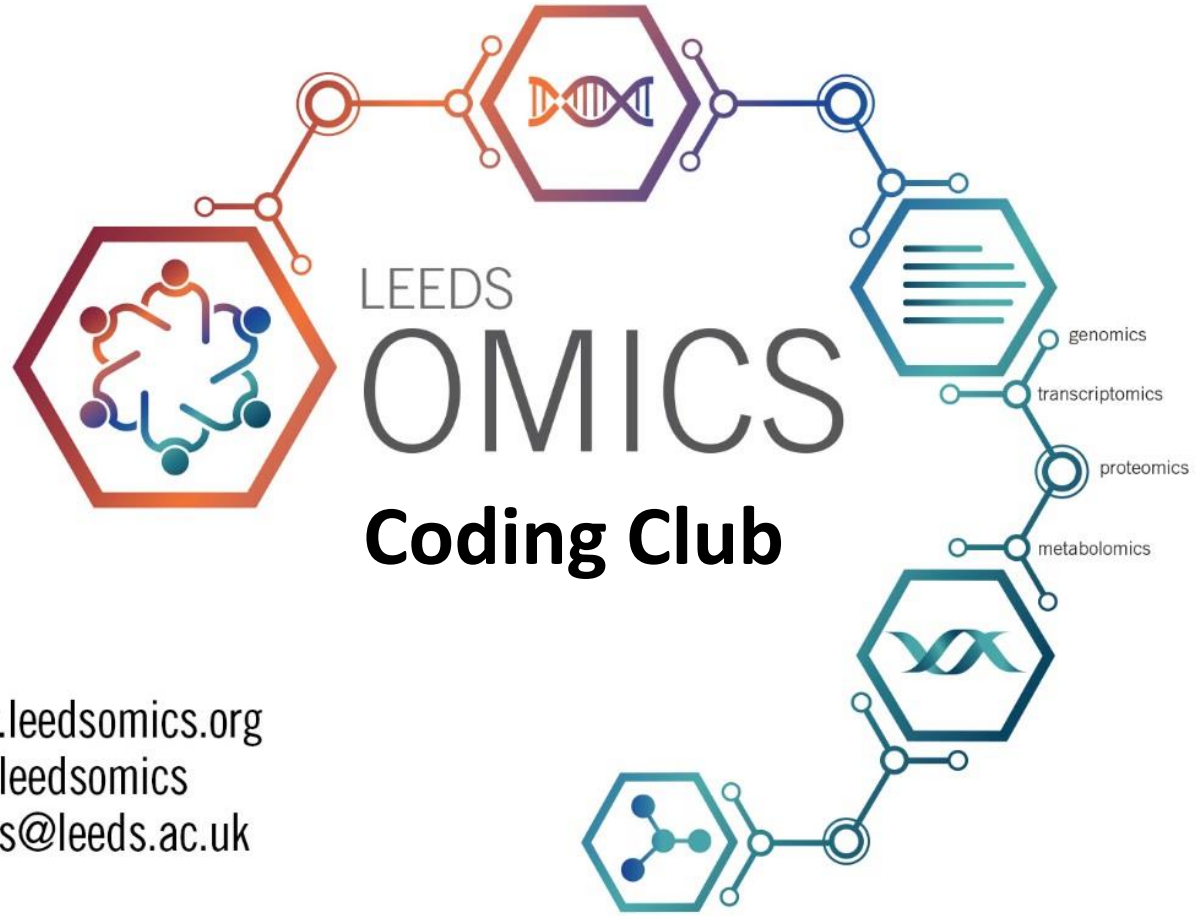


Coding Club

www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

Regular Expression

Club Moderators: Elton Vasconcelos, Peter Mulhair, Euan McDonell, Chew Cheng, and Dapeng Wang

Topics to be addressed - **Survey Result**

Regular Expression

~~$Log_2(FC) = 0$~~ → Not what it is about!

Computer Science: computational language for representing patterns on textual strings
→ interpreted by most programming languages (Java, Python, R, **PERL** …)

Metacharacters

char	meaning
^	beginning of string
$	end of string
.	any character except newline
*	match 0 or more times
+	match 1 or more times
?	match 0 or 1 times; *or*: shortest match
|	alternative
( )	grouping; "storing"
[ ]	set of characters
{ }	repetition modifier
\	quote or special

To present a metacharacter as a data character standing for itself, precede it with \ (e.g. \. matches the full stop character . only).

Repetition

*a**	zero or more *a*'s
a+	one or more *a*'s
a?	zero or one *a*'s (i.e., optional *a*)
a{m}	exactly *m* *a*'s
a{m,}	at least *m* *a*'s
a{m,n}	at least *m* but at most *n* *a*'s
repetition?	same as *repetition* but the *shortest* match is taken

Read the notation *a*'s as "occurrences of strings, each of which matches the pattern *a*". Read *repetition* as any of the repetition expressions listed above it. Shortest match means that the shortest string matching the pattern is taken. The default is "greedy matching", which finds the longest match. The *repetition*? construct was introduced in Perl version 5.

Special notations with \

Single characters

\t	tab
\n	newline
\r	return (CR)
\x*hh*	character with hex. code *hh*

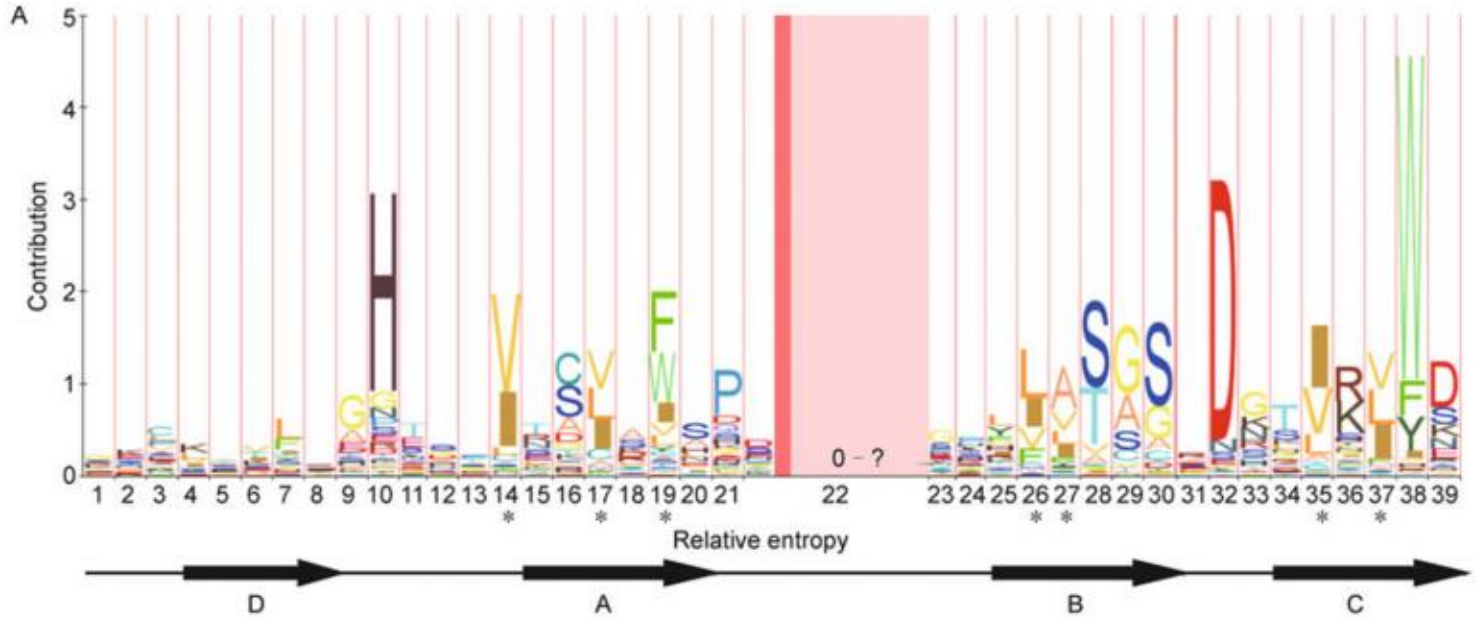
"Zero-width assertions"

\b	"word" boundary
\B	not a "word" boundary

Matching

\w	matches any *single* character classified as a "word" character (alphanumeric or "_")
\W	matches any non-"word" character
\s	matches any whitespace character (space, tab, newline)
\S	matches any non-whitespace character
\d	matches any digit character, equiv. to [0-9]
\D	matches any non-digit character

http://jkorpela.fi/perl/regexp.html



https://www.ncbi.nlm.nih.gov/pubmed/21468892

```
$ wget ftp://ftp.ensembl.org/pub/release-98/fasta/homo_sapiens/pep/Homo_sapiens.GRCh38.pep.all.fa.gz
$ gunzip Homo_sapiens.GRCh38.pep.all.fa.gz
$ perl fasta-single-line.pl Homo_sapiens.GRCh38.pep.all.fa >Homo_sapiens.GRCh38.pep.all.fa2
$ grep -B1 -P '[A-Z]{8}GH[A-Z]{21}D[A-Z]{5}WD' Homo_sapiens.GRCh38.pep.all.fa2 >WD40-proteins-Hsa.fa
```

Download fasta-single-line.pl script from here: https://github.com/eltonjrv/bioinfo.scripts/blob/master/fasta-single-line.pl

Bring your issues on!