

Benchmark of multiple sequence alignment (MSA) methods applied to third-generation long reads

Coralie ROHMER¹, Hélène TOUZET¹, Antoine LIMASSET¹
1. Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRISTAL, F-59000 Lille, France

Which MSA tools should you use?

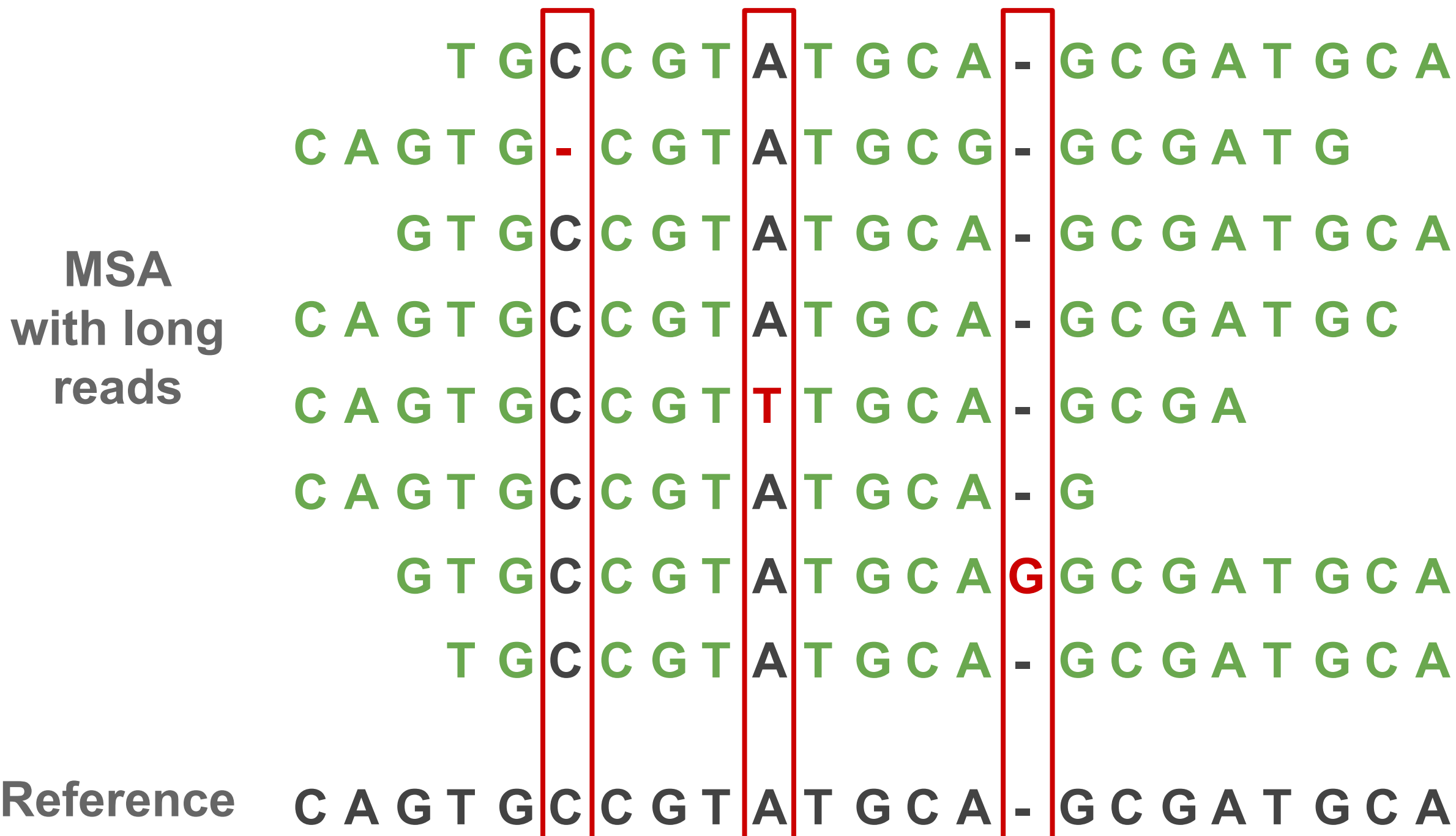


To find the poster

In what extend can we use the classical MSA tools to extract the signal present in the long reads?

Context

Third-generation sequencing is radically changing the way we think about accessing genomic information because it allows for long reads of tens or hundreds of kilobases. However, these reads have a large amount of erroneous bases, including deletions and insertions. Multiple sequence alignment (MSA) tools can identify and correct these errors. However, MSA tools were not initially designed for this type of data. How well can existing MSA tools adapt to the error profile and length of long reads? What is the best tool to use in this context?



Difference between reads from 2nd and 3rd generation

	Short read (2nd)	Long read (3rd)
Size	100 – 300 pb	10 – 100 kb
Error rate	<1%	5 to 17%
Error type	only substitutions	lots of insertions and deletions, some substitutions

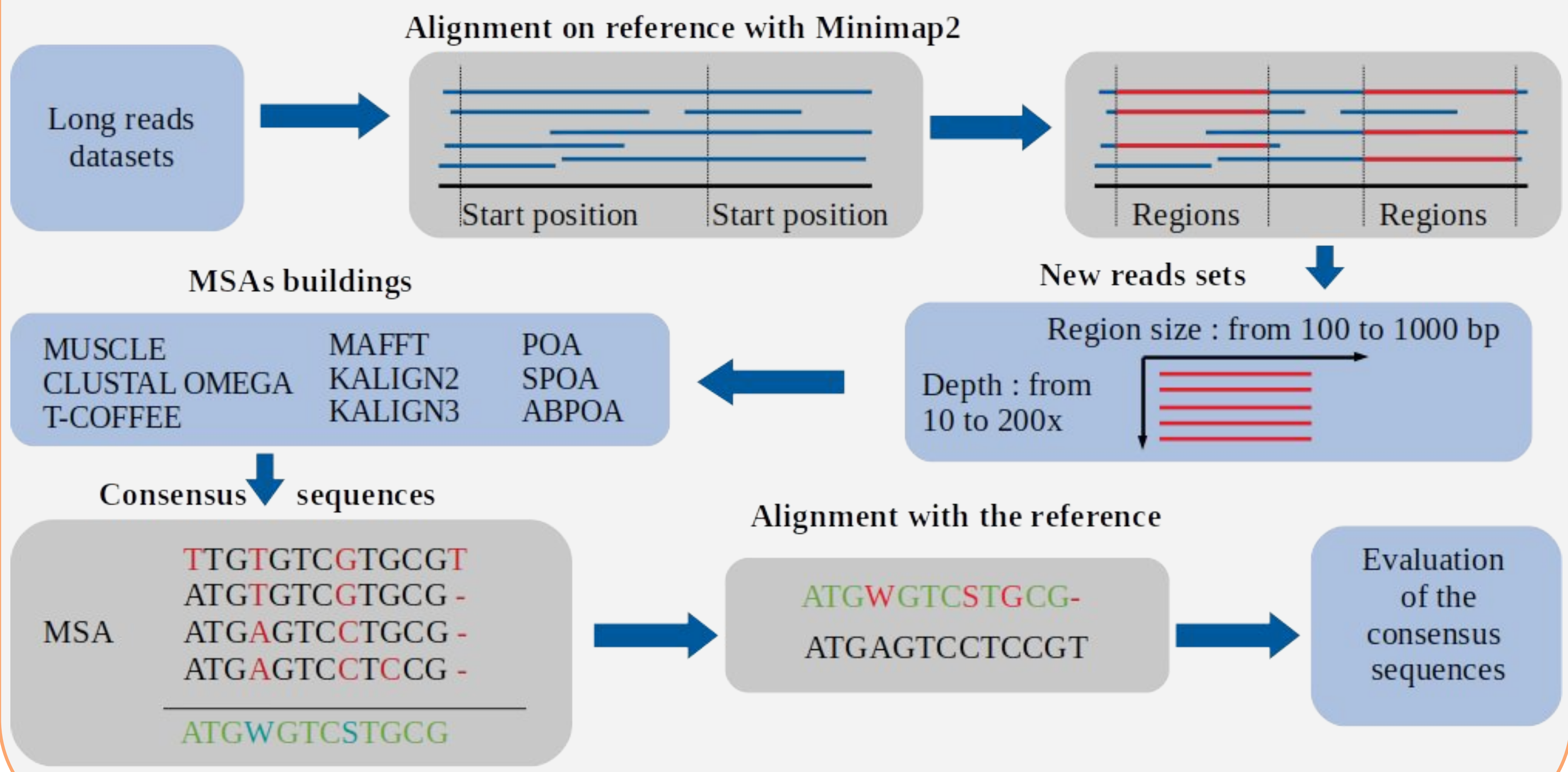
Multiple Sequence Alignment Tools

Created for	Genes	Evolutionary distance	Genomic variation	No redundancy	Substitutions, insertions and deletions
How we want to use it	Long reads	Sequencing error	High Error Rate	Redundant sequences	Mainly insertions and deletions

Metrics

- Identity rate
- Error rate
- Match rate
- Ambiguous characters rate (IUPAC)
- Type of error
- Time
- Memory
- Sequences length

Analysis pipeline



MSA tools

Muscle (Mu)	Mafft (Ma)	Poa (Po)
Spoa (Sp)	Abpoa (Ab)	Kalign2 (K2)
Kalign3 (K3)	Clustal Omega	T-Coffee

Technical



Repository

- Python
- Snakemake
- Conda

<https://gitlab.cristal.univ-lille.fr/crohmer/msa-limit.git>

Results:

Except for T-Coffee which is too expensive and Clustal Omega which does not have consistent results on this types of data, the other MSA tools are all usable and are able to reduce the noise. The size of the region hardly affects the quality of the results, only the time and memory.

Evaluation of the impact of the reads error rate

Region size: 500bp, Depth: 50x

Ranking	1st	2nd	3rd	4th	5th	6th	7th	min	max	mean
Human (ER: 6%)	K2	Sp	Ma	Mu	Ab	Po	K3	76.4%	100%	99.6% (+/- 2.1)
Yeast (ER: 10%)	Sp	Mu	Ma	K2	Ab	Po	K3	84.8%	100%	98.6% (+/- 1.4)
E. coli (ER: 16%)	Sp	Mu	Ma	Po	Ab	K2	K3	88.0%	99.6%	97.0% (+/- 1.6)

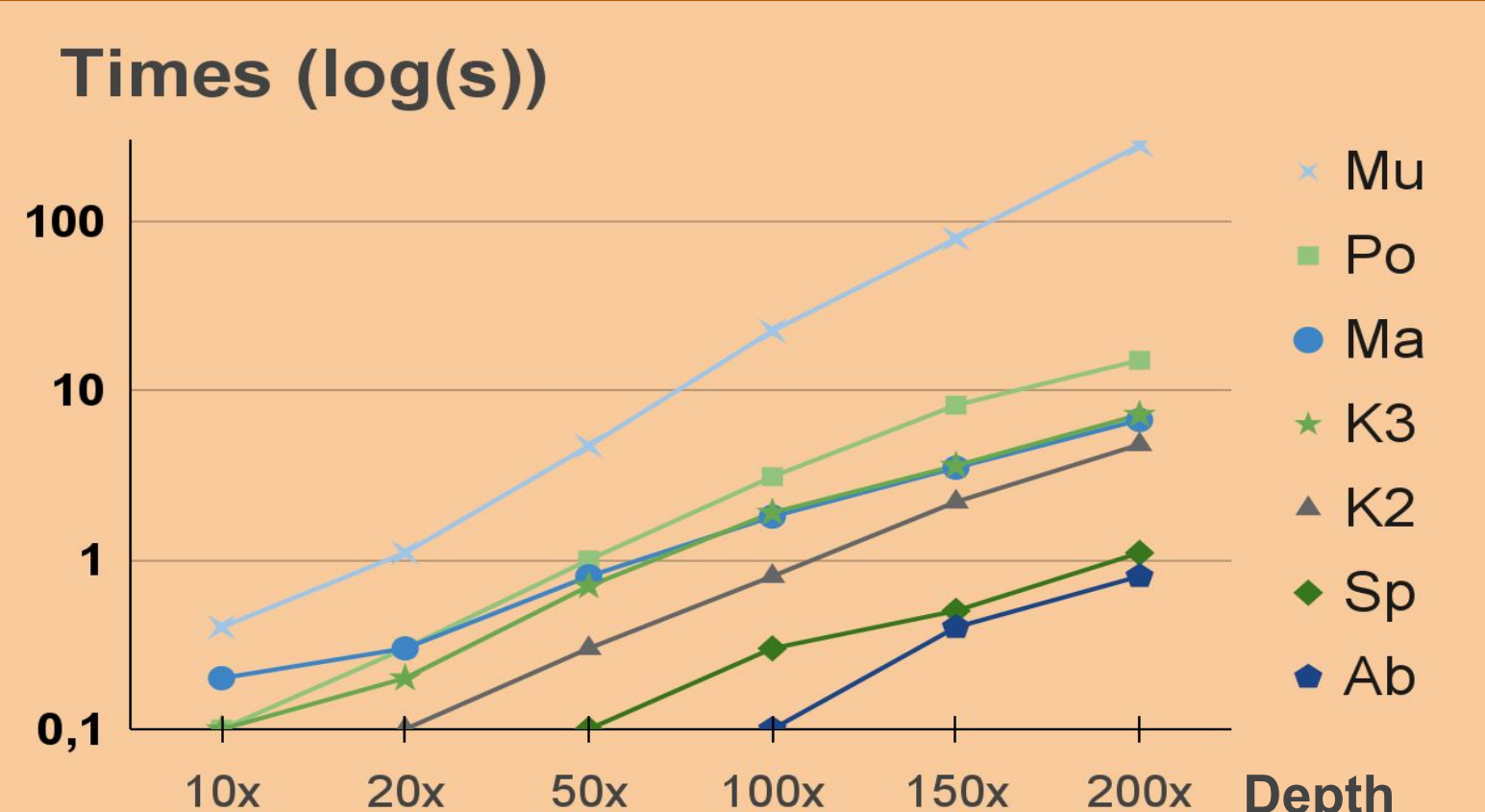
Evaluation of the impact of the sequencing depth

Region size: 500bp, Data set: Human (ER: 6%)

Ranking	1st	2nd	3rd	4th	5th	6th	7th	min	max	mean
10x	Sp	Po	Ab	Mu	Ma	K2	K3	74.8%	100%	99.5% (+/- 2.3)
20x	Sp	Mu	Po	Ab	Ma	K2	K3	72.6%	100%	99.5% (+/- 2.6)
50x	K2	Sp	Ma	Mu	Ab	Po	K3	76.4%	100%	99.6% (+/- 2.1)
100x	K2	Ma	Sp	Ab	Mu	K3	Po	75.1%	100%	99.6% (+/- 2.3)

Time comparison

Region size: 500bp, Data set: Human (ER: 6%)



ER: Error rate, Mu: Muscle, Ma: Mafft, Po: Poa, Sp: Spoa, Ab: Abpoa, K2: Kalign2, K3: Kalign3
In the first two tables, tools are ranked according to the identity rate between the consensus sequence obtained and the reference sequence over 100 regions.