Fast Multiple Sequence Alignment

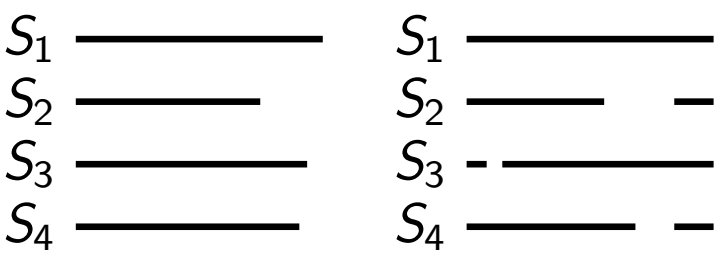
Fabian Klötzl

MPI for Evolutionary Biology, Plön

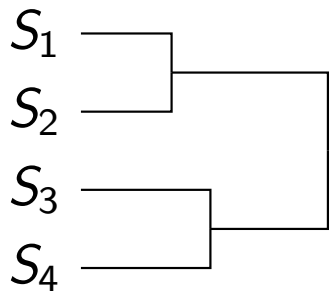
IMPRS selection week, 2016-06-28

Reconstructing Evolution

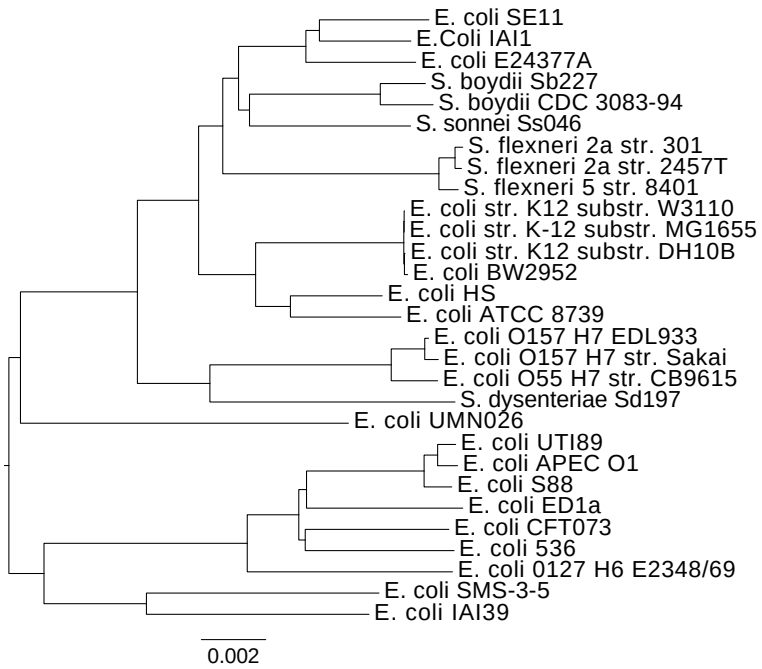
Genomes $\xrightarrow{\text{slow}}$ Alignment $\xrightarrow{\text{fast}}$ Distance Matrix $\xrightarrow{\text{fast}}$ Tree



	S_1	S_2	S_3	S_4
S_1	0			
S_2	0.1	0		
S_3	0.4	0.4	0	
S_4	0.4	0.4	0.2	0



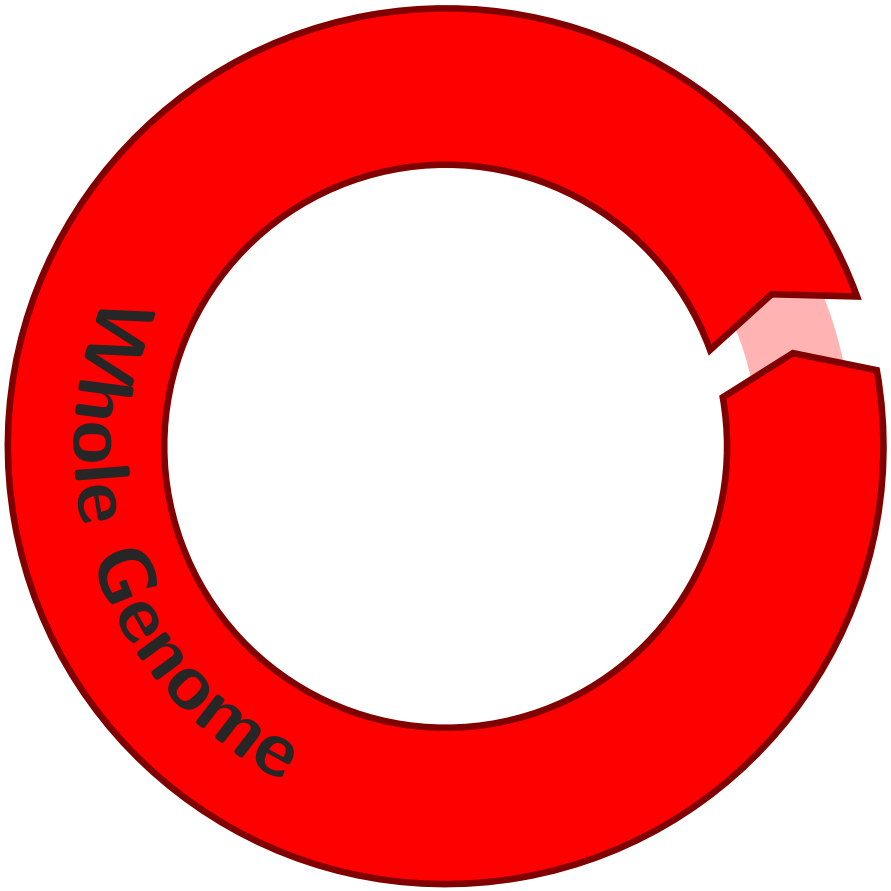
Multiple Sequence Alignment



Eco29

- 29 *E. coli/ Shigella* genomes
- file size: 138 MByte
- Multiz & TBA (UCSC Genome Browser): 1 d 3 h
- Mugsy: 2 h 49 min
- andi: 45 s
- Muscle, MAFFT, HAlign exceed memory limit
- KAlign: 4 weeks and running

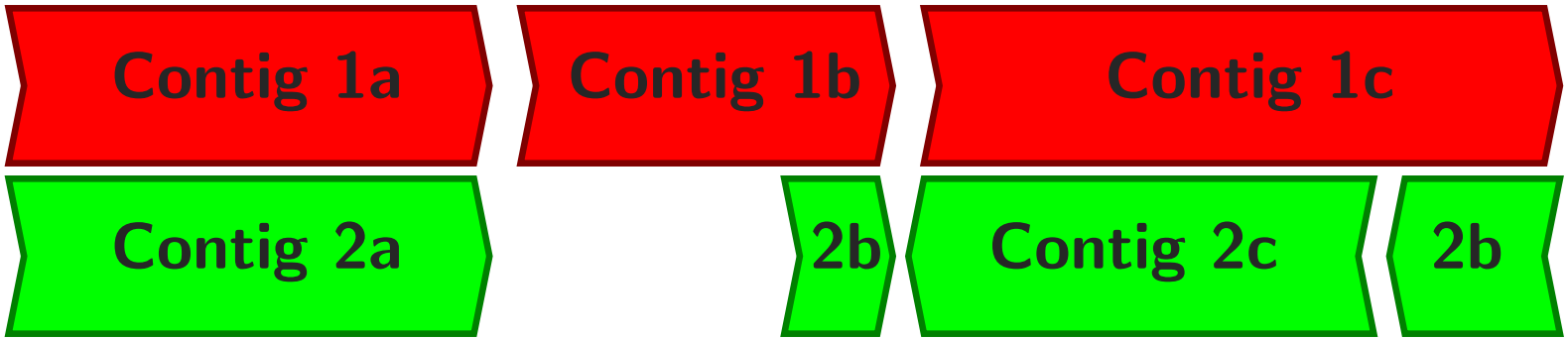
Assembly
What you won't get



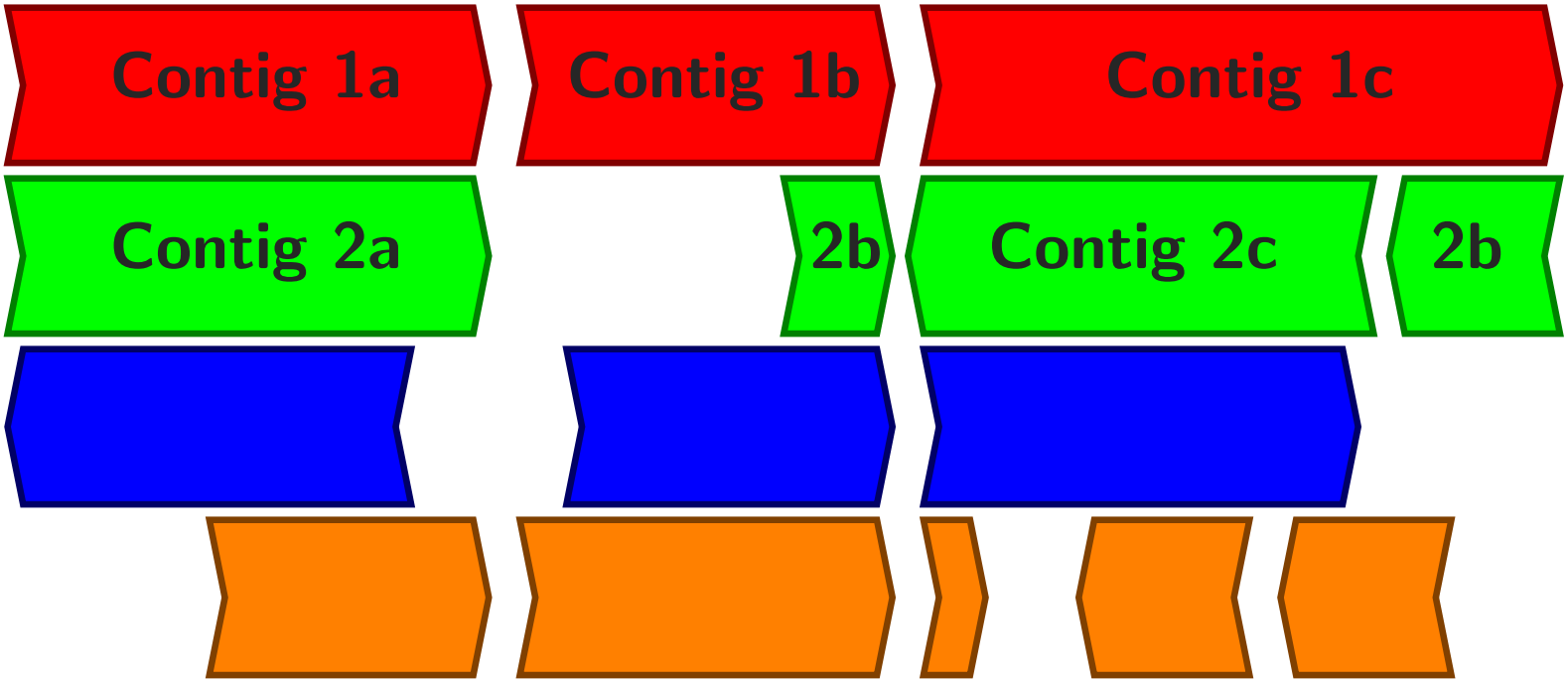
Contigs

Contigs

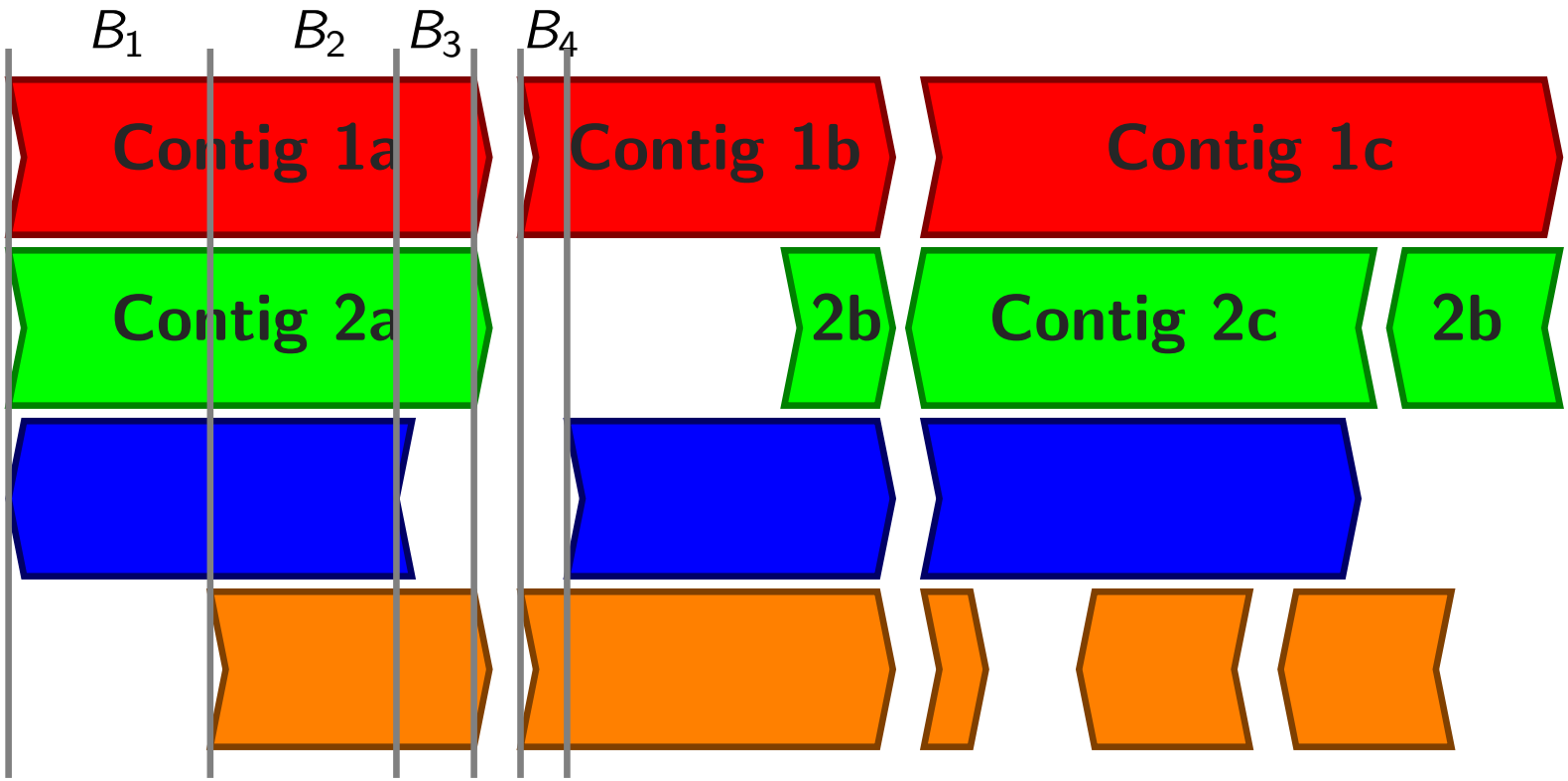
Reference Alignment



More Reference Alignment



Blockwise Reference Alignment



Blocking

Lessons learned from andi

Gaps are hard to compute. What happens if we leave them out?

Gapped Block

S1	A	C	C	G	T	T
S2	A	C	-	G	T	A
S3	A	T	C	C	T	T

$\Rightarrow$

Block 1

S1	A	C
S2	A	C
S3	A	T

Block 2

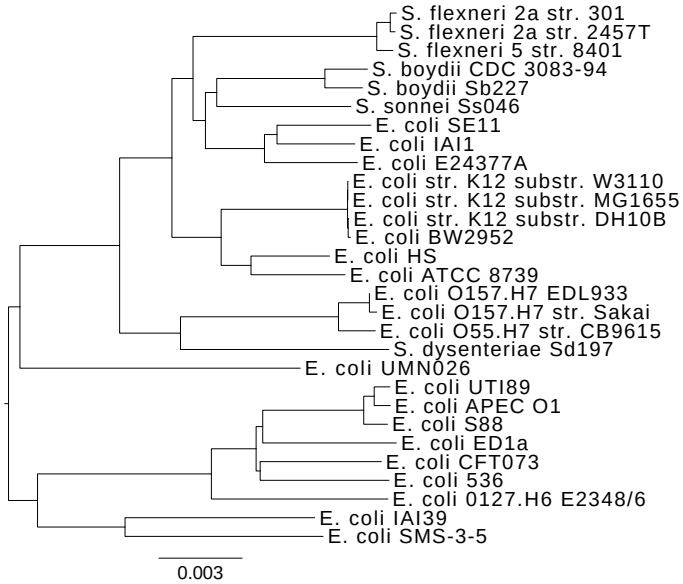
S1	C
S3	C

Block 3

S1	G	T	T
S2	G	T	A
S3	C	T	T

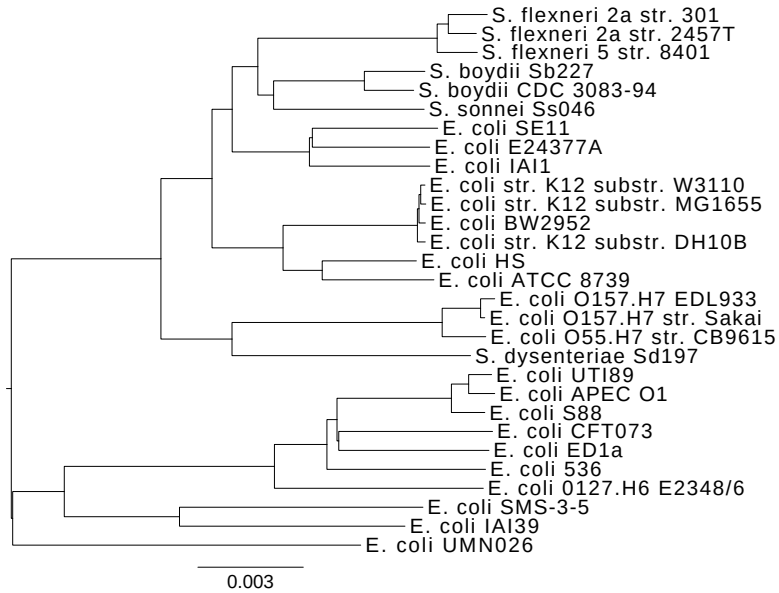
Eco29 again

Multiz & TBA



1d 3h

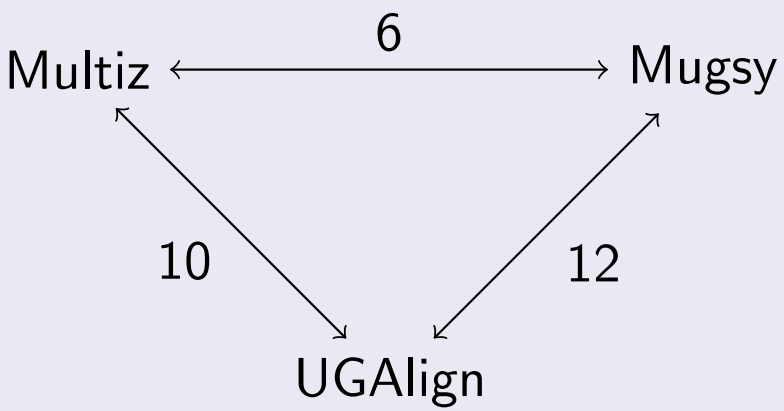
UGAlign



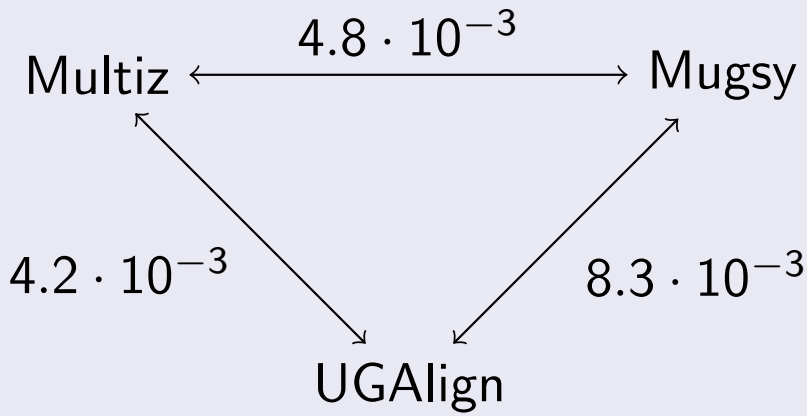
3.5 s

Accuracy on Eco29

Symmetric (Robinson-Foulds)



Branch Score



Pairwise coverage

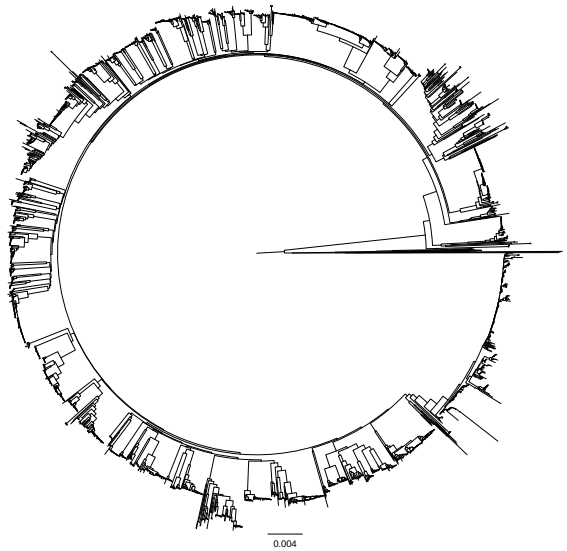
Multiz: 82% Mugsy: 81% UGAlign: 76%

New Frontiers

Pneu3085

- 3085 *Streptococcus pneumoniae* genomes
- multiple contigs per genome
- 2 million nucleotides per genome
- a total filesize of 6.8 GB

- andi: 4 h 59 min, 9.8 GB RAM
- UGAlign: 2 min, 13 GB RAM, alignment file: 35 GB



UGAlign

Pros

- super-duper fast
- reasonably accurate
- $O(n \cdot l \cdot \log(n \cdot l))$
- handles massive data

Cons

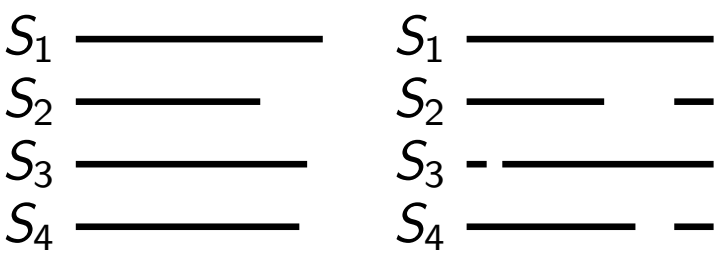
- ungapped
- tiny block sizes
- reference-based
- sensitive w.r.t reference-choice
- heuristic: longest

The next step

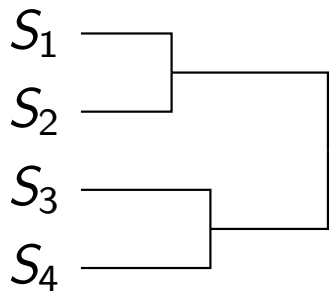
Stitch blocks back together by inserting gaps

A near future?

Genomes $\xrightarrow{\text{fast}}$ Alignment $\xrightarrow{\text{slow}}$ Distance Matrix $\xrightarrow{\text{fast}}$ Tree



	S_1	S_2	S_3	S_4
S_1	0			
S_2	0.1	0		
S_3	0.4	0.4	0	
S_4	0.4	0.4	0.2	0



Thank you for your attention

Bernhard Haubold

Julien Y. Dutheil

- Previous projects: `github.com/kloetzl`
- Available in a repository near you (soon)
- Random information: `kloetzl.info`