Fast Computation of Genome Distances

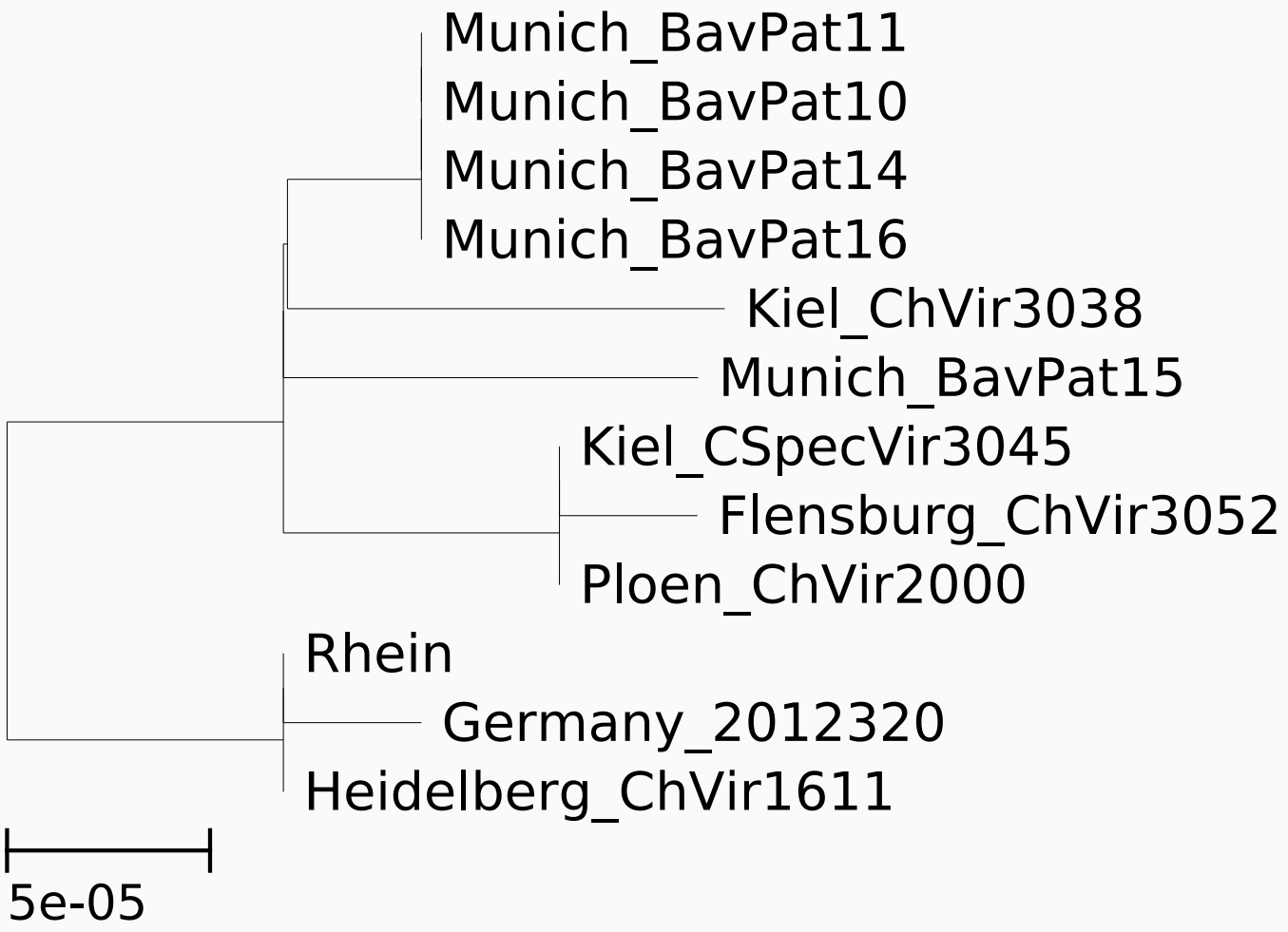
Fabian Klötzl & Bernhard Haubold

Aquavit 2020

MPI für Evolutionsbiologie

Slides: `kloetzl.info`

Sars-CoV-2


Munich_BavPat11
Munich_BavPat10
Munich_BavPat14
Munich_BavPat16
Kiel_ChVir3038
Munich_BavPat15
Kiel_CSpecVir3045
Flensburg_ChVir3052
Ploen_ChVir2000
Rhein
Germany_2012320
Heidelberg_ChVir1611
5e-05

Multiple Sequence Alignment

```
A   AACGTTGTGCA
B   AATG-TGAGC-
C   -ACGTTGTG--
D   AAC-TTGTGA-
```

L: Length of genome sequence
N: Number of sequences

Multiple Sequence Alignment: $N \cdot L$
Computation of optimal MSA: NP-Complete,
think: $O(2^{NL})$

Example

$L = 11\,\text{bp}, \quad N = 4$
$NL = 44, \quad 2^{NL} = 1.7 \cdot 10^{13}$

Multiple Sequence Alignment

```
A   AACGTTGTGCA
B   AATG-TGAGC-
C   -ACGTTGTG--
D   AAC-TTGTGA-
```

L: Length of genome sequence
N: Number of sequences

Multiple Sequence Alignment: $N \cdot L$
Computation of optimal MSA: NP-Complete,
think: $O(2^{NL})$

Example

$L = 11$ bp, $\quad N = 4$
$NL = 44, \quad 2^{NL} = 1.7 \cdot 10^{13}$

Sars-CoV-2

$L \approx 30\,000$ bp
$N = 12$
$NL = 3.6 \cdot 10^5$
$2^{NL} = 6 \cdot 10^{108370}$

Multiple Sequence Alignment

```
A   AACGTTGTGCA
B   AATG-TGAGC-
C   -ACGTTGTG--
D   AAC-TTGTGA-
```

L: Length of genome sequence
N: Number of sequences

Multiple Sequence Alignment: $N \cdot L$
Computation of optimal MSA: NP-Complete, think: $O(2^{NL})$

Example

$L = 11\,\text{bp}, \quad N = 4$
$NL = 44, \quad 2^{NL} = 1.7 \cdot 10^{13}$

Sars-CoV-2

$L \approx 30\,000\,\text{bp}$
$N = 12$
$NL = 3.6 \cdot 10^{5}$
$2^{NL} = 6 \cdot 10^{108370}$

Ensembl *E. coli*

$L \approx 5\,\text{Mb}$
$N = 2681$
$NL = 1.3 \cdot 10^{10}$
$2^{NL} \equiv \texttt{Inf}$

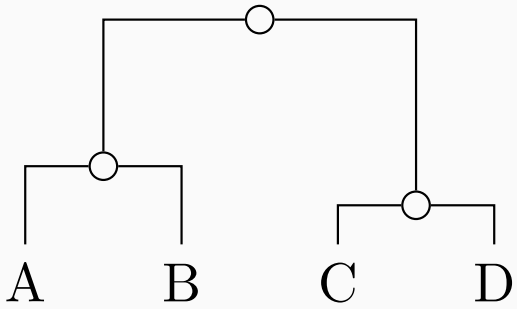
Faster Phylogeny Reconstruction

```
>A
AACGTTGTGCA
>B
AATGTGAGC
>C
ACGTTGTG
>D
AACTTGTGA
```

$\Rightarrow$

$$\begin{pmatrix} 0 & 0.1 & 0.25 & 0.3 \\ 0.1 & 0 & 0.3 & 0.3 \\ 0.25 & 0.3 & 0 & 0.05 \\ 0.3 & 0.3 & 0.05 & 0 \end{pmatrix}$$

$\Rightarrow$



$O(N \cdot L) \overset{?}{\Rightarrow} O(N^2) \overset{O(N^3)}{\Rightarrow} O(N)$

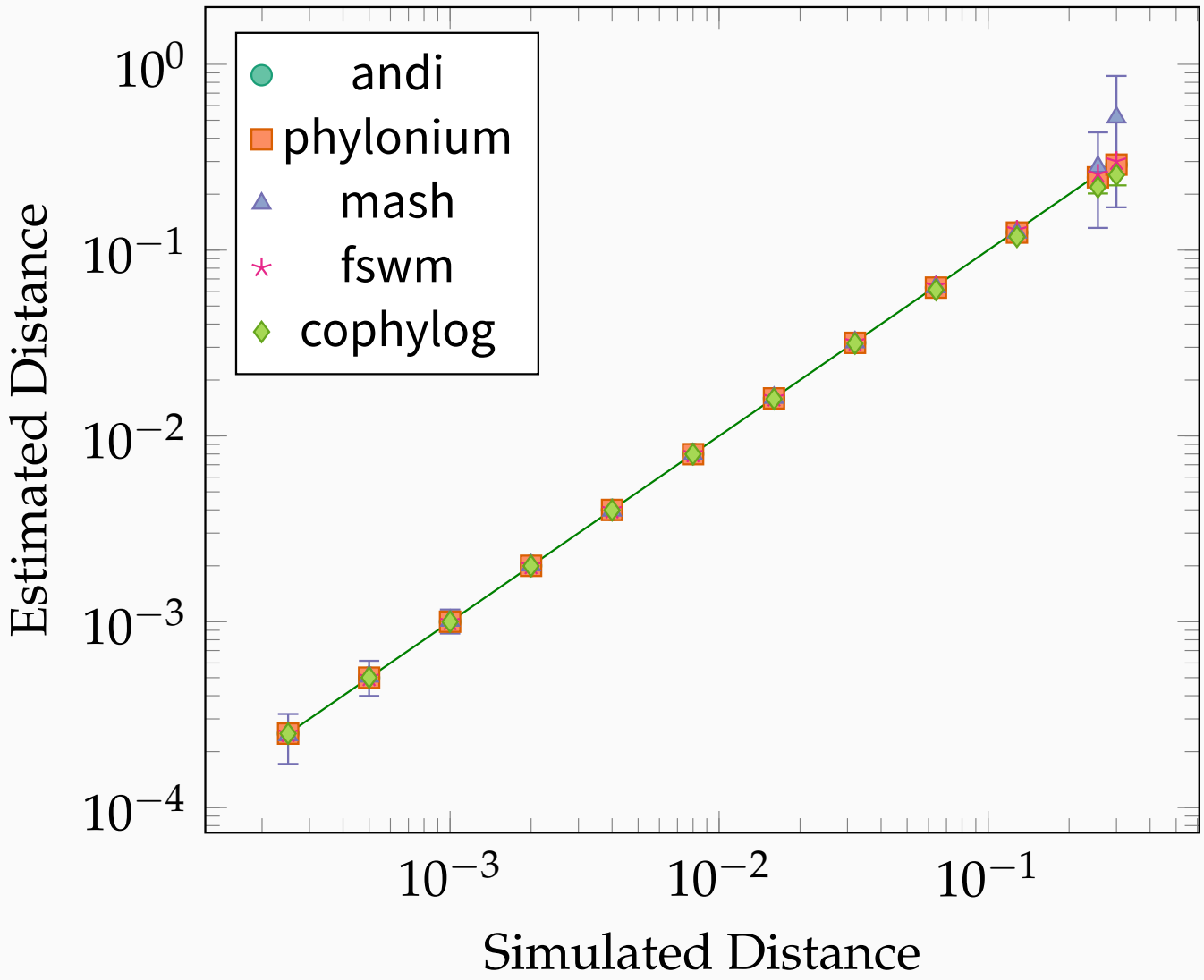
Pairwise Comparison

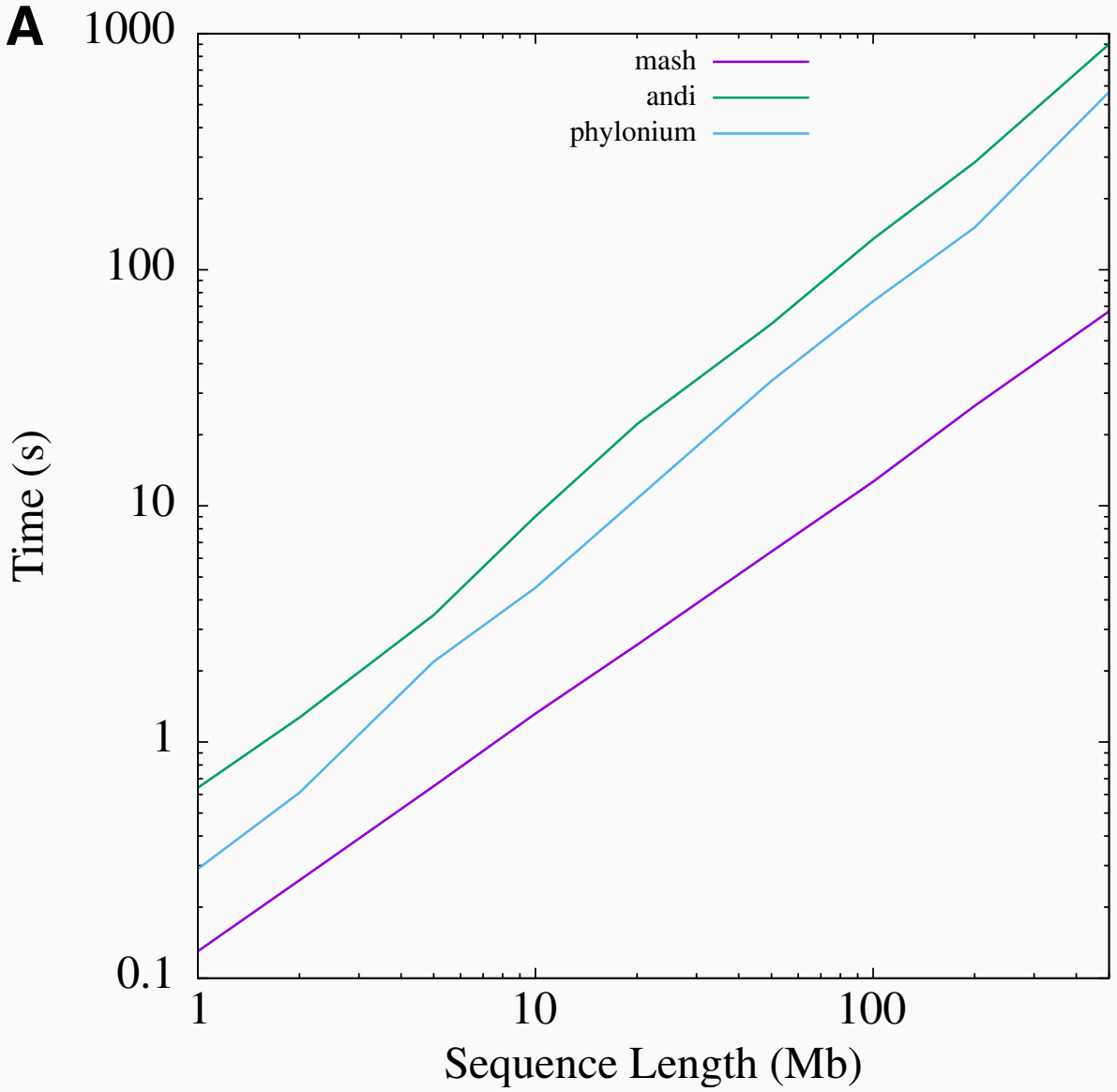
```
A       AACGTTGTGCA
B       AATG  TGAGC
```

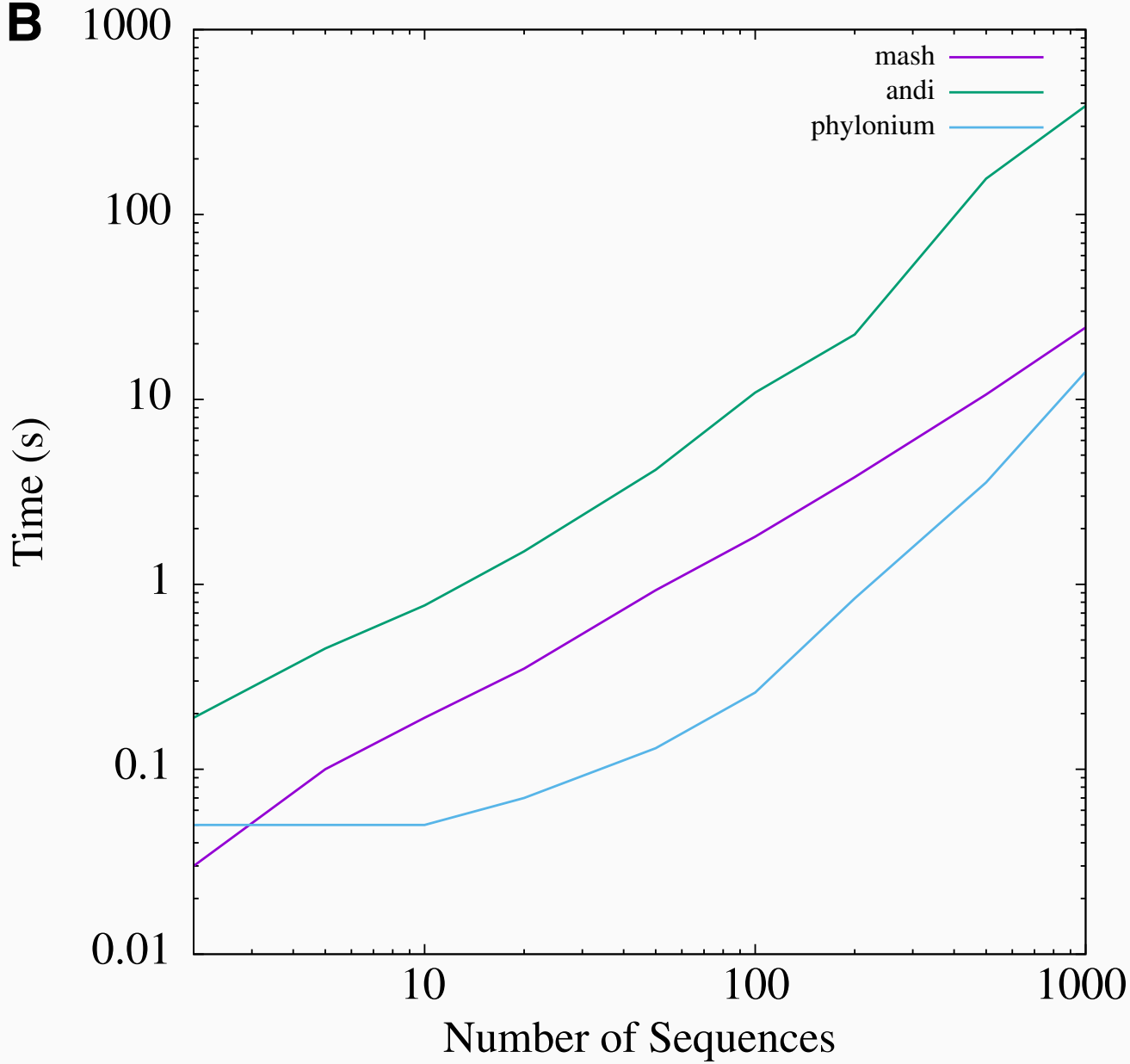
- Align long exact matches
- Disregard indels
- Don't focus on every substitution
- Count substitutions, quickly

$$O(L_1 + L_2) \text{ comparison}$$

Accuracy


andi
phylonium
mash
fswm
cophylog
Estimated Distance
Simulated Distance
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-3}
10^{-2}
10^{-1}


A
1000
100
10
1
0.1
Time (s)
1
10
100
Sequence Length (Mb)
mash
andi
phylonium


B
1000
100
10
1
0.1
0.01
Time (s)
10
100
1000
Number of Sequences
mash
andi
phylonium

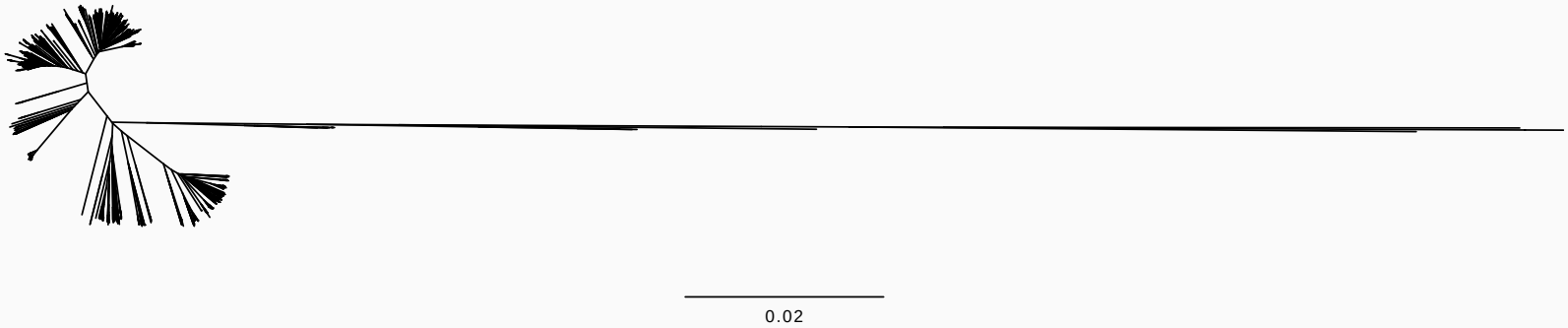
A simplified MSA

```
A      AACGTTGTGCA
B      AATG  TGAGC
C        ACGTTGTG
D      AAC  TTGTGA
```

- Align long exact matches against a single reference
- Disregard indels
- Don’t focus on every substitution
- Count substitutions, quickly

$$O(N^2 L) \text{ comparison}$$

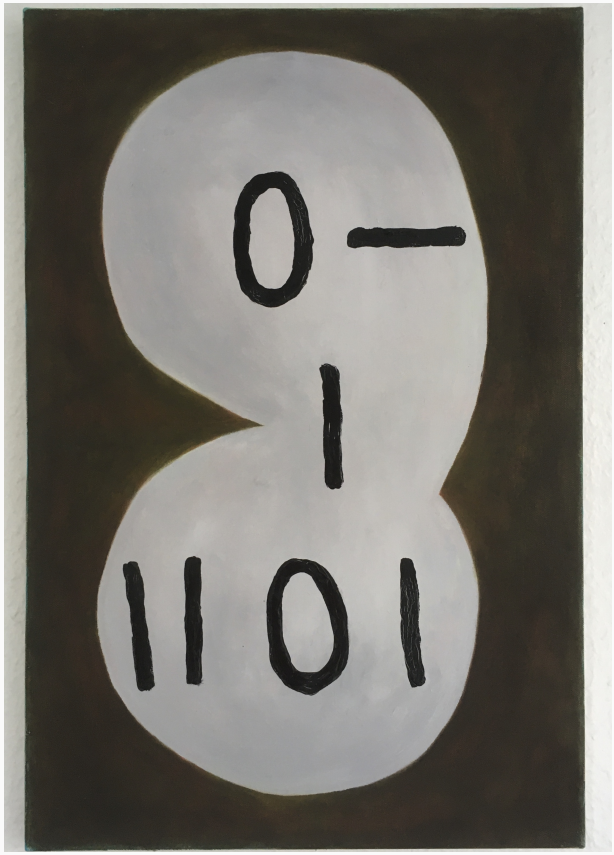
Ensembl *E. coli*



$L \approx 5$ Mb $\quad N = 2681$

13 min $\quad$ 14 GB

Phylonium



Art: Christian Rothmaler

Paper:

- doi.org/10.1093/bioinformatics/btz903

Get it:

- apt install phylonium
- brew install science/phylonium
- aura -A phylonium
- github.com/evolbioinf/phylonium