Aligning raw reads directly

Philip Clausen
Recap from yesterday
Minor variants and Malaria
Mapping and aligning raw reads directly
BLAST recap

Yesterday we saw that BLAST used seed and extend to speed up alignment, where the query sequences were indexed on the fly.

This was smart when BLAST was developed, as the query sequences were limited while the target databases were growing.

This gave BLAST the run-time of $O(m + n)$, under the assumption that the query could be indexed in one go.
The BLAST problem

\[ m + n < m \frac{n}{b} + n \]
Next generation alignment

\[ m = \text{total number of bases in database sequences} \]
\[ n = \text{total number of bases in input raw reads} \]

Assumptions:

- \( m \) is rarely changing
- We have a lot of samples with size \( n \)

How can we then improve our alignment performance.

**Baseline performance:**
\[ m \frac{n}{b} + n \]

**Hint:**
The optimal performance can be done with \( n \) operations.
\[ n << m \frac{n}{b} + n \]

**Solution:**
We could index our database once, and then reuse that.
**K-Mer Alignment (KMA)**

1. Mapping of sequences.

2. Alignment of sequences.
Mapping
Mapping
Alignment

1. Seed and extend

2. Needleman-Wunsch
Mapping and aligning raw reads directly
Oh God! they are targeting us

Guys!...Guys! CALM DOWN! We are the target, not you!

Oh! shit!