Reproducible Research
RegEx

Niklaus Zemp

June 2020

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

Replace pattern

>E1_L96

AATTATTACTTTATGACACT

>E2_L119

CGAATTCGTCATTTGAAACCGATTCGG
CTAGAATT

>E4_L96

TTTTACTTACATGGTGAAAAAATAGAAT
ACGTATTCTCTGCCAAGATTCATTAACT
CAAAAGAGAATTTTTTGAGTTAATGCA
GAGGATACGAATT

>E1

AATTATTACTTTATGACACT

>E2

CGAATTCGTCATTTGAAACCGATTCGG
CTAGAATT

>E4

TTTTACTTACATGGTGAAAAAATAGAAT
ACGTATTCTCTGCCAAGATTCATTAACT
CAAAAGAGAATTTTTTGAGTTAATGCA
GAGGATACGAATT

Regular expression (RegEx)

Groups and Ranges	
.	Any character except new line (\n)
(a|b)	a or b
(...)	Group
(?:...)	Passive (non-capturing) group
[abc]	Range (a or b or c)
[^abc]	Not (a or b or c)
[a-q]	Lower case letter from a to q
[A-Q]	Upper case letter from A to Q
[0-7]	Digit from 0 to 7
\x	Group/subpattern number "x"

Ranges are inclusive.

Quantifiers			
*	0 or more	{3}	Exactly 3
+	1 or more	{3,}	3 or more
?	0 or 1	{3,5}	3, 4 or 5

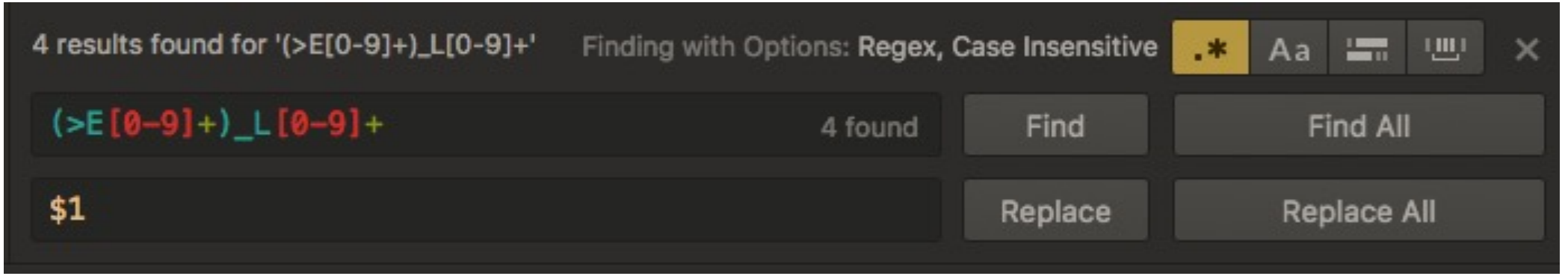
Add a ? to a quantifier to make it ungreedy.

Anchors	
^	Start of string, or start of line in multi-line pattern
\A	Start of string
$	End of string, or end of line in multi-line pattern
\Z	End of string
\b	Word boundary
\B	Not word boundary
\<	Start of word
\>	End of word

Tutorial:

RegEx with Atom

```
>E1_L96
AATTATTACTTTATGACACTGACACTGACACTGACACTGACATACAGAAATGAATTAAGTCAAGAACCAAAGCGGAGGAAGCGCTTCTAGAGAATT
>E2_L119
CGAATTCGTCATTTGAAACCGATTCAATGAGTTTTAGACTTGAGTTCACGAAGAAGTTTAATGAACTTAAAAACACCCTAGTTCTACTCTTCAAAT
>E4_L96
TTTTACTTACATGGTGAAAAAATAGAATACGTATTCTCTGCCAAGATTCATTAACTCAAAAGAGAATTTTTTGAGTTAATGCAGAGGATACGAATT
>E14_L135
CGAATGTCTCCTGGGACTTCTTGGTAGCTTGACCTTCATTCCACATCGTCTGCATTAACTCAAAAGAGTATCTTTTGAGTTAATACCTATCGGCTG
```



(>E[0-9]+)_L[0-9]+