Welcome back

Today's schedule

TIME	Wednesday 21st of March 2018
9:00-9:15	Welcome back and intro
9:15-9:30	The Bacterial Analysis Pipeline and batch upload
9:30-9:45	
9:45-10:00	Exercise 5 Computer work - batch upload, pipeline
10:00-10:15	
10:15-10:30	
10:30-10:45	
10:45-11:00	Coffee
11:00-11:15	Wrap-up of computer work
11:15-11:30	CGE and the Global Microbial Initiative (Ole Lund, DTU)
11:30-11:45	
11:45-12:00	Q/A to CGE and GMI
12:00-12:15	Lunch
12:15-12:30	
12:30-12:45	
12:45-13:00	
13:00-13:15	Metagenomics
13:15-13:30	
13:30-13:45	Exercise 6 Computerwork - MGMapper
13:45-14:00	
14:00-14:15	
14:15-14:30	Wrap-up of computer work
14:30-14:45	Coffee
14:45-15:00	Real World Applications (Henrik Hasman, Statens Serum Institut)
15:00-15:15	
15:15-15:30	
15:30-15:45	Q/A to Real World Applications
15:45-16:00	Evaluation and workshop ends

End of day

- Please remember to pick up a course certificate before you leave
- Please evaluate the workshop (https://www.goseqit.com/workshop-evaluation/)

Yesterday's multiple choice questions

5: Which sequencing technique is most likely to have produced the below reads?

```
@QOWBR:00027:00153
CGGTTTTTTTTTTTGCATCATCGGGTAGTCACTGTGTCCCTGATCGAGGGCACGGCGGATGTAGTCCAGTAACAATTCTGCTGAGTTATCGCGC
+
22,2:::::::::&333:;;<444*3<<?A??@BCDDDD=BBB777777+7BBB@BBADBBB999?<?AAB=@@<@377@DDDDDD?DDDDDAA
@QOWBR:00027:00179
CCGTTTTTTTTTTTCCCC
+
?8>>AAABAA@==&333(
@QOWBR:00027:00189
CATTTCAGCATAATCGCCAGCGTATAGCGGAACAGCGGAGAGAAATCGCCCTG
+
323<+2222722*288:9@@@???>?B@B=B=?@@@@3776??@<AA@@@:@?
@QOWBR:00028:00154
CTTTTTTTTTTTCTTTTTTTTTTT
+
CAAB@@ABBBB&@@BBBBA@>>>&
```

Answer:

First, look at the reads and answer if this is short/long reads? And if all reads have approximately the same length?

Then answer, which sequencing technology is characterised by short reads of uneven lengths?

6: A file with raw sequence reads in FASTQ format contains 1.000.000 lines. How many reads does it contain?

A: 4.000.000
B: 1.000.000
C: 500.000
D: 250.000
E: 100.000

Answer:

Since each read in a FASTQ file covers 4 lines, you just have to divide total no. of lines with 4:

1.000.000/4 = 250.000 (D)

8. Below, the output when running whole genome sequence data from an *S. aureus* isolate through the MLST web-service, is shown. Although perfectly matching alleles are identified for all loci, the sequence type is reported as unknown. What is the likely cause of this?

MLST-1.6 Server - Typing Results

Sequence Type: ***Unknown ST***

Locus	% Identity	HSP Length	Allele Length	Gaps	Allele
arcc	100.00	456	456	0	*arcc-7*
aroe	100.00	456	456	0	*aroe-6*
glpf	100.00	465	465	0	*glpf-28*
gmk	100.00	417	417	0	*gmk_-27*
pta	100.00	474	474	0	*pta_-11*
tpi	100.00	402	402	0	*tpi_-40*
yqil	100.00	516	516	0	*yqil-27*

A: The chosen MLST configuration does not match the species. This is likely not a *S. aureus* isolate.

B: This can only be explained by a bug in the method.

C: This could never occur.

D: This is likely due to too low initial quality of the sequence data. This can be confirmed by examining the N50 value of the draft genome.

E: Although all the alleles have been seen before and are included in the MLST database, the *combination* of alleles is previously unseen and hence reported as "unknown".