Informatics on High-throughput Sequencing Data

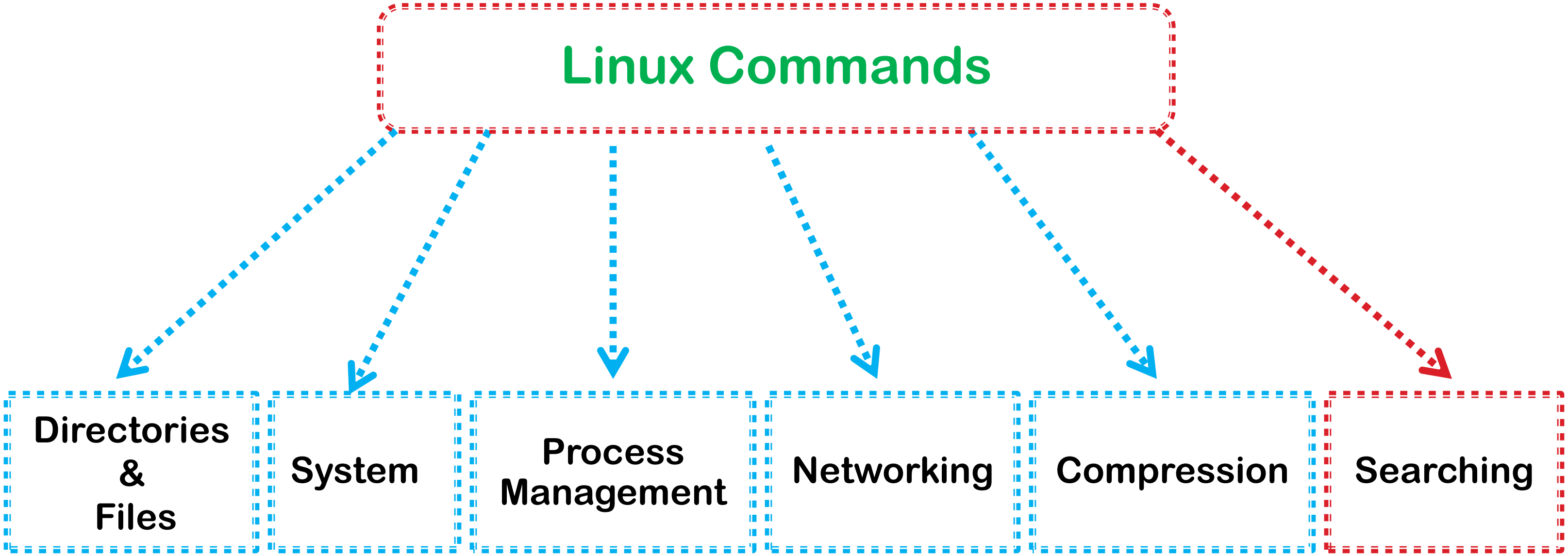
(Summer Course 2020)

Day 5

Agenda

- Unix-based systems.
- Why Linux!
- Let’s start!
- Linux Commands for:
  - Files & Directories.
  - System.
  - Process Management.
  - Networking.
  - Compression.
  - Searching.
- Piping output.
- Wildcard character.
- Redirecting output.
- Stream Editor (Sed).
- Linux tools for text files processing.
- Shell Scripting

Getting Started !!



Linux Commands

Directories & Files

System

Process Management

Networking

Compression

Searching

Searching

```
grep pattern files
```

Search for pattern in files.

```
grep -r pattern Bio
```

Search recursively for pattern in Bio.

```
grep -rn pattern Bio
```

Search recursively for pattern in Bio and show the line number found.

```
grep -rn --colour pattern Bio
```

Search recursively for pattern in Bio and colored the matched patterns.

Searching

```
grep –i pattern File
```

Search ignoring the case.

```
grep –in --colour TTTTT Reads1.fastq
```

```
grep –w ‘word’ Bio
```

Search for the whole word

```
grep –w  ‘chr1’ athal.genes.gtf

grep –iw  ‘chr1’ athal.genes.gtf

grep Chr1 athal.genes.gtf
```

Searching

```
grep -c pattern File
```

Prints number of occurrence.

```
grep -c Chr1 athal.genes.gtf

grep -w -B 1 -A 2 ‘word’ File
```

Search for the whole word and prints 1 line before it and 2 lines after it.

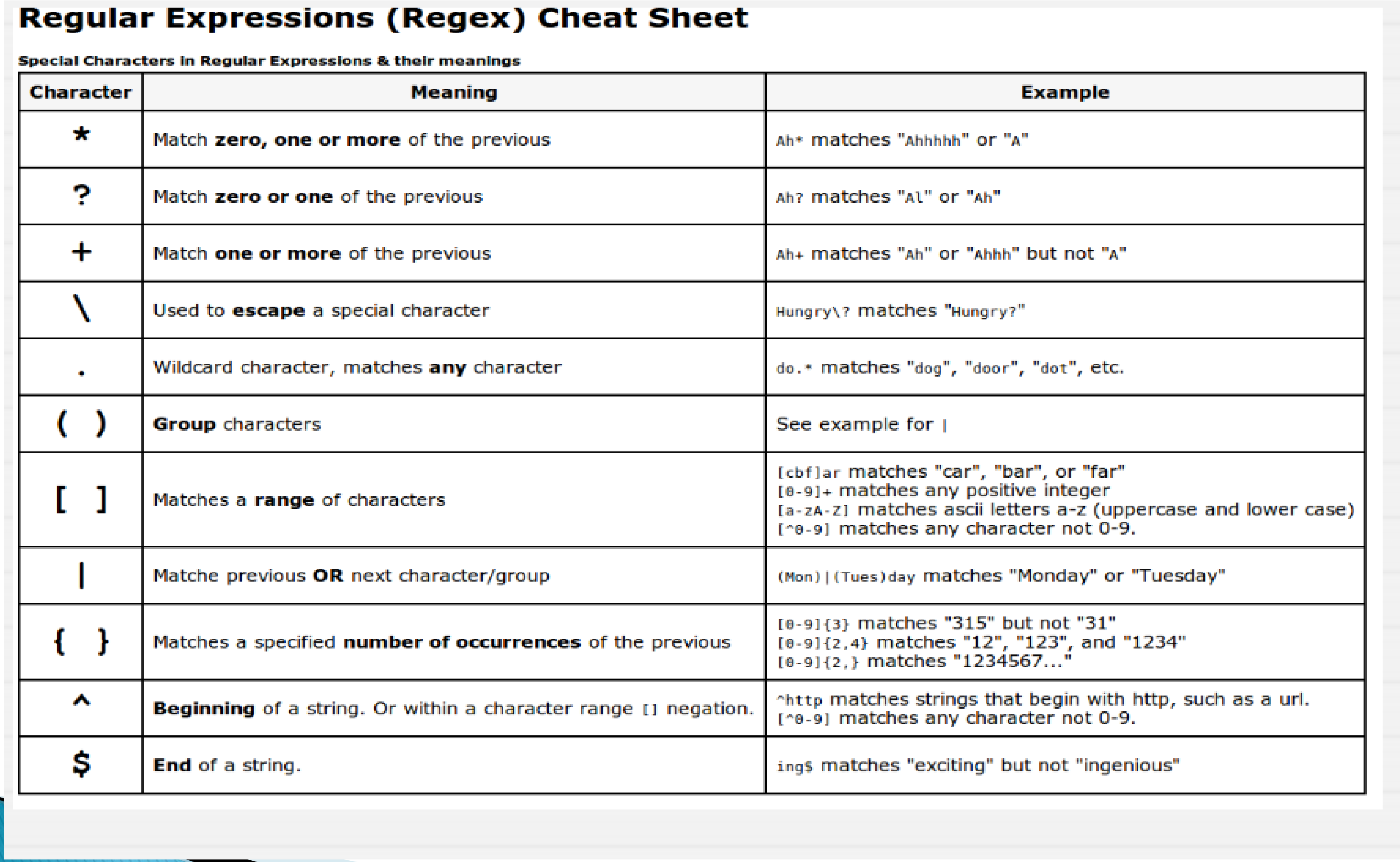
```
grep -v  ‘Chr1’ athal.genes.gtf

grep -vi  ‘Chr1’ athal.genes.gtf

grep -viw  ‘Chr1’ athal.genes.gtf

grep -w -f reads.txt reads.fastq
```

Searching



Regular Expressions (Regex) Cheat Sheet

Special Characters in Regular Expressions & their meanings

Character	Meaning	Example
*	Match **zero, one or more** of the previous	Ah* matches "Ahhhhh" or "A"
?	Match **zero or one** of the previous	Ah? matches "Al" or "Ah"
+	Match **one or more** of the previous	Ah+ matches "Ah" or "Ahhh" but not "A"
\	Used to **escape** a special character	Hungry\? matches "Hungry?"
.	Wildcard character, matches **any** character	do.* matches "dog", "door", "dot", etc.
( )	**Group** characters	See example for |
[ ]	Matches a **range** of characters	[cbf]ar matches "car", "bar", or "far" [0-9]+ matches any positive integer [a-zA-Z] matches ascii letters a-z (uppercase and lower case) [^0-9] matches any character not 0-9.
|	Matche previous **OR** next character/group	(Mon)|(Tues)day matches "Monday" or "Tuesday"
{ }	Matches a specified **number of occurrences** of the previous	[0-9]{3} matches "315" but not "31" [0-9]{2,4} matches "12", "123", and "1234" [0-9]{2,} matches "1234567..."
^	**Beginning** of a string. Or within a character range [] negation.	^http matches strings that begin with http, such as a url. [^0-9] matches any character not 0-9.
$	**End** of a string.	ing$ matches "exciting" but not "ingenious"

https://blog.mycode.website/regular-expressions/

Searching

```
grep –r “>” reads
```

- grep will be used to examine whether the fasta files you downloaded contain a properly formatted title line.

```
grep –E ‘^@.+/1’ reads > SeqIDs.txt
less SeqIDs.txt
```

- grep will be used to examine whether the fastq files you downloaded contain a properly formatted title line.

Searching

```
grep –r “>” reads | wc –l
```

- count how many lines start with “>”

```
grep –E ‘^@.+/1’ reads | wc -l
```

- count how many lines start with “@”

Piping output

- Pipes are represented by the | character.
- It is possible to send the output of one program to another program as input.

```
history |less
```

List all remembered commands page by page.

```
history |grep Bio
```

List all remembered commands containing string “Bio”.

References

- **Paul Stothard, An Introduction to Linux for bioinformatics , 2016.**
- **Robert Bukowski, Linux for Biologists- Part 1.**
- **Steve Pederson, Introduction To Linux/Ubuntu & Sell Scripting, 2014.**
- https://bioinformatics.uconn.edu/unix-basics/#
- https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/
- https://coding4medicine.com/Members/pages/home/
- https://open.oregonstate.education/computationalbiology/chapter/patterns-regular-expressions/
- https://bioinformaticsworkbook.org/Appendix/Unix/unix-basics-3grep.html#gsc.tab=0
- https://datacarpentry.org/shell-genomics/04-redirection/

Thanks!

// | ?