Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Regular Expressions

HS2020

Bioinformatics - R

Sample Lane Cell Archive

JL-2-76_S7_L001_R2_001.fastq.gz

Saample Read Format

JL-2-76_S7_L001_R2_001.fastq.gz

JL-2-76_R2.fq.gz

Bioinformatics - R

```
mv JL-2-76_S7_L001_R2_001.fastq.gz JL-2-76_R2.fq.gz
```

move stdin stdout

```
z_tmp/JL-2-76_S7_L001_R2_001.fastq.gz
z_tmp/JL-2-86_S18_L001_R1_001.fastq.gz
z_tmp/JL-2-86_S18_L001_R2_001.fastq.gz
z_tmp/JL-2-96_S34_L001_R1_001.fastq.gz
z_tmp/JL-2-96_S34_L001_R2_001.fastq.gz
z_tmp/JL-3-53_S40_L001_R1_001.fastq.gz
z_tmp/JL-3-53_S40_L001_R2_001.fastq.gz
z_tmp/JL-3-63_S8_L001_R1_001.fastq.gz
z_tmp/JL-3-63_S8_L001_R2_001.fastq.gz
z_tmp/JL-3-73_S19_L001_R1_001.fastq.gz
z_tmp/JL-3-73_S19_L001_R2_001.fastq.gz
z_tmp/JL-4-20_S35_L001_R1_001.fastq.gz
z_tmp/JL-4-20_S35_L001_R2_001.fastq.gz
z_tmp/JL-4-23_S41_L001_R1_001.fastq.gz
z_tmp/JL-4-23_S41_L001_R2_001.fastq.gz
z_tmp/JL-4-33_S9_L001_R1_001.fastq.gz
z_tmp/JL-4-33_S9_L001_R2_001.fastq.gz
z_tmp/JL-4-50_S20_L001_R1_001.fastq.gz
z_tmp/JL-4-50_S20_L001_R2_001.fastq.gz
...
```

JL-2-76_S7_L001_R2_001.fastq.gz

```
z_tmp\/(\w+-\w+-\w+)_S\d+_L001_(R[12])_001.fastq.gz
```

```
mv z_tmp/$1_S*_L001_$2_001.fastq.gz a_data/gz/\$1_$2.fq.gz
```

Bioinformatics - R

```
#!/bin/bash

## Rename Sequencing Files
## Hubert J. Farnsworth (17.01.2045)

mv z_tmp/JL-2-76_S*_L001_R1_001.fastq.gz a_data/gz/JL-2-76_R1.fq.gz
mv z_tmp/JL-2-76_S*_L001_R2_001.fastq.gz a_data/gz/JL-2-76_R2.fq.gz
mv z_tmp/JL-2-86_S*_L001_R1_001.fastq.gz a_data/gz/JL-2-86_R1.fq.gz
mv z_tmp/JL-2-86_S*_L001_R2_001.fastq.gz a_data/gz/JL-2-86_R2.fq.gz
mv z_tmp/JL-2-96_S*_L001_R1_001.fastq.gz a_data/gz/JL-2-96_R1.fq.gz
mv z_tmp/JL-2-96_S*_L001_R2_001.fastq.gz a_data/gz/JL-2-96_R2.fq.gz
mv z_tmp/JL-3-53_S*_L001_R1_001.fastq.gz a_data/gz/JL-3-53_R1.fq.gz
mv z_tmp/JL-3-53_S*_L001_R2_001.fastq.gz a_data/gz/JL-3-53_R2.fq.gz
mv z_tmp/JL-3-63_S*_L001_R1_001.fastq.gz a_data/gz/JL-3-63_R1.fq.gz
mv z_tmp/JL-3-63_S*_L001_R2_001.fastq.gz a_data/gz/JL-3-63_R2.fq.gz
mv z_tmp/JL-3-73_S*_L001_R1_001.fastq.gz a_data/gz/JL-3-73_R1.fq.gz
mv z_tmp/JL-3-73_S*_L001_R2_001.fastq.gz a_data/gz/JL-3-73_R2.fq.gz
mv z_tmp/JL-4-20_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-20_R1.fq.gz
mv z_tmp/JL-4-20_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-20_R2.fq.gz
mv z_tmp/JL-4-23_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-23_R1.fq.gz
mv z_tmp/JL-4-23_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-23_R2.fq.gz
mv z_tmp/JL-4-33_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-33_R1.fq.gz
mv z_tmp/JL-4-33_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-33_R2.fq.gz
mv z_tmp/JL-4-50_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-50_R1.fq.gz
mv z_tmp/JL-4-50_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-50_R2.fq.gz
```

Bioinformatics - R

```
2019-10-26 app[java.2]: 126.0.132.125
```

```
.* (.*)\[(.*)\]:.*
```

```
[12]\d{3}-[01]\d-[0-3]\d ([^] \[]*?)\[([^\]]*?)\]:.*
```

speed

(~42x faster)

Limits of (simple) find and replace

find: the

✔ Whole word only

O Captain! my Captain! rise up and hear the bells;
Rise up-for you the flag is flung-for you the bugle trills;
For you bouquets and ribbon'd wreaths-for you the shores a-crowding;
For you they call, the swaying mass, their eager faces turning;
Here Captain! dear father!
This arm beneath your head;
It is some dream that on the deck,
You've fallen cold and dead.

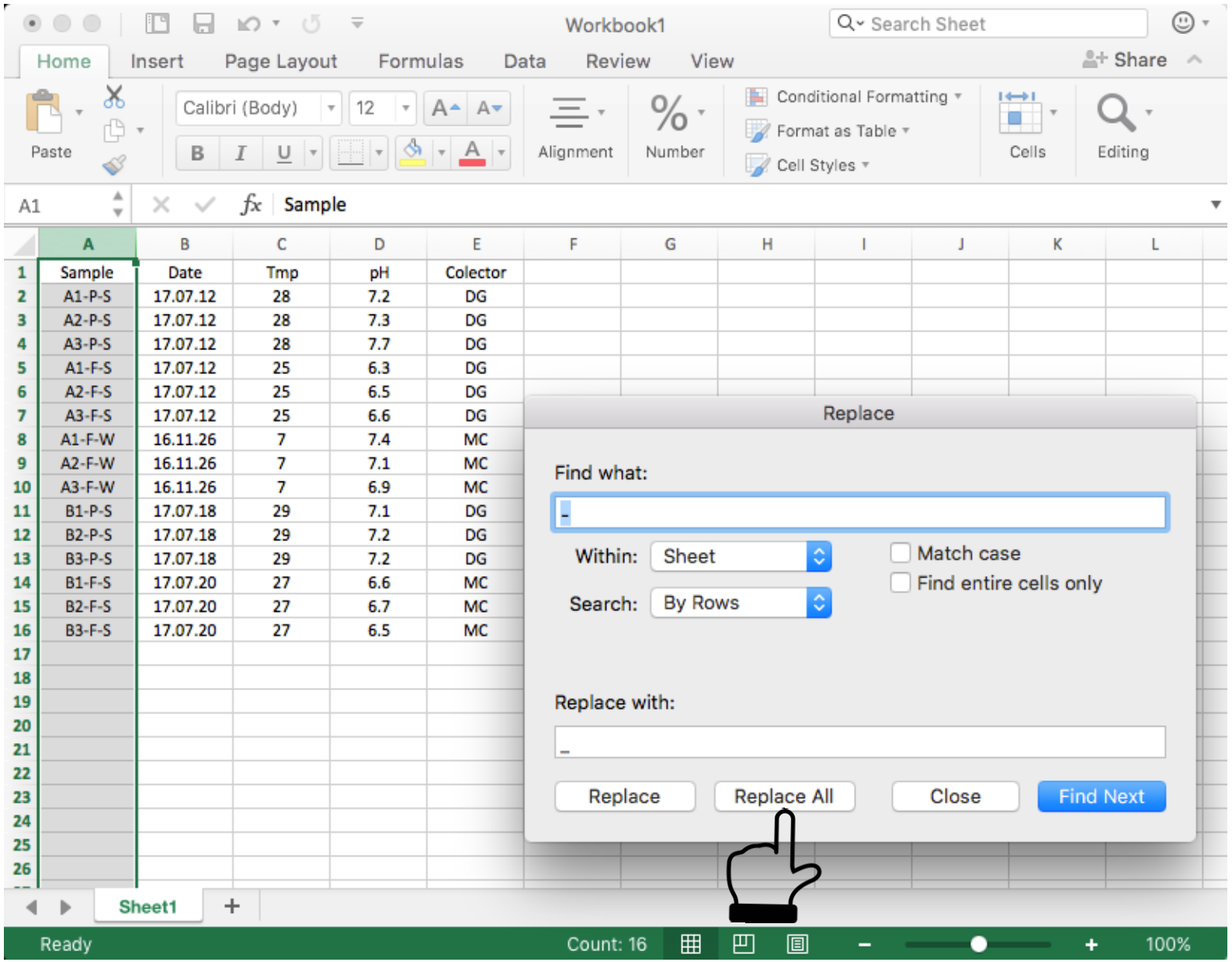
by Walt Whitman

Bioinformatics - RegEx

find & replace

Sample		Sample
A1-P-S	→	A1_P_S
A2-P-S	→	A2_P_S
A3-P-S	→	A3_P_S
A1-F-S	→	A1_F_S
A2-F-S	→	A2_F_S
A3-F-S	→	A3_F_S
A1-F-W	→	A1_F_W
A2-F-W	→	A2_F_W
A3-F-W	→	A3_F_W
B1-P-S	→	B1_P_S
B2-P-S	→	B2_P_S
B3-P-S	→	B3_P_S
B1-F-S	→	B1_F_S
B2-F-S	→	B2_F_S
B3-F-S	→	B3_F_S

Bioinformatics - RegEx



Bioinformatics - RegEx

find & replace

Sample		Sample
A1-P-S	→	A1_P_S
A2-P-S	→	A2_P_S
A3-P-S	→	A3_P_S
A1-F-S	→	A1_F_S
A2-F-S	→	A2_F_S
A3-F-S	→	A3_F_S
A1-F-W	→	A1_F_W
A2-F-W	→	A2_F_W
A3-F-W	→	A3_F_W
B1-P-S	→	B1_P_S
B2-P-S	→	B2_P_S
B3-P-S	→	B3_P_S
B1-F-S	→	B1_F_S
B2-F-S	→	B2_F_S
B3-F-S	→	B3_F_S

regular expression?

Sample		Sample
A1-P-S	→	Sample_PS_A-1
A2-P-S	→	Sample_PS_A-2
A3-P-S	→	Sample_PS_A-3
A1-F-S	→	Sample_FS_A-1
A2-F-S	→	Sample_FS_A-2
A3-F-S	→	Sample_FS_A-3
A1-F-W	→	Sample_FW_A-1
A2-F-W	→	Sample_FW_A-2
A3-F-W	→	Sample_FW_A-3
B1-P-S	→	Sample_PS_B-1
B2-P-S	→	Sample_PS_B-2
B3-P-S	→	Sample_PS_B-3
B1-F-S	→	Sample_FS_B-1
B2-F-S	→	Sample_FS_B-2
B3-F-S	→	Sample_FS_B-3

Bioinformatics - RegEx

Sample		Sample
A1-P-S	→	Sample_PS_A-1
A2-P-S	→	Sample_PS_A-2
A3-P-S	→	Sample_PS_A-3
A1-F-S	→	Sample_FS_A-1
A2-F-S	→	Sample_FS_A-2
A3-F-S	→	Sample_FS_A-3
A1-F-W	→	Sample_FW_A-1
A2-F-W	→	Sample_FW_A-2
A3-F-W	→	Sample_FW_A-3
B1-P-S	→	Sample_PS_B-1
B2-P-S	→	Sample_PS_B-2
B3-P-S	→	Sample_PS_B-3
B1-F-S	→	Sample_FS_B-1
B2-F-S	→	Sample_FS_B-2
B3-F-S	→	Sample_FS_B-3

find: (\w)(\d)-(\w)-(\w)

replace*1: Sample_$3$4_$1-$2

replace*2: Sample_\3\4_\1-\2

*1 example for e.g. Atom
*2 example for e.g. TextWrangler

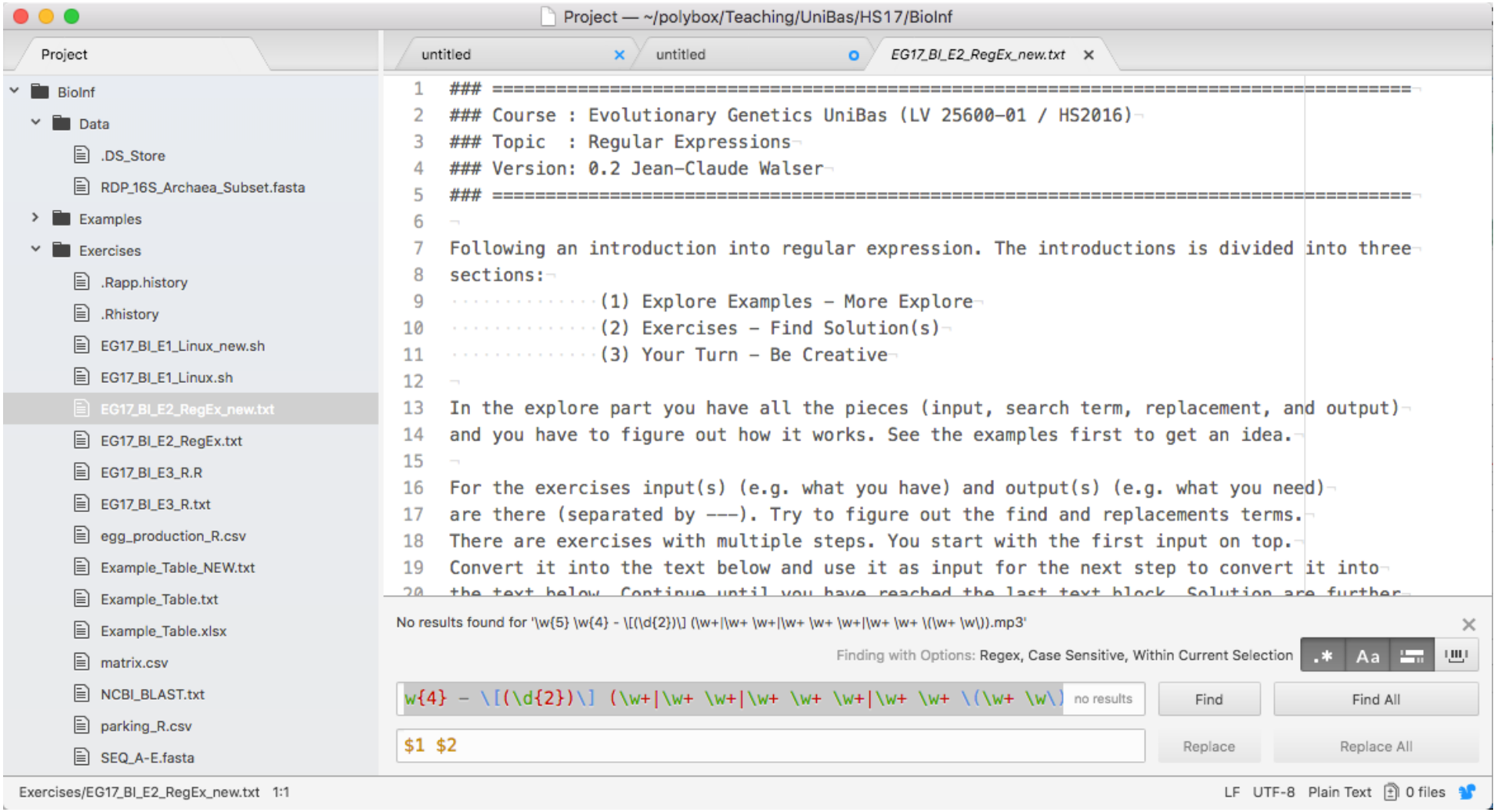
Bioinformatics - RegEx

A **regular expression** (**regex** or **regexp** for short) is a special **text string for describing a search pattern**. You can think of regular expressions as wildcards on steroids.

```
\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b
```

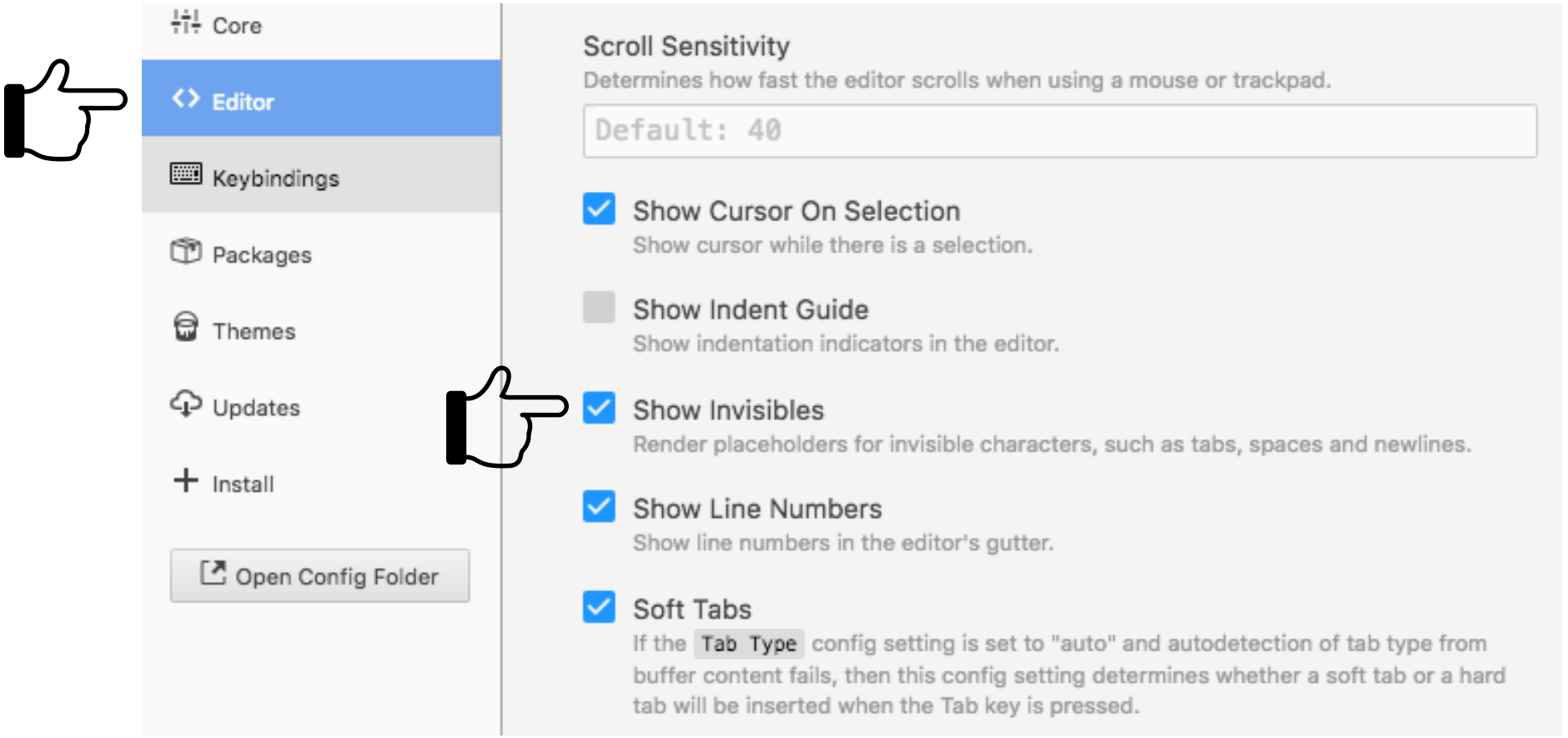
A regular expression "**engine**" is a piece of software that can process regular expressions, trying to match the pattern to the given string. Usually, the engine is part of a larger application and you do not access the engine directly.

Bioinformatics - RegEx



https://atom.io

[Atom] > [Preferences…] > Settings



Bioinformatics - RegEx

[Find] > [Find in Buffer] or [cmd] & [F]

Enable regex mode:

Bioinformatics - RegEx

regex	meaning
\b	word boundary

find: the\b

O Captain! my Captain! rise up and hear the bells;
Rise up-for you the flag is flung-for you the bugle trills;
For you bouquets and ribbon'd wreaths-for you the shores a-crowding;
For you they call, the swaying mass, their eager faces turning;
Here Captain! dear father!
This arm beneath your head;
It is some dream that on the deck,
You've fallen cold and dead.

by Walt Whitman

Bioinformatics - RegEx

regex	meaning
\w	word character (including letters, numbers, and underscore)
\W	not word character
\w+	one or more words
\d	digit
\D	not digit
\d+	one or more digits
\b	word boundary
\t	tab
\s	white space
\r	end-of-line (sometimes \n)

Bioinformatics - RegEx

regex	meaning
^A	begin with A
A$	end with A
[ ]	character class/set
[ACGT]	set of character
[A-Za-z]	2 sets (ranges with metacharacter dash)
[0-9\.]	set of digits and decimal point
[^TG]	negation(s)
{}	number of matches
{n}	n matches
{n,m}	n-m matches
{n,}	n and more matches

Bioinformatics - RegEx

regex	meaning
()	search set
(|)	search set alternatives
\	escape character
\s	white space
\S	not white space
.	any character except new line
*	0 or more (e.g. .* whole line)
+	1 or more
?	0 or 1

Bioinformatics - RegEx

match	skip
can	dan
man	ran
fan	pan

[cmf]an

[cmf].

^[cmf]

Bioinformatics - RegEx

match	skip
hog	bog
dog	

[hd]og

^[hd]

Bioinformatics - RegEx

match	skip
Ana	aax
Bob	bby
Cpc	zcz

[ABC]

[A-C]

^[A-C]

^[^a-z]

^[^abz]

Bioinformatics - RegEx

match	skip	
wazzzzup	wazup	waz{2,}
wazzzup		
		waz{3}
		wazz. (!!!)

Bioinformatics - RegEx

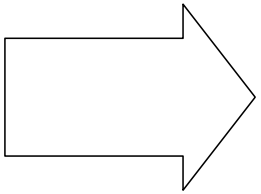
match	skip
test.tmp	seq.fa
seq.tmp	tmp.tex
help.tmp	help.txt

*.tmp

tmp$

(\w+).tmp$

Mus musclus
Agalma elegans
Frillagalma vitiazi
Cordagalma tottoni
Shortia galacifolia



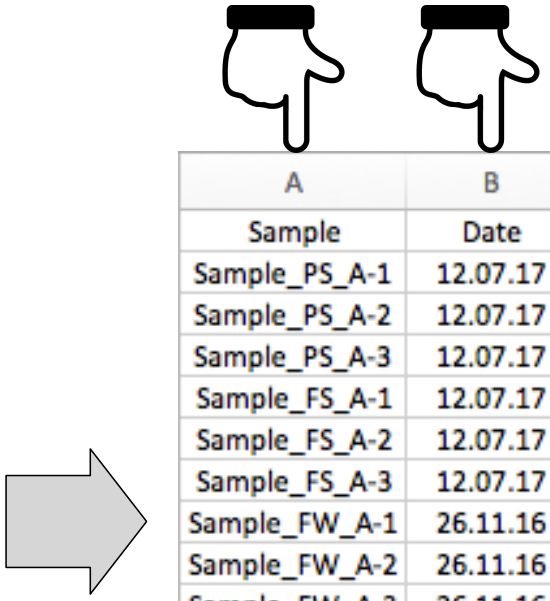
M. musclus
A. elegans
F. vitiazi
C. tottoni
S. galacifolia

Find: **(\w)\w+ (\w+)**

Replace: **$1. $2**

Bioinformatics - RegEx

A	B	C	D	E
Sample	Date	Tmp	pH	Colector
A1-P-S	17.07.12	28	7.2	DG
A2-P-S	17.07.12	28	7.3	DG
A3-P-S	17.07.12	28	7.7	DG
A1-F-S	17.07.12	25	6.3	DG
A2-F-S	17.07.12	25	6.5	DG
A3-F-S	17.07.12	25	6.6	DG
A1-F-W	16.11.26	7	7.4	MC
A2-F-W	16.11.26	7	7.1	MC
A3-F-W	16.11.26	7	6.9	MC
B1-P-S	17.07.18	29	7.1	DG
B2-P-S	17.07.18	29	7.2	DG
B3-P-S	17.07.18	29	7.2	DG
B1-F-S	17.07.20	27	6.6	MC
B2-F-S	17.07.20	27	6.7	MC
B3-F-S	17.07.20	27	6.5	MC



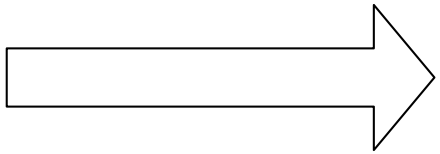
A	B	C	D	E
Sample	Date	Tmp	pH	Collector
Sample_PS_A-1	12.07.17	28	7.2	DG
Sample_PS_A-2	12.07.17	28	7.3	DG
Sample_PS_A-3	12.07.17	28	7.7	DG
Sample_FS_A-1	12.07.17	25	6.3	DG
Sample_FS_A-2	12.07.17	25	6.5	DG
Sample_FS_A-3	12.07.17	25	6.6	DG
Sample_FW_A-1	26.11.16	7	7.4	MC
Sample_FW_A-2	26.11.16	7	7.1	MC
Sample_FW_A-3	26.11.16	7	6.9	MC
Sample_PS_B-1	18.07.17	29	7.1	DG
Sample_PS_B-2	18.07.17	29	7.2	DG
Sample_PS_B-3	18.07.17	29	7.2	DG
Sample_FS_B-1	20.07.17	27	6.6	MC
Sample_FS_B-2	20.07.17	27	6.7	MC
Sample_FS_B-3	20.07.17	27	6.5	MC

Bioinformatics - RegEx

Example_Table.xls

A	B	C	D	E
Sample	Date	Tmp	pH	Colector
A1-P-S	17.07.12	28	7.2	DG
A2-P-S	17.07.12	28	7.3	DG
A3-P-S	17.07.12	28	7.7	DG
A1-F-S	17.07.12	25	6.3	DG
A2-F-S	17.07.12	25	6.5	DG
A3-F-S	17.07.12	25	6.6	DG
A1-F-W	16.11.26	7	7.4	MC
A2-F-W	16.11.26	7	7.1	MC
A3-F-W	16.11.26	7	6.9	MC
B1-P-S	17.07.18	29	7.1	DG
B2-P-S	17.07.18	29	7.2	DG
B3-P-S	17.07.18	29	7.2	DG
B1-F-S	17.07.20	27	6.6	MC
B2-F-S	17.07.20	27	6.7	MC
B3-F-S	17.07.20	27	6.5	MC

Save As…
Tab Delimited Text (.txt)

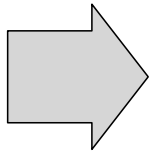


Example_Table.txt

```
Sample	Date	Tmp	pH	Collector
A1-P-S	17.07.12	28	7.2	DG
A2-P-S	17.07.12	28	7.3	DG
A3-P-S	17.07.12	28	7.7	DG
A1-F-S	17.07.12	25	6.3	DG
A2-F-S	17.07.12	25	6.5	DG
A3-F-S	17.07.12	25	6.6	DG
A1-F-W	16.11.26	7	7.4	MC
A2-F-W	16.11.26	7	7.1	MC
A3-F-W	16.11.26	7	6.9	MC
B1-P-S	17.07.18	29	7.1	DG
B2-P-S	17.07.18	29	7.2	DG
B3-P-S	17.07.18	29	7.2	DG
B1-F-S	17.07.20	27	6.6	MC
B2-F-S	17.07.20	27	6.7	MC
B3-F-S	17.07.20	27	6.5	MC
```

Bioinformatics - RegEx

Sample	Date	Tmp	pH	Collector
A1-P-S	17.07.12	28	7.2	DG
A2-P-S	17.07.12	28	7.3	DG
A3-P-S	17.07.12	28	7.7	DG
A1-F-S	17.07.12	25	6.3	DG
A2-F-S	17.07.12	25	6.5	DG
A3-F-S	17.07.12	25	6.6	DG
A1-F-W	16.11.26	7	7.4	MC
A2-F-W	16.11.26	7	7.1	MC
A3-F-W	16.11.26	7	6.9	MC
B1-P-S	17.07.18	29	7.1	DG
B2-P-S	17.07.18	29	7.2	DG
B3-P-S	17.07.18	29	7.2	DG
B1-F-S	17.07.20	27	6.6	MC
B2-F-S	17.07.20	27	6.7	MC
B3-F-S	17.07.20	27	6.5	MC



Sample	Date	Tmp	pH	Collector
Sample_PS_A-1	12.07.17	28	7.2	DG
Sample_PS_A-2	12.07.17	28	7.3	DG
Sample_PS_A-3	12.07.17	28	7.7	DG
Sample_FS_A-1	12.07.17	25	6.3	DG
Sample_FS_A-2	12.07.17	25	6.5	DG
Sample_FS_A-3	12.07.17	25	6.6	DG
Sample_FW_A-1	26.11.16	7	7.4	MC
Sample_FW_A-2	26.11.16	7	7.1	MC
Sample_FW_A-3	26.11.16	7	6.9	MC
Sample_PS_B-1	18.07.17	29	7.1	DG
Sample_PS_B-2	18.07.17	29	7.2	DG
Sample_PS_B-3	18.07.17	29	7.2	DG
Sample_FS_B-1	20.07.17	27	6.6	MC
Sample_FS_B-2	20.07.17	27	6.7	MC
Sample_FS_B-3	20.07.17	27	6.5	MC

find:	**(\w)(\d)-(\w)-(\w)**\t(\d{2}).(\d{2}).(\d+)\t(\d+)\t(\d.\d)\t(\w+)
replace:	**Sample_$3$4_$1-$2**\t$7.$6.$5\t$8\t$9\t$10