Quality assessment

Raw read QC

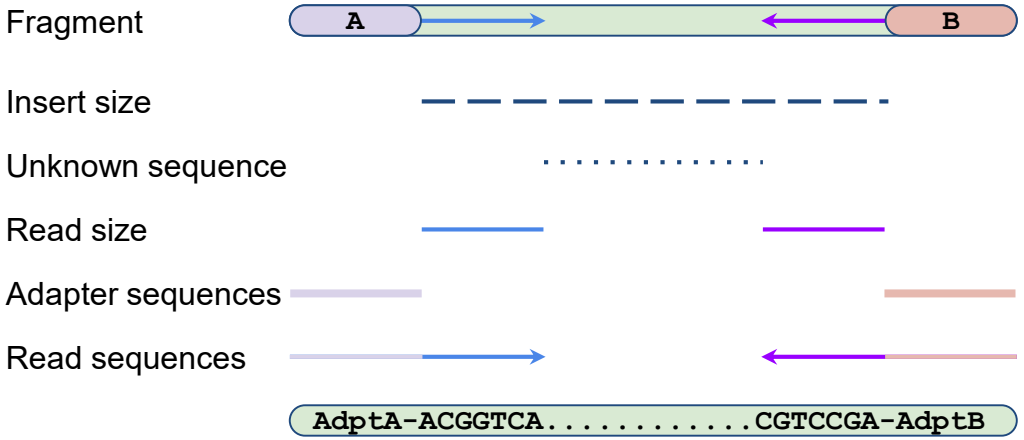
Aims ITOLs

- Understand the relation between read sizes and library preparation
- Overview of read count in relation to sequencing depth and error assessment
- Insights into overall read quality on the basis of base quality scoring
- Introduction to assessment of base contents.

What matters to quality?

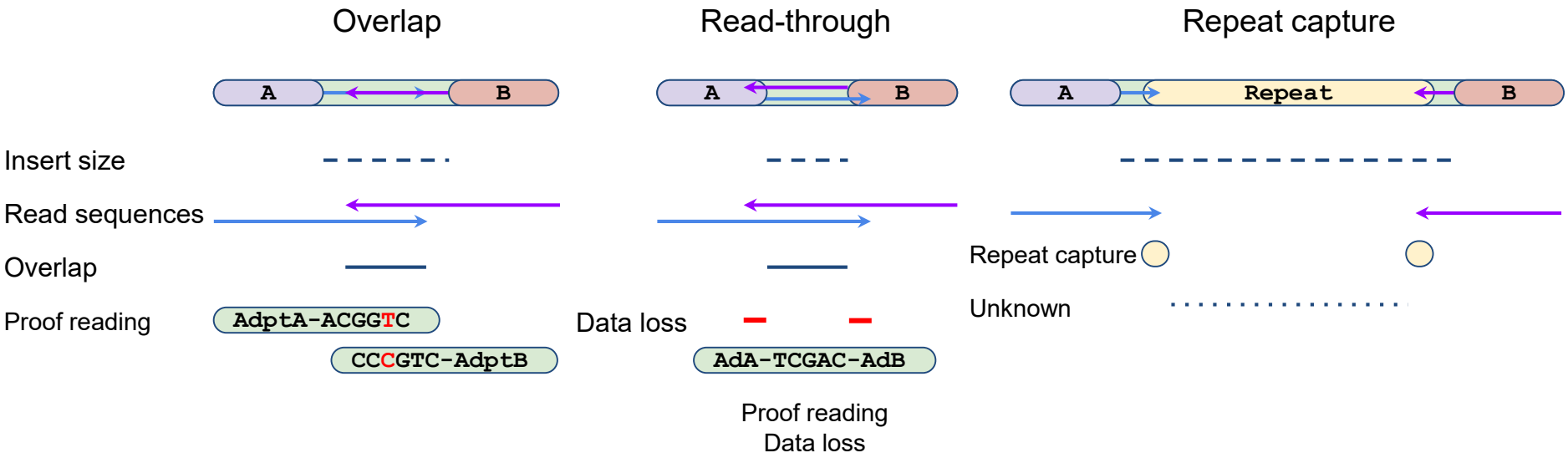
- Read size
- Read count
- Base quality score
- Base contents

Read size

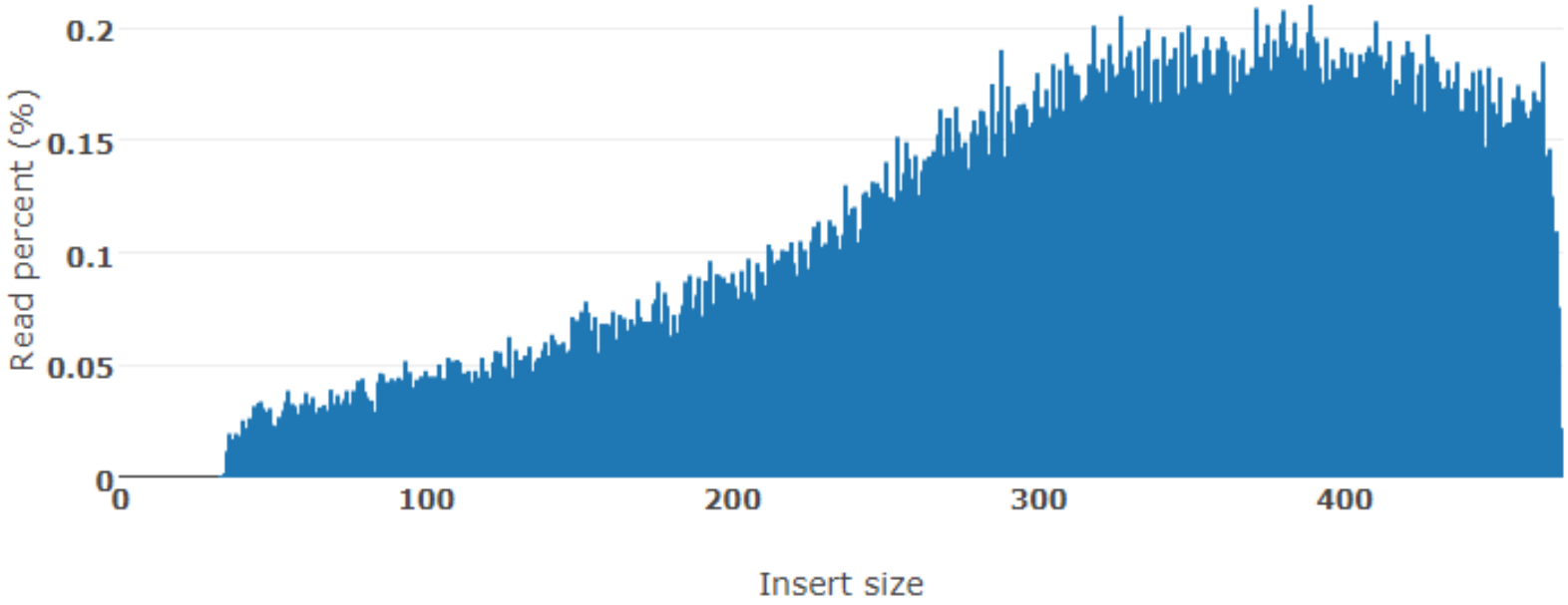



```
        @sample1_Mate1
AdptA-ACGGTCA
        ...
        @sample1_Mate2
AdptB-AGCCTGC
        ...
```

Scenarios


Overlap
Read-through
Repeat capture
A
B
A
B
A
Repeat
B
Insert size
Read sequences
Overlap
Proof reading
AdptA-ACGGTC
CCCGTC-AdptB
Data loss
AdA-TCGAC-AdB
Proof reading
Data loss
Repeat capture
Unknown

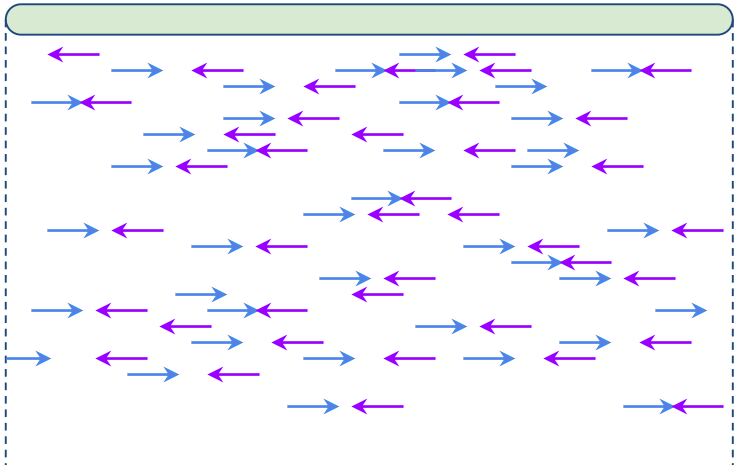
What library size do you think is aimed for?



Read size should correlate with the library strategy

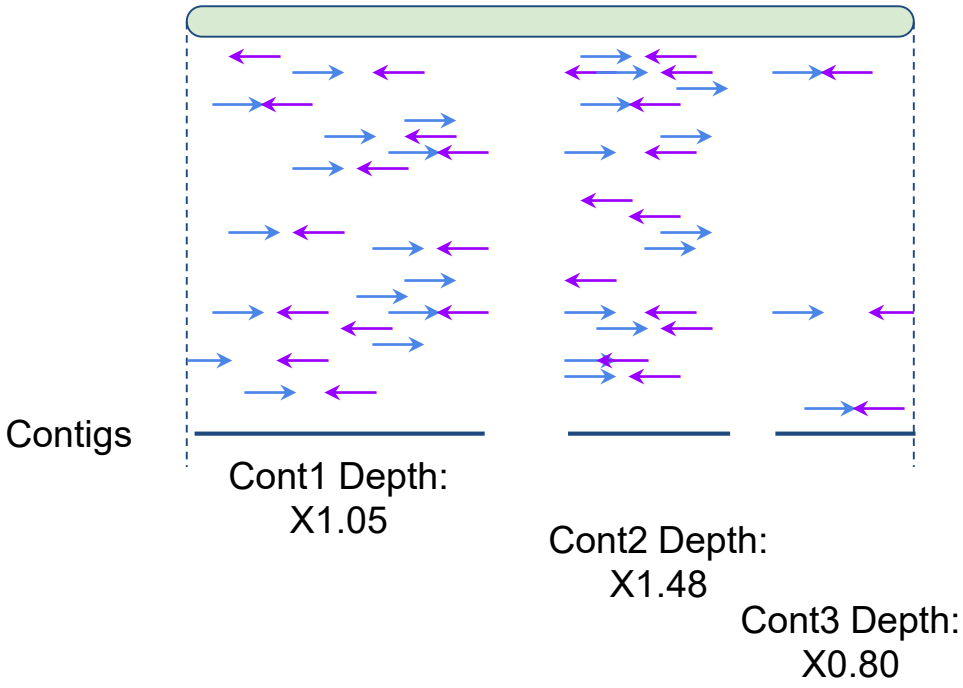
Read count

Read count



Depth: X2.8

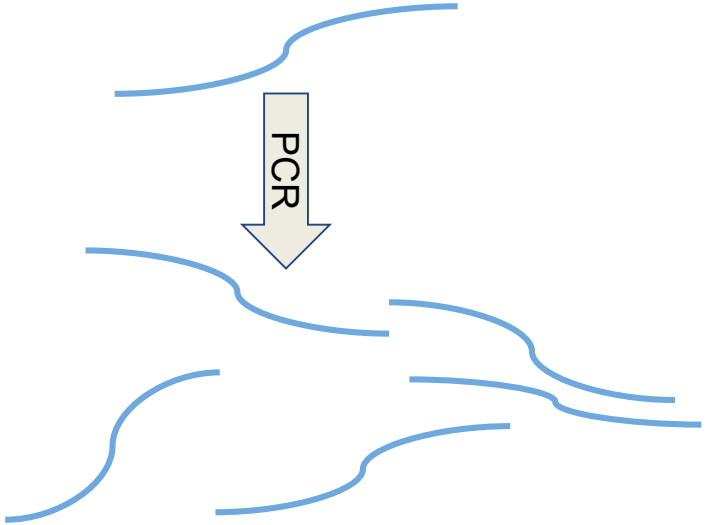
Read count

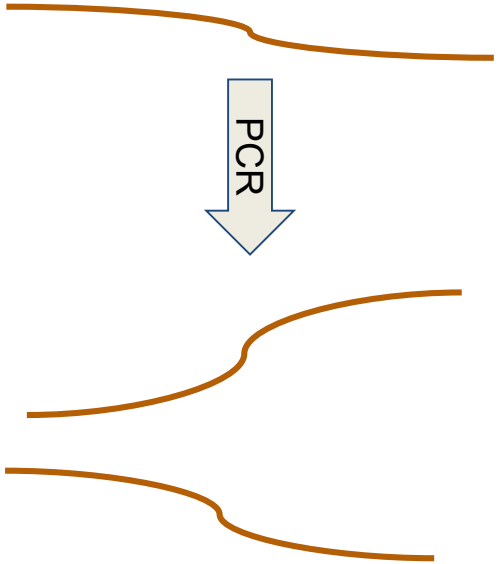


Read counts dictates the sequencing depth.

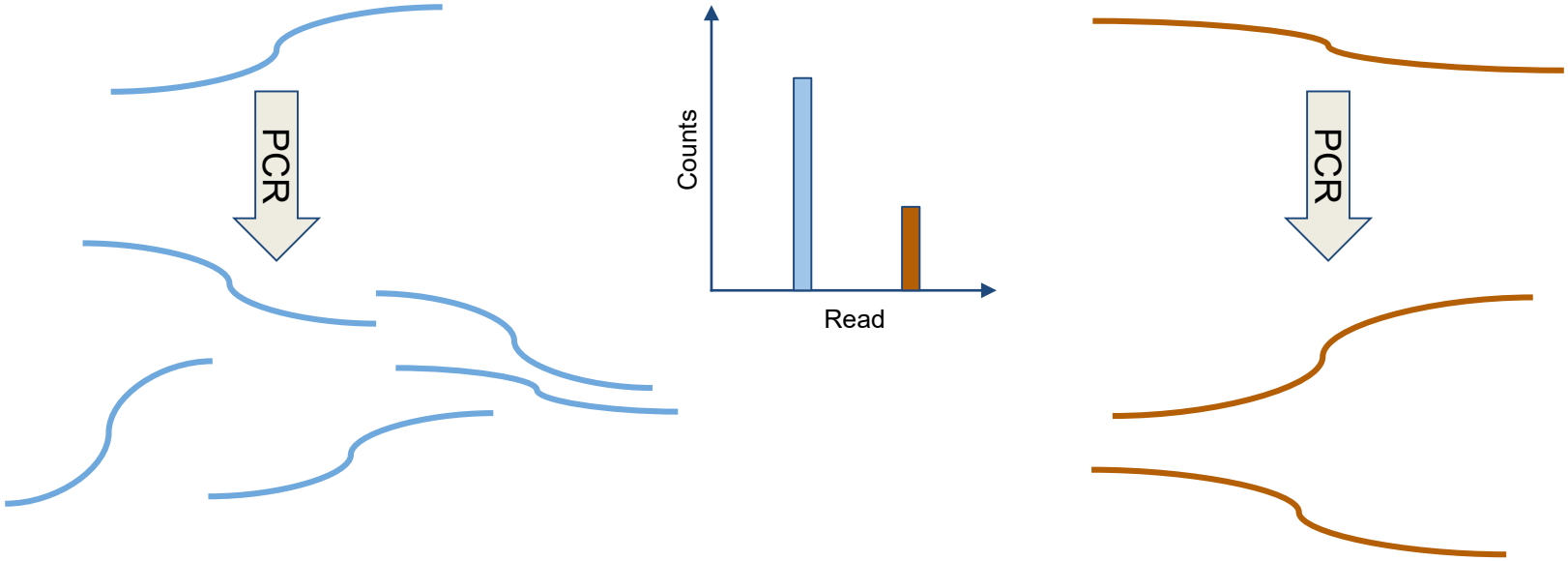
Overrepresentation

Overrepresentation

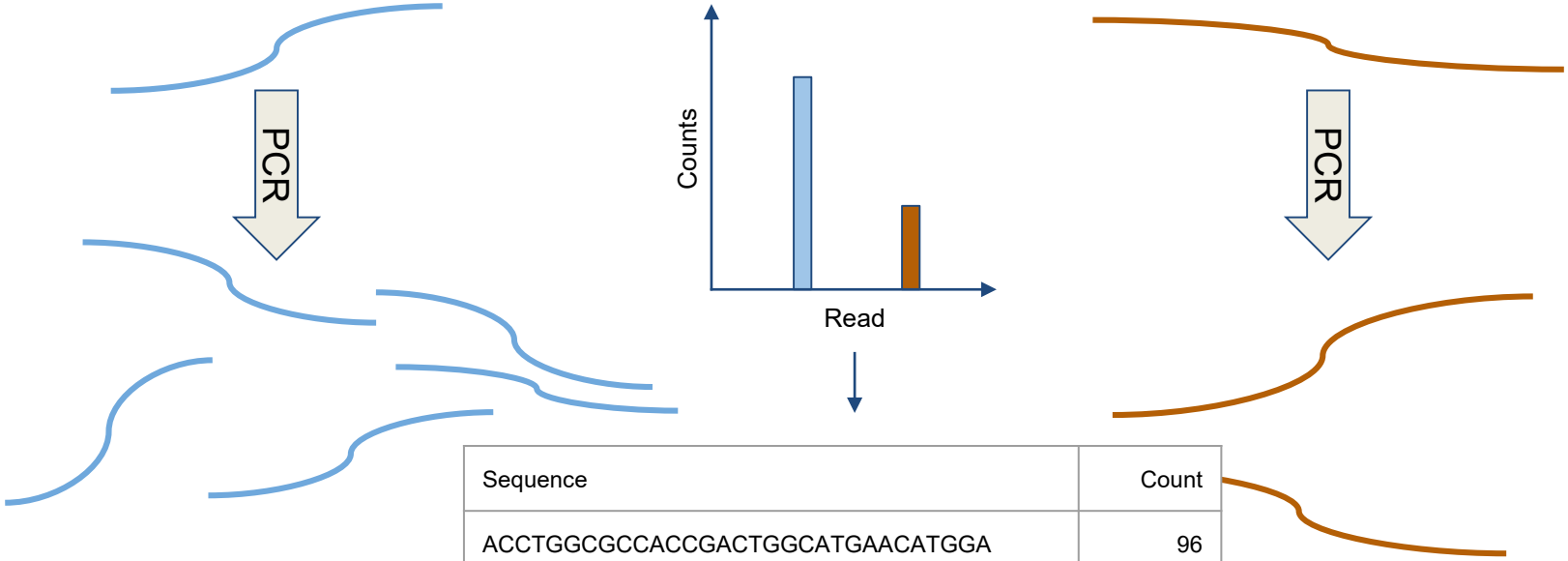




Overrepresentation



Overrepresentation



Sequence	Count
ACCTGGCGCCACCGACTGGCATGAACATGGA	96
NNNNN	57
ACCTTGGC	48

Pop quiz time!!!

Sequence	Count
ACCTGGCGCCACCGACTGGCATGAACATGGA	96
NNNNN	57
ACCTTGGC	48

- Adapter?
- Biological?
- Technical?

Sequence	Count
ACCTGGCGCCACCGACTGGCATGAACATGGA	96
NNNNN	57
ACCTTGGC	48

- Adapter?
- Biological?
- Technical?

Sequence	Count
ACCTGGCGCCACCGACTGGCATGAACATGGA	96
NNNNN	57
ACCTTGGC	48

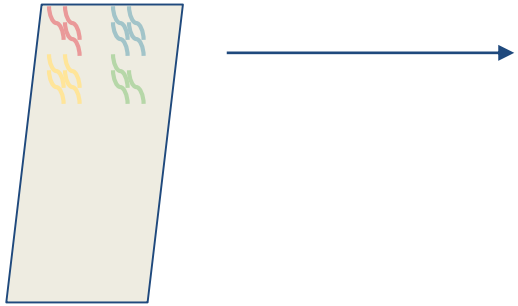
- Adapter?
- Biological?
- Technical?

Sequence	Count
ACCTGGCGCCACCGACTGGCATGAACATGGA	96
NNNNN	57
ACCTTGGC	48

- Adapter?
- Biological?
- Technical?

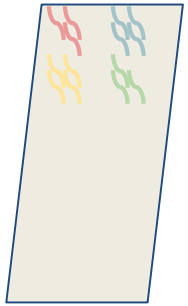
Overrepresented reads indicates issues during **library prep** or **sequencing**.

Base quality



High quality

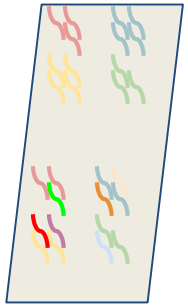
Base quality



High quality

ACCGTGGTTA
+

Base quality



High quality

Low quality

ACCGTGGTTA
+

Overall base quality

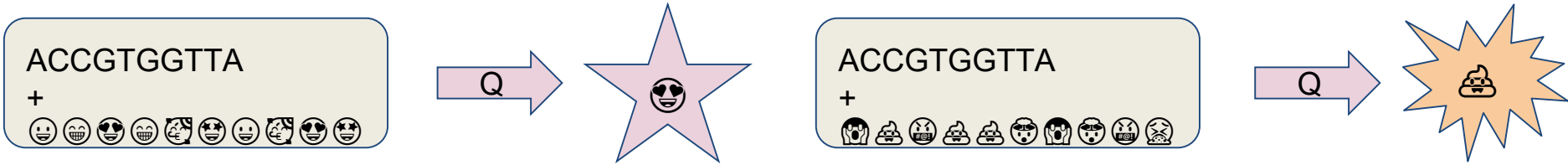
ACCGTGGTTA
+

Q

ACCGTGGTTA
+

Q

Overall base quality



What to do with the crappy bases and reads?

What to do with the crappy bases and reads?

Throw it out!!

Base quality and read quality usable to extract highest quality sequencing data.

Base contents

>NC_007795.1 Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome
CGATTAAAGATAGAAATACACGATGCGAGCAATCAAATTTCATAACATCACCATGAGTTTGGTCCGAAGC
ATGAGTGTTTACAATGTTCGAACACCTTATACAGTTCTTATACATACTTTATAAATTATTTCCCAA
...

Base contents

>NC_007795.1 Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome
CGATTAAAGATAGAAATACACGATGCGAGCAATCAAATTTCATAACATCACCATGAGTTTGGTCCGAAGC
ATGAGTGTTTACAATGTTCGAACACCTTATACAGTTCTTATACATACTTTATAAATTATTTCCCAA
...

Base contents

>NC_007795.1 Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome
CGATTAAAGATAGAAATACACGATGCGAGCAATCAAATTTCATAACATCACCATGAGTTTGGTCCGAAGC
ATGAGTGTTTACAATGTTCGAACACCTTATACAGTTCTTATACATACTTTATAAATTATTTCCCAA
...

Base	Count	Percentage
A	49	36.03
C	26	19.12
G	19	13.97
T	42	30.88

GC contents

S. aureus: ~33.09%

E. coli: X%

... Y%

...

You survived!

Congratz...

Acknowledgements

The creation of this training material was commissioned by ECDC to Statens Serum Institut with the direct involvement of Kasper Thystrup Karstensen
The revision and update of this training material was commissioned by ECDC to Statens Serum Institut with the direct involvement of Kasper Thystrup Karstensen, Astrid Rasmussen, and Søren Hallstrøm