A new sketching method for metagenome assembly

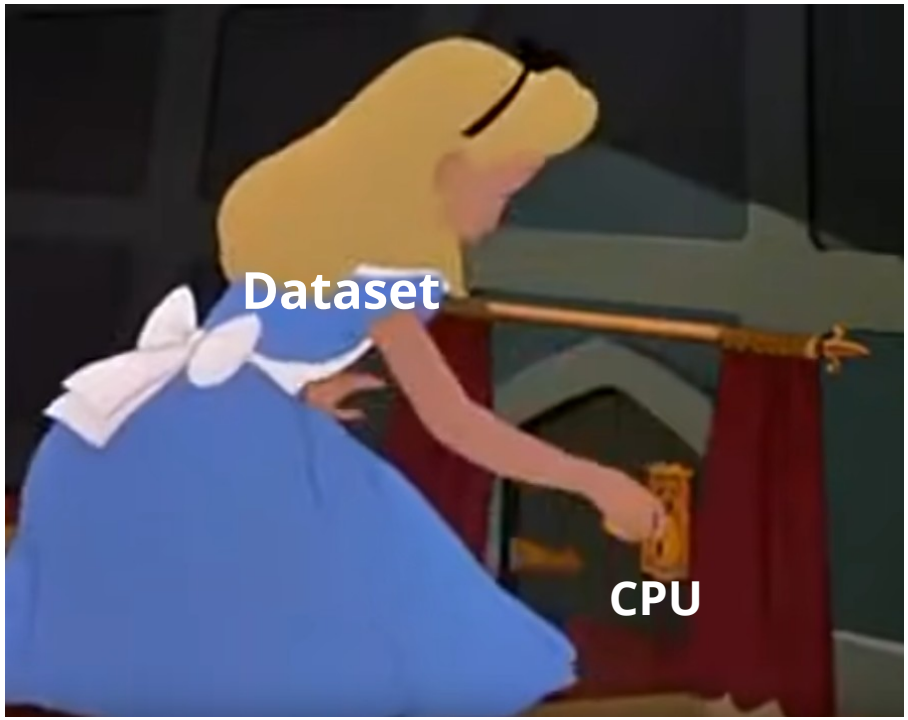
Roland Faure[1,2], Baptiste Hilaire[2], Jean-François Flot[1], Dominique Lavenier[2]

[1]Université libre de Bruxelles (ULB) - Belgium

[2]Université de Rennes, IRISA - France

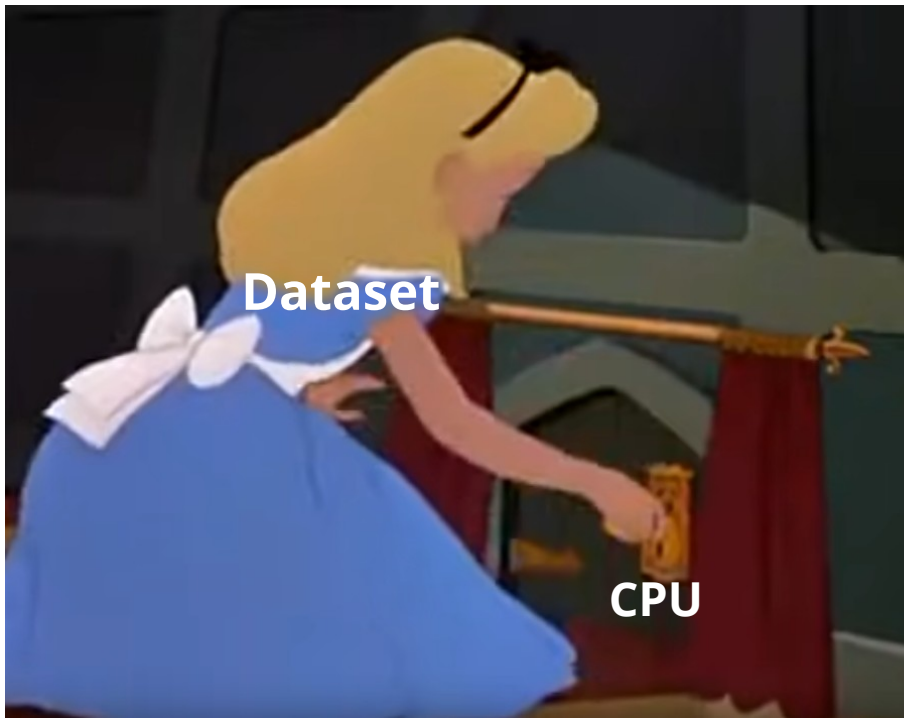
Genome Informatics 2024

Sequencing data is big



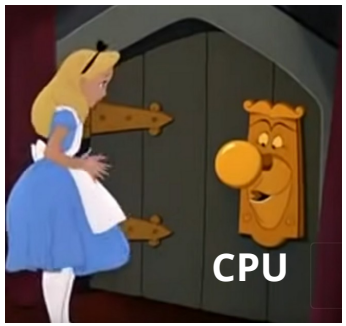
Credits: Alice in Wonderland, Lewis, Disney

Sequencing data is big

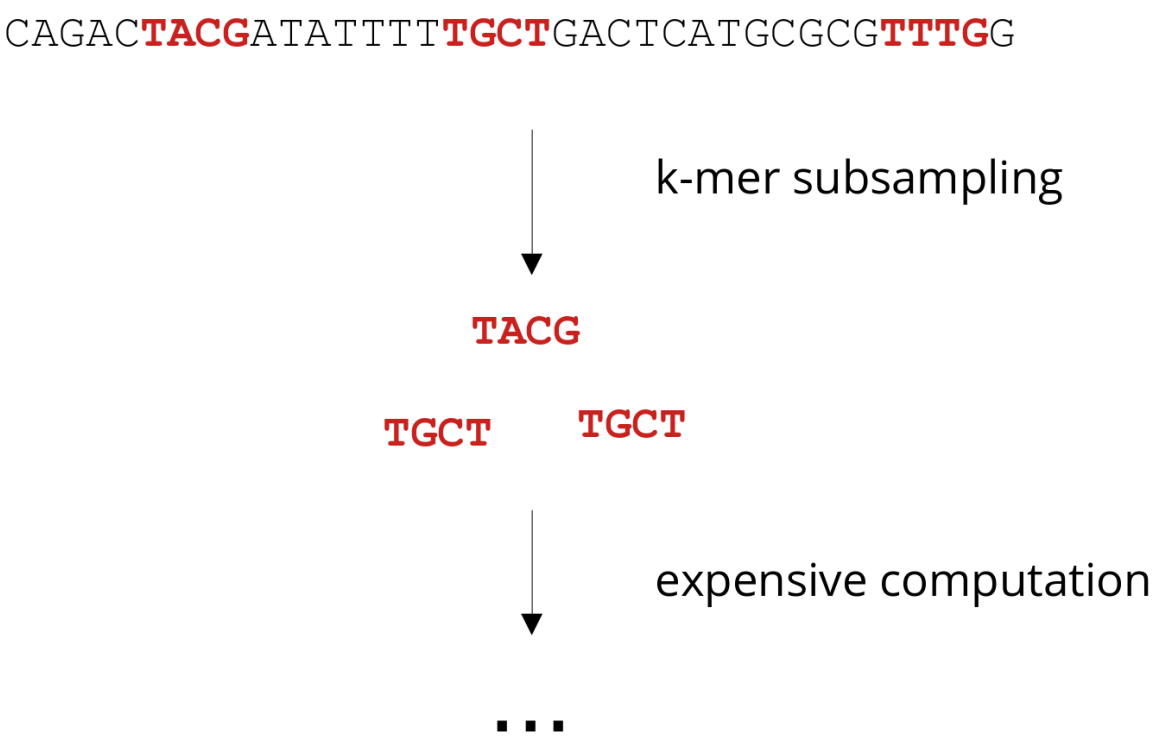


Credits: Alice in Wonderland, Lewis, Disney

Drink-me potion

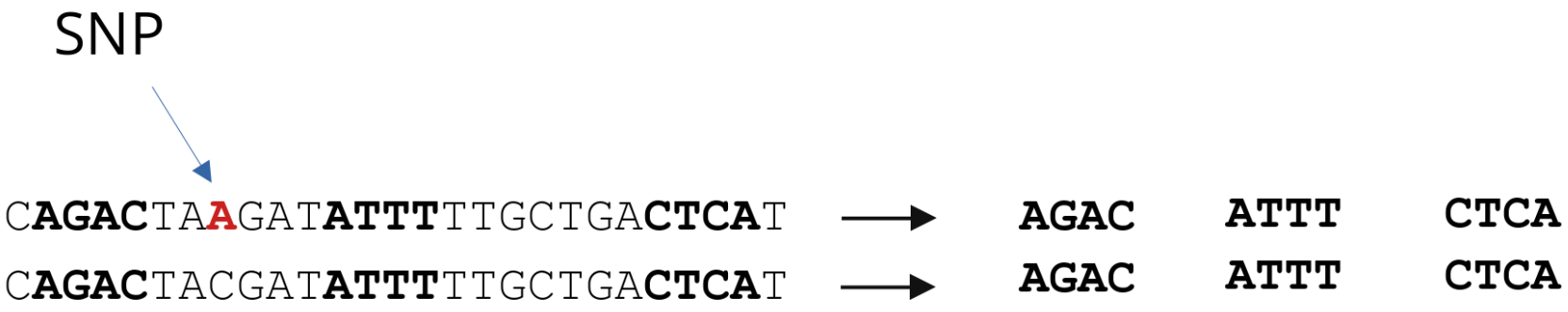


Sketching with sequence subsampling

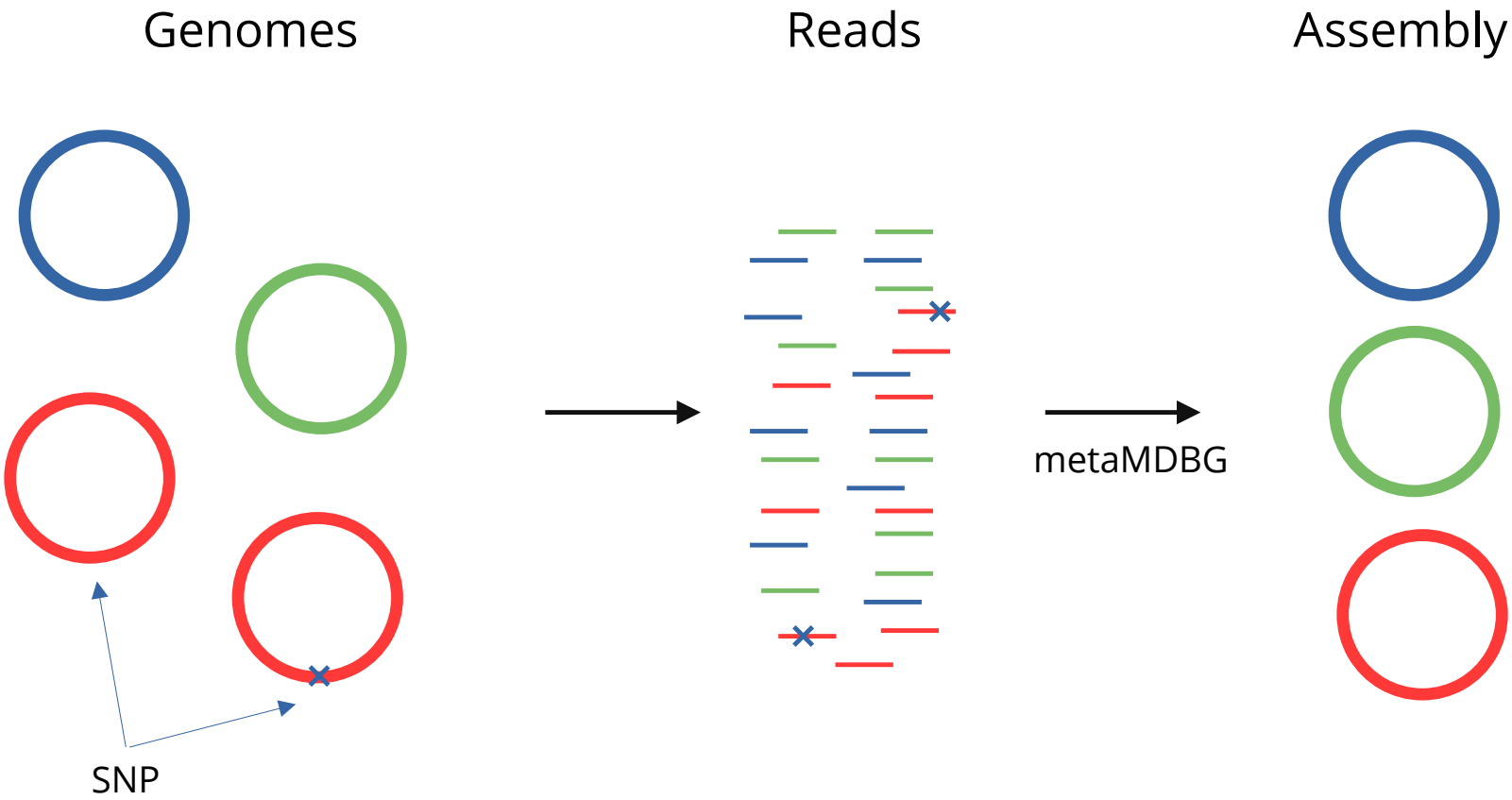


- minimizers, FracMinHash, seed-chain, syncmers...
- minimap2, Mash, BLAST, metaMDBG...

Sequence subsampling does not preserve SNPs

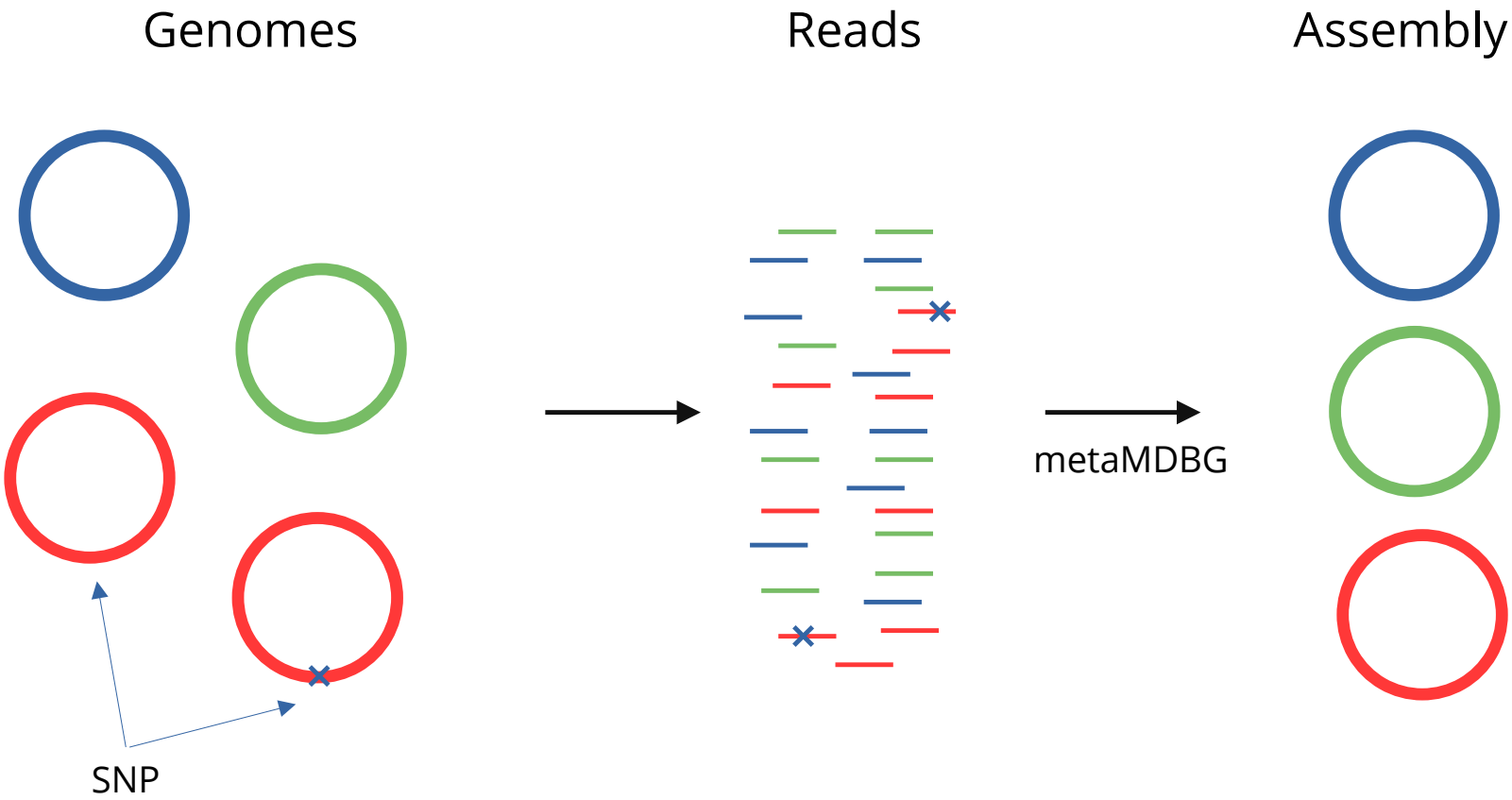


My problem: complete metagenome assembly



- metaMDBG is very fast, but some variants are lost!

My problem: complete metagenome assembly



- metaMDBG is very fast, but some variants are lost!
- Let me introduce MSR sketching

Introducing a MSR sketch

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

Introducing a MSR sketch

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence `CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG`

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

sequence **CAGTATGGAT**ACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**CAGTATGGAT**)= 0.0023

f(**CAGTATGGAT**)= A

sketch A

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(AGTATGGATA)= 0.624

f(AGTATGGATA)= $\emptyset$

sketch A

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

sequence CA**GTATGGATAC**AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GTATGGATAC**)= 0.124

f(**GTATGGATAC**)= G

sketch A G

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

sequence CAG**TATGGATACA**GATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TATGGATACA**)= 0.88

f(**TATGGATACA**)= $\emptyset$

sketch A G

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence CAGT**ATGGATACAG**ATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATGGATACAG**)= 0.32

f(**ATGGATACAG**)= $\emptyset$

sketch A G

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

sequence CAGTA**TGGATACAGA**TGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TGGATACAGA**)= 0.19

f(**TGGATACAGA**)= T

sketch A G T

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence CAGTAT**GGATACAGAT**GGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GGATACAGAT**)= 0.214

f(**GGATACAGAT**)= $\emptyset$

sketch A G T

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

sequence CAGTATG**GATACAGATG**GAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GATACAGATG**)= 0.678

f(**GATACAGATG**)= $\emptyset$

sketch A G T

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

sequence CAGTATGG**ATACAGATGG**AGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATACAGATGG**)= 0.669

f(**ATACAGATGG**)= $\emptyset$

sketch A G T

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

hash(**TGTACCAGAG**)= 0.06

f(**TGTACCAGAG**)= C

sketch A G T T C C G T C

Introducing a MSR sketch

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

order (l)

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

compression ratio (c)

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch A G T T C C G T C

MSRs keep SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	**CAGTATGGAT**ACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
sketch1	A
sequence2	**CAGTATGGAT**ACAGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG
sketch2	A

MSRs keep SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	C**AGTATGGATA**CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
sketch1	A
sequence2	C**AGTATGGATA**CAGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG
sketch2	A

MSRs keep SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	CA**GTATGGATAC**AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G
sequence2	CA**GTATGGATAC**AGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG
sketch2	A G

MSRs keep SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	CAGTATGGATACAG**ATGGAGATAT**CATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T
sequence2	CAGTATGGATACAG**ATGGAGATATG**ATCGAGTAGGGGCACTGTACCAGAG
sketch2	A G T

MSRs keep SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	CAGTATGGATACAGA**TGGAGATATC**ATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T T
sequence2	CAGTATGGATACAGA**TGGAGATATG**ATCGAGTAGGGGCACTGTACCAGAG
sketch2	A G T

MSRs keep SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	CAGTATGGATACAGAT**GGAGATATCA**TCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T T
sequence2	CAGTATGGATACAGAT**GGAGATATGA**TCGAGTAGGGGCACTGTACCAGAG
sketch2	A G T G

MSRs keep SNPs

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2$$

sequence1	CAGTATGGATACAGATG**GAGATATCAT**CGAGTAGGGGCACTGTACCAGAG
sketch1	A G T T C
sequence2	CAGTATGGATACAGATG**GAGATATGAT**CGAGTAGGGGCACTGTACCAGAG
sketch2	A G T G

MSRs keep SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	CAGTATGGATACAGATGG**AGATATCATC**GAGTAGGGGCACTGTACCAGAG
sketch1	A G T T C
sequence2	CAGTATGGATACAGATGG**AGATATGATC**GAGTAGGGGCACTGTACCAGAG
sketch2	A G T G

MSRs keep SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

sequence1	CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**
sketch1	A G T T C C G T C
sequence2	CAGTATGGATACAGATGGAGATAT**G**ATCGAGTAGGGGCAC**TGTACCAGAG**
sketch2	A G T G A C G T C

MSRs keep and amplify SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

- If $l \cdot c > 1$, SNPs are amplified

MSR = **Mapping-friendly** Sequence Reductions [1]

- Sketches are reduced sequences
- Mapping-friendly

[1] Blassel, Luc & Medvedev, Paul & Chikhi, Rayan. (2022). Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience.

MSR = **Mapping-friendly** Sequence Reductions [1]

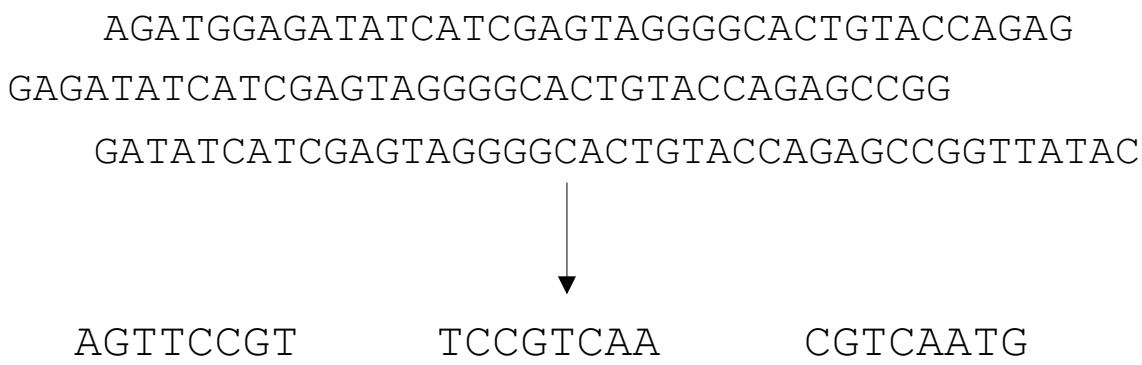
- Sketches are reduced sequences
- Mapping-friendly

- Reverse-complement property not shown

[1] Blassel, Luc & Medvedev, Paul & Chikhi, Rayan. (2022). Mapping-friendly sequence reductions: Going beyond homopolymer compression. iScience.

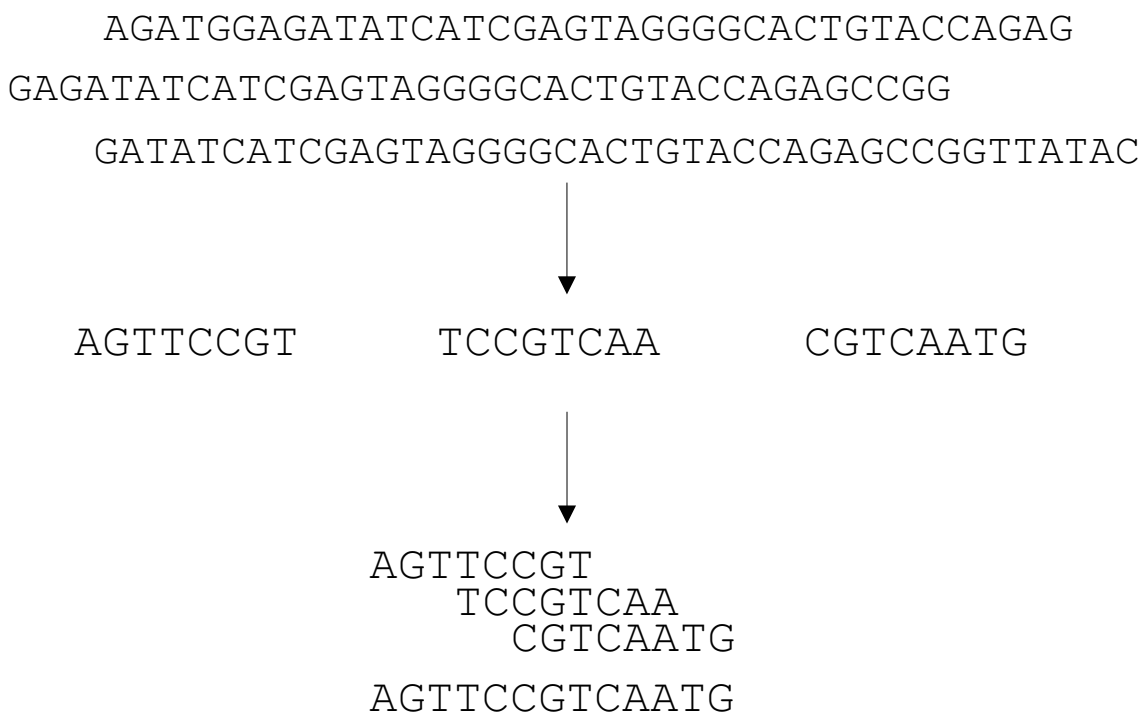
An MSR assembler

MSR sketching



An MSR assembler

MSR sketching



Assembly

An MSR assembler

AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
GAGATATCATCGAGTAGGGGCACTGTACCAGAGCCGG
GATATCATCGAGTAGGGGCACTGTACCAGAGCCGGTTATAC

MSR sketching

AGTTCCGT TCCGTCAA CGTCAATG

Assembly

AGTTCCGT
TCCGTCAA
CGTCAATG

AGTTCCGTCAATG

Inflating

AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAGCCGGTTATAC

Going back to uncompressed space

- MSR is lossy $\rightarrow$ not strictly reversible

Going back to uncompressed space

- MSR is lossy $\rightarrow$ not strictly reversible
- Keep a record while compressing

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch **AG_____A_____G_____C_____A__**

record { AGAG → ATGGATACAGATGGAGATATCATCG,
GAGC → AGTATGGATACAGATGGAGATATCATCGAGTAGGGGC,
AGCA → GATGGAGATATCATCGAGTAGGGGCACTGTAC }

Going back to uncompressed space

- MSR is lossy → not strictly reversible
- Keep a record while compressing

sequence `CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG`

sketch **`AG_____A_____G_____C_____A__`**

record
```
{AGAG → ATGGATACAGATGGAGATATCATCG,
 GAGC → AGTATGGATACAGATGGAGATATCATCGAGTAGGGGC,
 AGCA → GATGGAGATATCATCGAGTAGGGGCACTGTAC}
```

- Actually record 31-mers

The Alice assembler

1.
sketch the reads

2.
assemble the sketchs

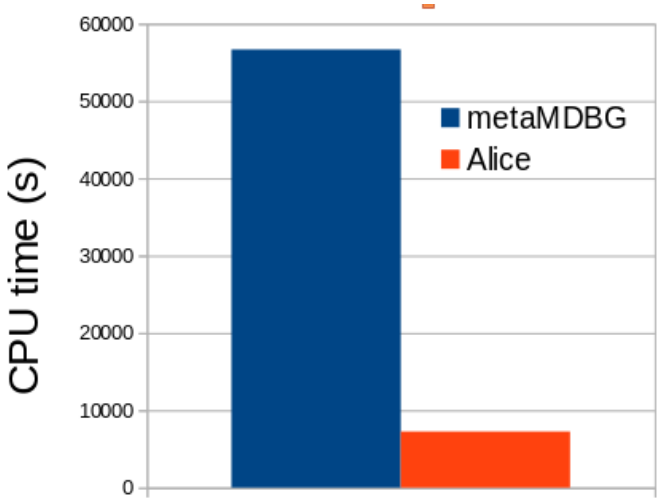
3.
inflate the assembly

Credits: Alice in Wonderland, Lewis, Disney

- Any assembler, by default BCALM2+tip-clipping
- `github.com/rolandfaure/alice-asm` (warning: immature code)

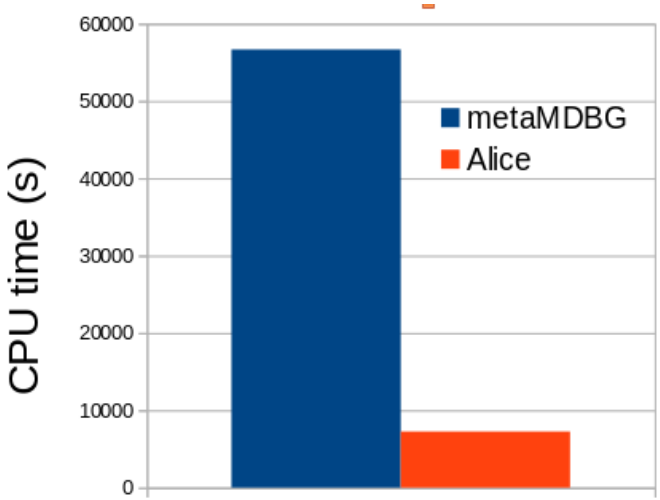
The Alice assembler: results

- Zymobiomics Gut Microbiome Standard with 5 strains of *E.coli*



The Alice assembler: results

- Zymobiomics Gut Microbiome Standard with 5 strains of *E.coli*



Genome fraction (%)

	metamdbg	alice
Escherichia_coli_B1109	78.408	92.039
Escherichia_coli_B3008	36.411	99.968
Escherichia_coli_B766	95.647	95.641
Escherichia_coli_JM109	38.211	96.334
Escherichia_coli_b2207	37.335	95.495

Measured using metaQUAST

- Strains are not collapsed

The dark side of MSR: errors

- Errors are amplified: Alice only works on highly accurate reads

MSR sketching: take-home messages

1.
sketch the reads

2.
assemble the sketchs

3.
inflate the assembly

Credits: Alice in Wonderland, Lewis, Disney

- MSR sketches are sequences

MSR sketching: take-home messages

1.
sketch the reads

2.
assemble the sketchs

3.
inflate the assembly

Credits: Alice in Wonderland, Lewis, Disney

- MSR sketches are sequences
- Differences between sequences are kept & amplified

MSR sketching: take-home messages

1.
sketch the reads

2.
assemble the sketchs

3.
inflate the assembly

Credits: Alice in Wonderland, Lewis, Disney

- MSR sketches are sequences
- Differences between sequences are kept & amplified
- Alice HiFi assembler out soon (`github.com/rolandfaure/alice-asm`)