_fasta_file = $ARGV[0];
5  open(my $input_file_handle, $input_fasta_file);

7  while ( <$input_file_handle> )
8  {
9      my $line = $_;
10     chomp($line);
11     if($line =~ /^>/)
12     {
13         next;
14     }

16     if($line =~ /[~ACGTNacgtn-]/)
17     {
18         die "The line contains a character we didnt expect\n";
19     }
20 }

22 print "The file is valid\n";

```

**Figure 16** Check FASTA file

This is a Perl script which takes in a FASTA file and will print an error if there is an unexpected character used in the nucleotide sequence.

- **Line 1** tells the computer which program to use to execute this file, in this case it is the **perl** program. Every Perl script that you write will always begin with Lines 1-2.
- **Line 4** reads in a file name thats been passed in from the command line.
- **Line 5** opens the FASTA file for reading, saving the file handle to a variable.
- **Line 7** loops over each line of the FASTA file.
- **Line 9** saves the current line to a variable for later use.
- **Line 10** removes the new line line character from the end of the line.
- **Line 11** looks at the line and looks for the pattern where the first character is '>', e.g. the name of the sequence. This is a regular expression.
- **Line 13** skips to the next line.
- **Line 16** looks for any character that is not in the list given. This is a regular expression.
- **Line 18** stops the script and prints an error message and exits.
- **Line 22** prints out a message to say the script worked.

In a terminal window, give the `check_characters_in_fasta.pl` script execution privileges and then run it.

```
chmod +x check_characters_in_fasta.pl
```

```
./check_characters_in_fasta.pl valid_file.fasta
```

The file is valid

```
./check_characters_in_fasta.pl invalid_file.fasta
```

The line contains a character we didnt expect

Modify the script `check_characters_in_fasta.pl` so that it prints out the name of the sequence which contains the invalid characters.

### BioPerl

BioPerl is a collection of Perl modules that facilitate the development of Perl scripts for bioinformatics applications. BioPerl has support for most file formats used in bioinformatics, usually providing a single interface (set of functionality) for multiple different formats, where the only difference is the name of the format. In the example below, you can change the input file format of the features file from BED to GFF by changing format string from 'BED' to 'GFF' in the script. It also allows you to read a file as input in one format, and output it in another format. Other common functionality is also provided for manipulating sequences such as translating from nucleotides to amino acids, reverse complement,..., and for querying databases. The list of functionality available is absolutely enormous, so you may be able to just plug in a few modules to create a complex script, rather than doing it all from scratch.

Many other programming languages have similar modules, such as BioPython, Bioconductor (R) and BioJava.

## Optional Exercise : Extract Genes from a multiple sequence alignment file

```

1  #!/usr/bin/perl
2  use strict;
3  use Bio::SeqIO;
4  use Bio::FeatureIO;

6  my $input_bed_file   = $ARGV[0];
7  my $input_fasta_file = $ARGV[1];

9  my $bed_file_object = Bio::FeatureIO->new(
10     -file   => $input_bed_file,
11     -format => 'bed');

13 while(my $feature = $bed_file_object->next_feature())
14 {
15     my $fasta_file_input_object = Bio::SeqIO->new(
16         -file   => $input_fasta_file,
17         -format => 'fasta');
18     my @name_of_gene = $feature->get_tag_values('Name');
19     my $fasta_file_output_object = Bio::SeqIO->new(
20         -file   => ">".$name_of_gene[0].'fa',
21         -format => 'fasta');

23     while(my $sequence = $fasta_file_input_object->next_seq() )
24     {
25         my $sequence_of_gene = $sequence->subseq(
26             $feature->start,
27             $feature->end);
28         my $newseq = Bio::Seq->new(
29             -seq => $sequence_of_gene,
30             -id  => $sequence->id );
31         $fasta_file_output_object->write_seq($newseq);
32     }
33 }

```

**Figure 17** Extract Genes from a multi sequence alignment file using BioPerl

A multiple sequence alignment looks very similar to a FASTA file, however all the sequences are the same length. For example, each sequence represents a sample, so to identify SNPs by eye, you can just look down a column and see where the bases differ, or you can use it to build phylogenetic trees. The perl script above takes in a multiple sequence alignment file and a BED file containing the locations of genes you're interested in. It then outputs a new file for every gene containing the sequence of that gene in each sample. For example, if you have lots of samples this script would allow you to extract out a particular gene, possibly because you already know that antibiotic resistance can be caused by a SNP at a particular position.

- **Line 3-4** include the BioPerl packages for parsing FASTA and BED files.
- **Line 6-7** read in filenames passed in from the command line.
- **Line 9-11** opens the BED file for reading with BioPerl.
- **Line 13** loops over all the features in the BED file.
- **Line 15-17** opens the input fasta file for reading with BioPerl.
- **Line 18-21** creates a new fasta file for writing out genes with BioPerl.
- **Line 23** loops over each sequence in the multisequence alignment file.
- **Line 25-27** extracts the gene sequence from the multisequence alignment file.
- **Line 28-31** writes out the gene sequence to a FASTA file.

In a terminal window, give the `extract_genes_from_multifasta_alignment.pl` script execution privileges and then run it.

```
chmod +x extract_genes_from_multifasta_alignment.pl

./extract_genes_from_multifasta_alignment genes.bed aligned_sequences.aln
```

Modify the script so that it outputs the protein sequences of the genes instead of the nucleotides. Details of how to do this can be found at:

<http://search.cpan.org/~cjfields/BioPerl-1.6.901/Bio/SeqIO.pm>

### **Beginning Perl for Bioinformatics Book.**

We have provided a copy of the e-book "Beginning Perl for Bioinformatics" as a pdf on the USB disk. If there are any topics or concepts that are covered in this section that you need further information or explanations for, please refer to this book. For example, regular expressions which are discussed later in this section and are explained in chapter 5 starting on page 67.

**Look for Perl documentation online.**

If you search for **split** at <http://perldoc.perl.org/> you will be directed to the corresponding page with documentation on this function:

<http://perldoc.perl.org/functions/split.html>.

A fast way to look up documentation using the command line is to use the **perldoc** command.

This will let you access the documentation for built-in functions (`perldoc -f split`) as well as for third party modules, e.g. from the BioPerl package, `perldoc Bio::SeqIO`.

## Useful resources

For those of you who may want to learn more Perl after this course, we have included a copy of the e-book "Beginning Perl for Bioinformatics" as a pdf which is a good starting point for anyone who wants to teach themselves Perl. This book is stored as a pdf file on the USB disk and is called "Beginning\_Perl\_for\_Bioinformatics.pdf".

Some further useful resources include:

### Books

- "Learning Perl" by Randall Schwartz
- "Mastering Perl for Bioinformatics" by James Tisdall
- "Perl programming for biologists" by D. Curtis Jamison
- "Mastering Regular Expressions" by Jeffrey Friedl

### Online resources

- <http://www.perl.com/pub/a/2000/10/begperl1.html>
- <http://www.bioperl.org>
- <http://search.cpan.org>

We hope you have fun learning to program!