KMA updates and computer-problems

Philip T.L.C. Clausen

Recap

ConClave

Three steps

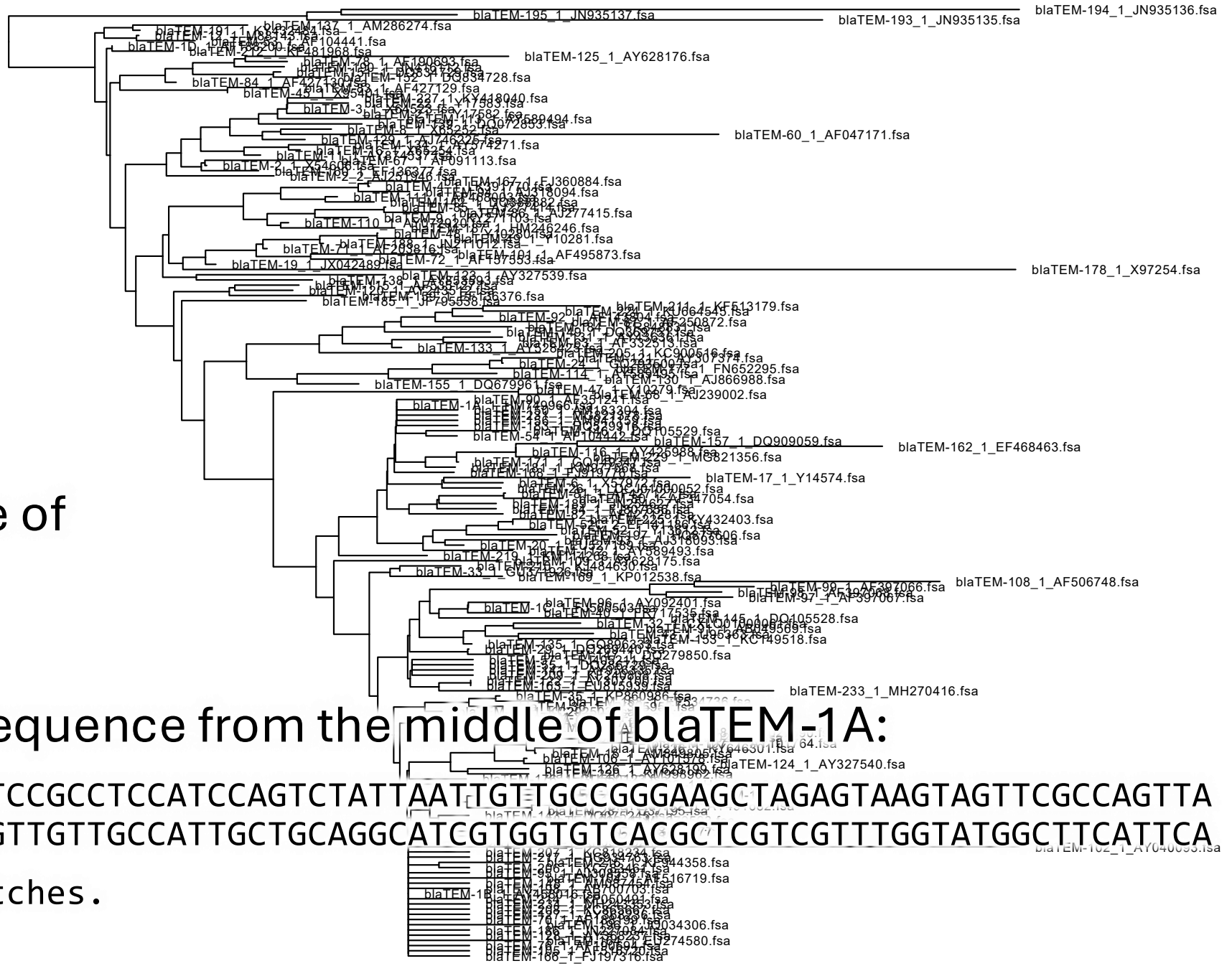
$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \tag{1}$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \tag{2}$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \tag{3}$$

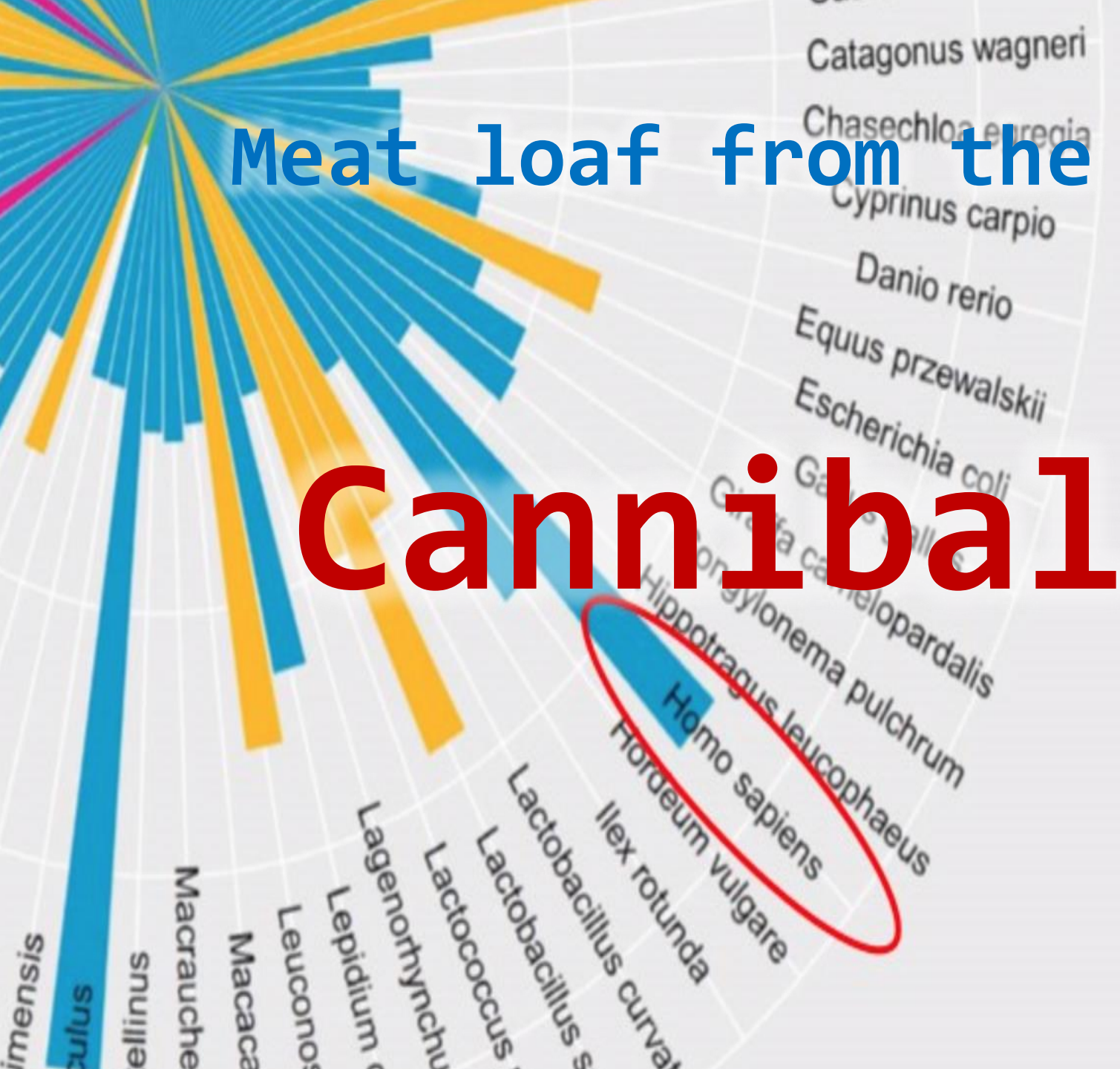
WHY?

- 183 blaTEM variants
- Max distance of 38 SNPs
- 150 bp subsequence from the middle of blaTEM-1A:
- GTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTA ATAGTTTGCGCAACGTTGTTGCCATTGCTGCAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCA
- 131 perfect matches.



Meat loaf from the FLI canteen

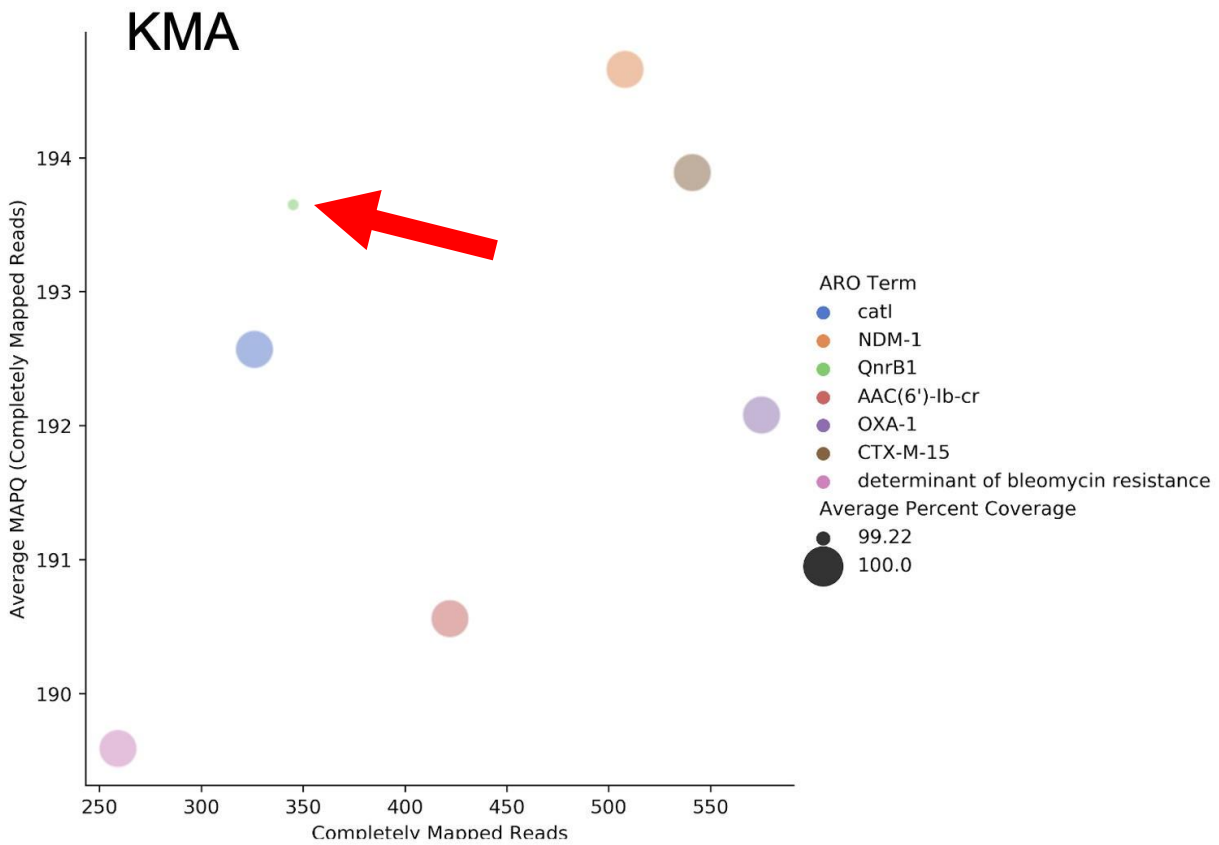
Cannibalism…

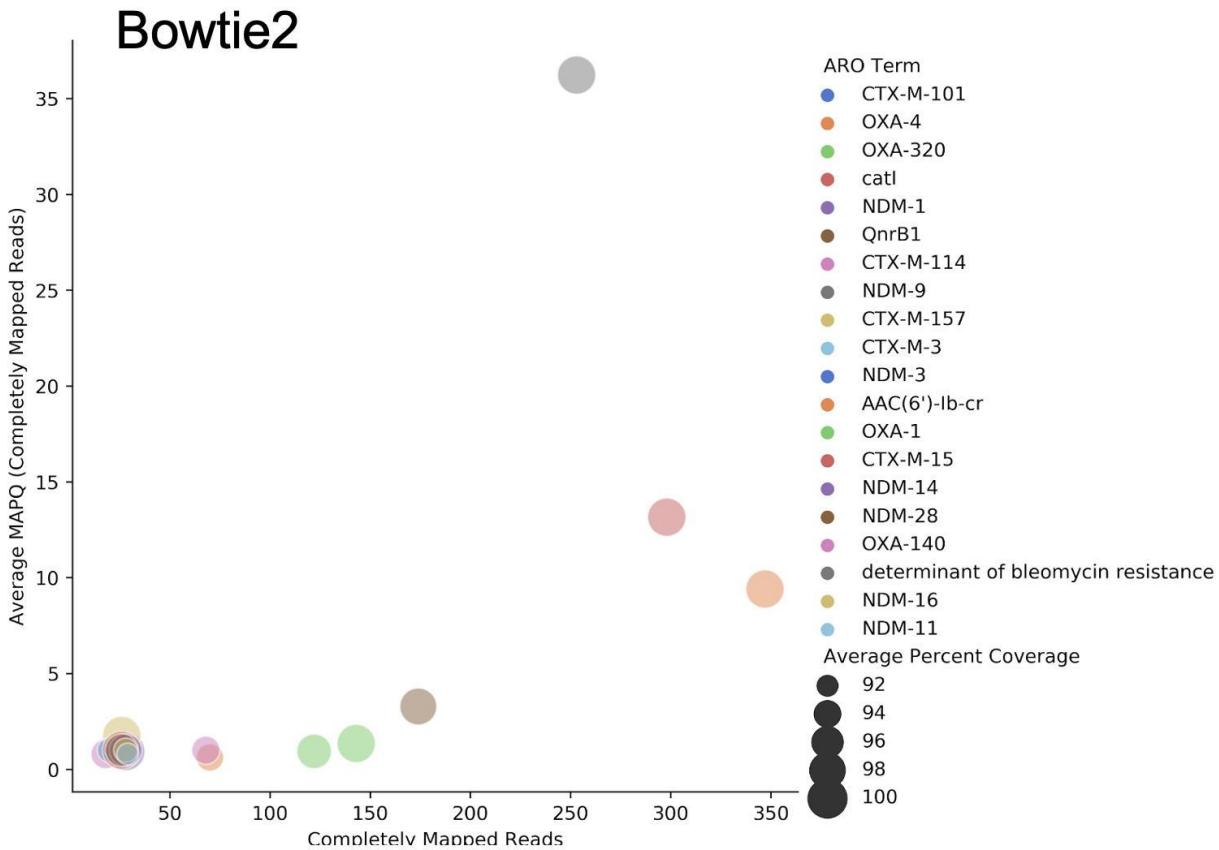


Höper D. et al. 2018

Does it work?

CARD/RGI test, 7 in 7 out





CARD/RGI GitHub

Limitations

- Sometimes miss-assign between closely related references. E.g. assigns all reads to *catl* instead of *catB3*.

- *What if resolution of the reference database is higher than the accuracy of the query sequences.*

Illumina

- High accuracy
- Usually <= 1 error pr. read
- Okay to assume zero errors in most cases
- Consensus sequences are usually correct, and can be used to reassign the hits.

ONT

- High error-rate
- Some say ~1%, usually 5-10%.
- But one read easily covers an entire gene.

ONT implications

- 1% error -> ~10 errors pr. gene
- ResFinder, CARD and AMRFinder has genes placed **1** SNP apart
- **Random errors will match actual alleles -> true allele will not be among the best scoring hits!**

Proximity scoring

- Consider hits with scores close to the best hit as candidate hits too.

- Weigh each hit according to the score they get.
  - I.e. higher scoring hits gets a higher weight.
  - E.g. A hit with a score of 100 will weigh 100, and a hit with a score of 90 will weigh 90.

ConClave Proximity scoring

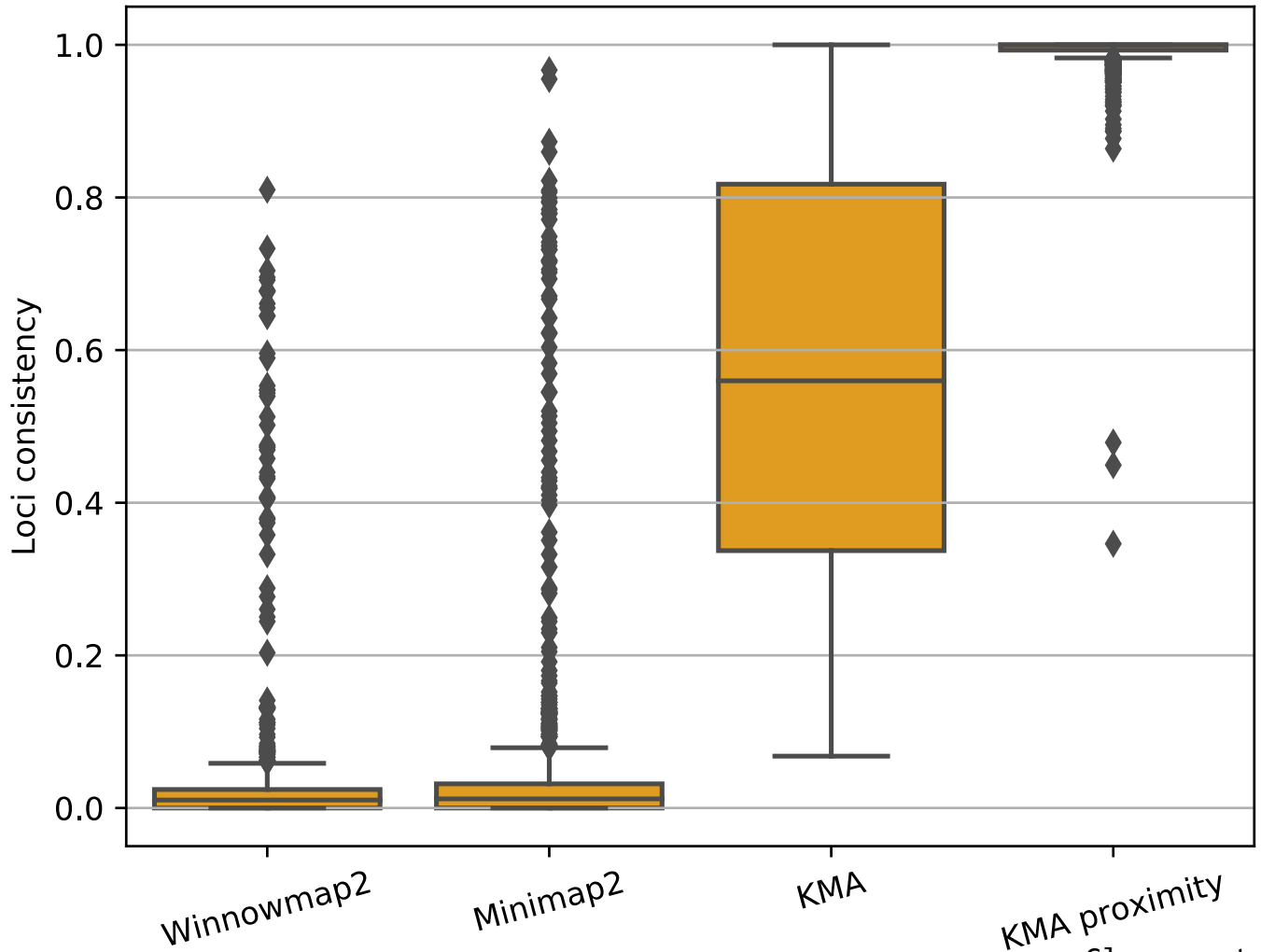
$$T_m(q) \in \underset{t \in T}{\operatorname{argmax}} \left\{ \begin{matrix} f(q,t) \\ \dfrac{f(q,t)}{\epsilon} \end{matrix} \right\} \quad \epsilon \in (0;1] \tag{1}$$

$$C(t) \in \sum_{q \in Q} \begin{cases} f(q,t) & \tau \leq f(t_k) \wedge t \in T_m(q) \\ 0 & else \end{cases}. \tag{2}$$

$$S_q \in \underset{t \in T_m(q)}{\operatorname{argmax}} \{ C(t) \} \tag{3}$$

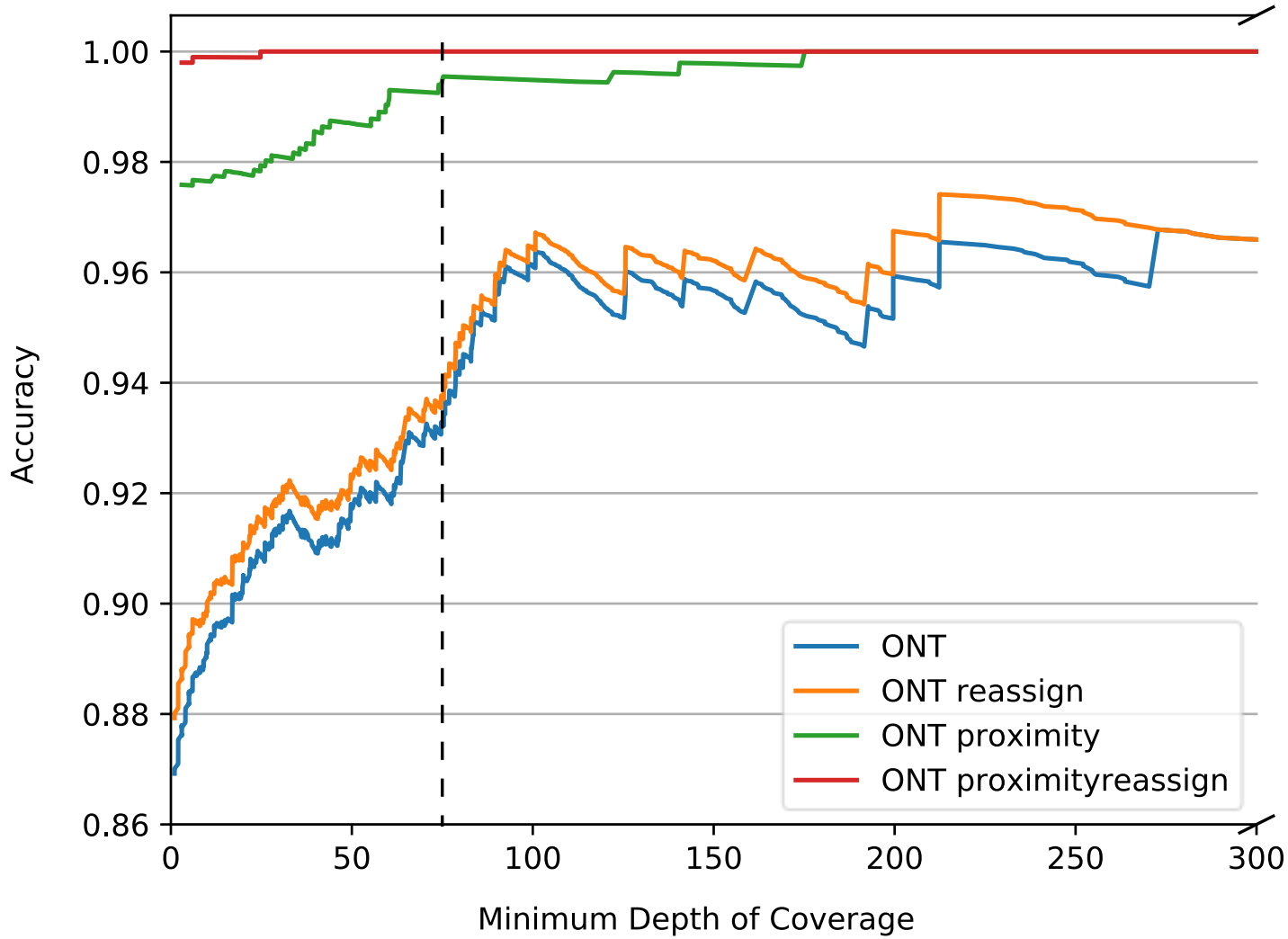
Does it work?

- MLST is at higher resolution than ResFinder
- We collected 142 samples sequenced on both Illumina and ONT
- Compared to state of the art aligners



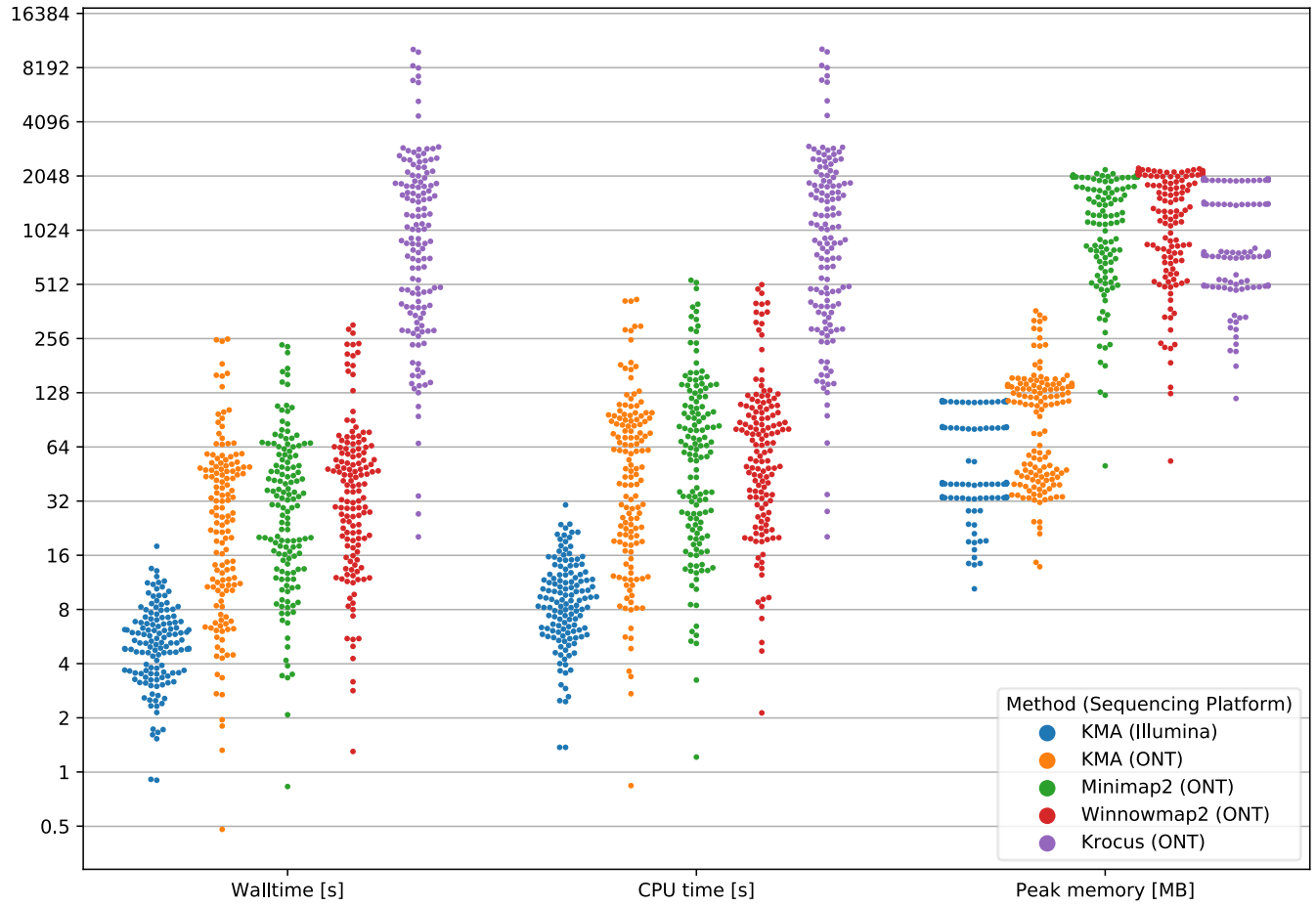
Clausen et al. 2025

Is it accurate?



Clausen et al. 2025

But it needs go over the input twice…



Clausen et al. 2025

But it needs go over the input twice…

- All hits needs to be saved.
  - One big temporary file.
- Then they need to be sorted.
  - Requires multiple smaller temporary files.
- Limitations:
  - Not enough space on device.
  - High memory consumption when sorting.
  - Too many files.

Basics of a computer

- CPU
  - Does the computation
- Memory
  - Stores what is frequently needed
  - Has a limited storage space, but is quick is access.
- Disk
  - Stores what is infrequently needed
  - Has a large storage space, but is slow to access.

Basics of a computer

- CPU
  - Is what you have in your hands
- Memory
  - Is what you have at your desk in front of you
- Disk
  - Is what you have in the closet at the other end of the room

Limitations

- Not enough space on device:
  - Default temporary locations usually have limited space.
  - Not a problem with small databases.
  - Big problem with large databases.

- Solution: Set the “-tmp” option,
  use local disks if possible.

Limitations

- High memory consumption when sorting.
  - Sorting is carried out in chunks which needs to stored in memory.
  - Decreasing the chunk size decreases memory requirements at this step, but increases the number of files.

- Solution: Set the “-mf” option down.

Limitations

- Too many files.
  - Might happen when a large query file is searched with many hits.
  - Or when too many programs create files at the same time.

- Solution: Set the “-mf” option up.

Nebula

- Okay hands (CPU) to work with.
- Desk (memory) is large enough for most stuff.
- Closet is big.
  - But it is placed inside a Walmart.
  - The highway to it is as undersized as E47 during rush hour.
  - And it is not large enough store irrelevant results.

Nebula

- Good computer for compute.
  - E.g. great for phylogenies and AMR detection

- Not so good when frequent disk access is required.
  - E.g. if too many I/O heavy jobs is running at once, it is like accessing that closet within Walmart on Black Friday.
  - Running more than one heavy KMA job at once is like splitting the bill into hundreds when paying at Walmart.
  - Running several assemblies or large BLAST jobs at once is like accessing Netto after Mette closed Denmark.
  - Reading and writing files without a buffer is like shoping without a cart.

In other words.

Not knowing what you are doing

Knowing what you are doing

Computerome

- Hands (CPU) like nebula.
- Desk (memory) like nebula.
- Closet is big.
  - But with better access compared to Nebula.
- It is a rental.
  - I.e. expensive to use.
  - Restricted access.

Laptop

- Okay hands (CPU) to work with.
  - Usually as fast as the ones on the cluster computers.
- Desk (memory) is small.
  - But large enough for smaller things.
- Closet is limited.
  - But it is placed closely, with autobahn-like connections.

Results shown earlier where all computed on a laptop.

New cluster

- Real handyman hands
  - But there is a limit to how many.
- Enough desk space for almost anything
  - But not everything.
- Big enough closet for now
  - With highway-like connections.
  - And autobahn-like connection to scratch disks.

I have tried to make KMA easier

- Presets:
  - -ont Preset when analyzing 3rd generation (ONT) data
  - -ill Preset when analyzing 2nd generation (Illumina) data
  - -asm Preset when analyzing assembly data

- -1t1
  - One query sequence should only match one reference sequence

- -mem_mode
  - Use mapping scores instead of alignment scores for ConClave, which reduces the required memory (smart for large DBs).

- -reassign
  - Re-assign imperfect matches, like the *catl* -> *catB3* earlier.

Remember the bi-weekly bioinformatics meetings, where anyone with computer- or bioinformatics-related questions can come by.