

Fast and strain-aware HiFi metagenome assembly: MSR sketching and the Alice assembler

Roland Faure^{1,2,3}, Jean-François Flot¹, Dominique Lavenier²

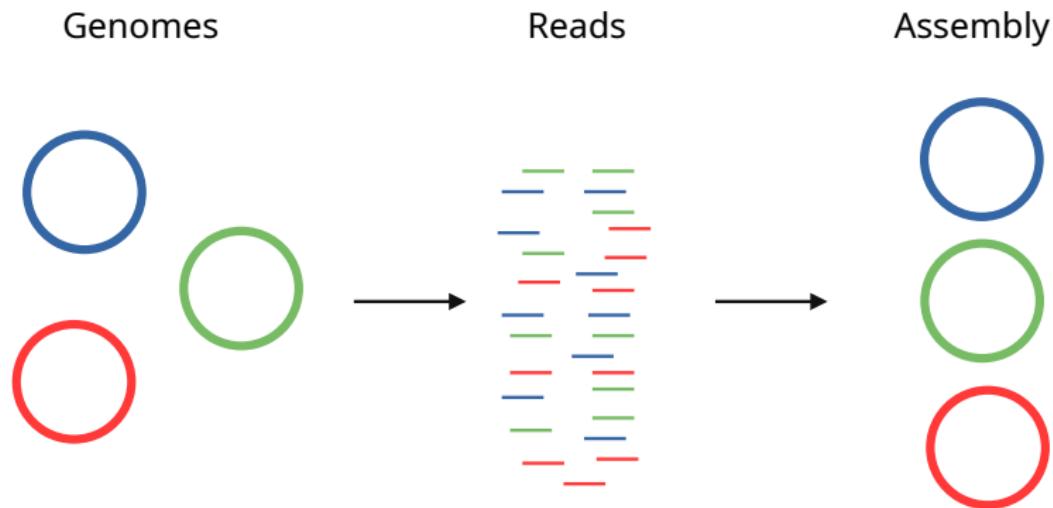
¹Université libre de Bruxelles (ULB) - Belgium

²Université de Rennes, IRISA - France

³Institut Pasteur, Paris - France

ISMB/ECCB 2025

Metagenome assembly



(Meta)genome assembly is a big computation

(Meta)genome assembly is a big computation

- ▶ Assembling a human gut metagenome (HiFi, 250Gpb)

(Meta)genome assembly is a big computation

- ▶ Assembling a human gut metagenome (HiFi, 250Gpb)

metaFlye

4 days, 256GB RAM



(Meta)genome assembly is a big computation

- ▶ Assembling a human gut metagenome (HiFi, 250Gpb)

metaFlye

4 days, 256GB RAM



hifiasm_meta

11 days, 454GB RAM



(Meta)genome assembly is a big computation

- ▶ Assembling a human gut metagenome (HiFi, 250Gpb)

metaFlye

4 days, 256GB RAM



hifiasm_meta

11 days, 454GB RAM



metaMDBG

19h, 10G RAM



metaMDBG: the trick is sketching input reads

CAGAC**TACG**ATATTT**TGCT**GACTCATGCGCG**TTTG**G



k-mer subsampling

TACG

TGCT **TGCT**

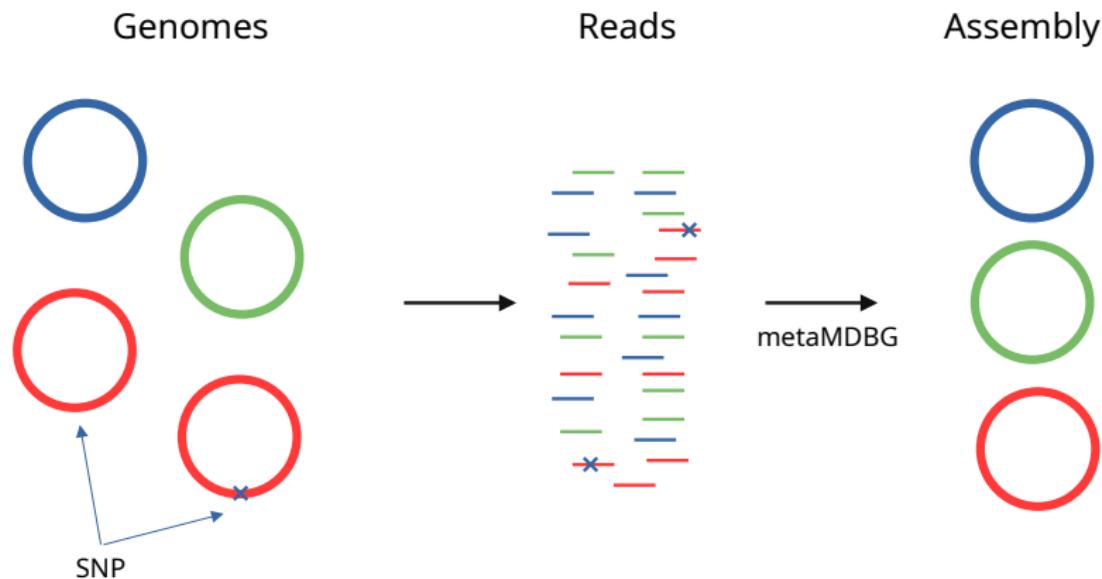


expensive computation

...

- ▶ minimizers, FracMinHash, seed-chain, strobemers...
- ▶ minimap2, Mash, BLAST, **metaMDBG**...

metaMDBG loses strain diversity



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

k-mer sketching loses SNPs

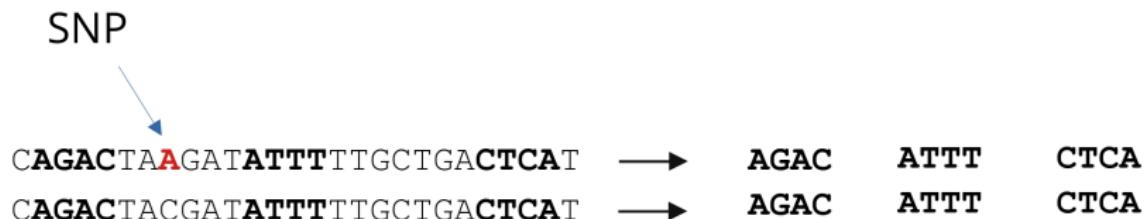
SNP



CAGACTA A GATATTTTGCTGACTCAT	→	AGAC	ATTT	CTCA
CAGACTACGAT ATTTTGCTGACTCAT	→	AGAC	ATTT	CTCA

k-mer sketching loses SNPs

SNP



CAGACTA**A**GATATTTTGCTGACTCAT → AGAC ATTT CTCA
CAGACTACGAT**T**TTTGCTGACTCAT → AGAC ATTT CTCA

- ▶ Is k-mer subsampling really the only way to sketch sequences ?

k-mer sketching loses SNPs

SNP

CAGACTA**A**GATATTTTGCTGACTCAT → AGAC ATTT CTCA
CAGACTACGAT**A**TTTGCTGACTCAT → AGAC ATTT CTCA

- ▶ Is k-mer subsampling really the only way to sketch sequences ?
- ▶ Bassel, Luc & Medvedev, Paul & Chikhi, Rayan. (2022). *Mapping-friendly sequence reductions: Going beyond homopolymer compression.* iScience.

Mapping-friendly Sequence Reductions

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

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

Mapping-friendly Sequence Reductions

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

$f(10-mer) \rightarrow A \text{ if } \text{hash}(10-mer) \in [0, 0.05]$

$f(10-mer) \rightarrow C \text{ if } \text{hash}(10-mer) \in [0.05, 0.1]$

$f(10-mer) \rightarrow G \text{ if } \text{hash}(10-mer) \in [0.1, 0.15]$

$f(10-mer) \rightarrow T \text{ if } \text{hash}(10-mer) \in [0.15, 0.2]$

$f(10-mer) \rightarrow \emptyset \text{ if } \text{hash}(10-mer) > 0.2$

sequence

CAGTATGGATAACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Mapping-friendly Sequence Reductions

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

$f(10-mer) \rightarrow A \text{ if } \text{hash}(10-mer) \in [0, 0.05]$

$f(10-mer) \rightarrow C \text{ if } \text{hash}(10-mer) \in [0.05, 0.1]$

$f(10-mer) \rightarrow G \text{ if } \text{hash}(10-mer) \in [0.1, 0.15]$

$f(10-mer) \rightarrow T \text{ if } \text{hash}(10-mer) \in [0.15, 0.2]$

$f(10-mer) \rightarrow \emptyset \text{ if } \text{hash}(10-mer) > 0.2$

sequence

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**CAGTATGGAT**) = 0.0023

$f(\text{CAGTATGGAT}) = A$

sketch

A

Mapping-friendly Sequence Reductions

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

$f(10-mer) \rightarrow A \text{ if } \text{hash}(10-mer) \in [0, 0.05]$

$f(10-mer) \rightarrow C \text{ if } \text{hash}(10-mer) \in [0.05, 0.1]$

$f(10-mer) \rightarrow G \text{ if } \text{hash}(10-mer) \in [0.1, 0.15]$

$f(10-mer) \rightarrow T \text{ if } \text{hash}(10-mer) \in [0.15, 0.2]$

$f(10-mer) \rightarrow \emptyset \text{ if } \text{hash}(10-mer) > 0.2$

sequence

C~~AGTATGGATA~~CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(~~AGTATGGATA~~) = 0.624

$f(\text{~~AGTATGGATA~~}) = \emptyset$

sketch

A

Mapping-friendly Sequence Reductions

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

$f(10-mer) \rightarrow A \text{ if } \text{hash}(10-mer) \in [0, 0.05]$

$f(10-mer) \rightarrow C \text{ if } \text{hash}(10-mer) \in [0.05, 0.1]$

$f(10-mer) \rightarrow G \text{ if } \text{hash}(10-mer) \in [0.1, 0.15]$

$f(10-mer) \rightarrow T \text{ if } \text{hash}(10-mer) \in [0.15, 0.2]$

$f(10-mer) \rightarrow \emptyset \text{ if } \text{hash}(10-mer) > 0.2$

sequence

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GTATGGATAC**) = 0.124

$f(\textbf{GTATGGATAC}) = G$

sketch

A G

Mapping-friendly Sequence Reductions

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

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

sequence CAG**TATGGATACA**GATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

$$\text{hash}(\textbf{TATGGATACA}) = 0.88$$
$$f(\textbf{TATGGATACA}) = \emptyset$$

sketch A G

Mapping-friendly Sequence Reductions

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

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

sequence CAGT**ATGGATACAG**ATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

$$\begin{aligned}\text{hash}(\textbf{ATGGATACAG}) &= 0.32 \\ f(\textbf{ATGGATACAG}) &= \emptyset\end{aligned}$$

sketch A G

Mapping-friendly Sequence Reductions

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

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

sequence CAGTA**TGGATACAGA**TGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

$$\begin{aligned}\text{hash}(\textcolor{red}{TGGATACAGA}) &= 0.19 \\ f(\textcolor{red}{TGGATACAGA}) &= T\end{aligned}$$

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTAT**GGATACAGAT**GGAGATATCATCGAGTAGGGCACTGTACCAGAG

$$\begin{aligned}\text{hash}(\textcolor{red}{GGATACAGAT}) &= 0.214 \\ f(\textcolor{red}{GGATACAGAT}) &= \emptyset\end{aligned}$$

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATG**GATACAGATG**GAGATATCATCGAGTAGGGCACTGTACCAGAG

$$\text{hash}(\textcolor{red}{GATACAGATG}) = 0.678$$
$$f(\textcolor{red}{GATACAGATG}) = \emptyset$$

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATGG**ATACAGATGG**AGATATCATCGAGTAGGGCACTGTACCAGAG

$$\begin{aligned}\text{hash}(\textbf{ATACAGATGG}) &= 0.669 \\ f(\textbf{ATACAGATGG}) &= \emptyset\end{aligned}$$

sketch A G T

Mapping-friendly Sequence Reductions

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

$f(10-mer) \rightarrow A \text{ if } \text{hash}(10-mer) \in [0, 0.05]$

$f(10-mer) \rightarrow C \text{ if } \text{hash}(10-mer) \in [0.05, 0.1]$

$f(10-mer) \rightarrow G \text{ if } \text{hash}(10-mer) \in [0.1, 0.15]$

$f(10-mer) \rightarrow T \text{ if } \text{hash}(10-mer) \in [0.15, 0.2]$

$f(10-mer) \rightarrow \emptyset \text{ if } \text{hash}(10-mer) > 0.2$

sequence

CAGTATGGATAACAGATGGAGATATCATCGAGTAGGGCAC**TGTACCAGAG**

$$\text{hash}(\textbf{TGTACCAGAG}) = 0.06$$

$$f(\textbf{TGTACCAGAG}) = C$$

sketch

A G T

T C

C G T C

Mapping-friendly Sequence Reductions

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

order (l) $\xrightarrow{\hspace{1cm}}$

$f(10-mer) \rightarrow A \text{ if } \text{hash}(10-mer) \in [0, 0.05]$

$f(10-mer) \rightarrow C \text{ if } \text{hash}(10-mer) \in [0.05, 0.1]$

$f(10-mer) \rightarrow G \text{ if } \text{hash}(10-mer) \in [0.1, 0.15]$

$f(10-mer) \rightarrow T \text{ if } \text{hash}(10-mer) \in [0.15, 0.2]$

$f(10-mer) \rightarrow \emptyset \text{ if } \text{hash}(10-mer) > 0.2$ $\xrightarrow{\hspace{1cm}}$ compression ratio (c)

sequence

CAGTATGGATAACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch

A G T T C C G T C

MSRs=Mapping-friendly Sequence Reductions

- ▶ MSR reductions are **mapping-friendly**

Diagram illustrating the mapping-friendliness of MSR reductions. Two DNA sequences are shown:

Top sequence: ATCATCGAGTAGGGGCACTGTACCATCAGAGCGCTTTAATGTAC

Bottom sequence: CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCATCAGAG

A red box highlights a 10-base pair segment from positions 5 to 14.

Below the top sequence, labels A G T are aligned under the first three bases, and C G T C are aligned under the last four bases.

Below the bottom sequence, labels A G T C are aligned under the first four bases.

- ▶ original sequences align \iff reduced sequences align

MSRs=Mapping-friendly Sequence Reductions

- ▶ MSR reductions are mapping-friendly

	C	G	T	C	CC	A
	ATCATCGAGTAGGGGCACTGTACCAGAGCGCTTAATGTAC					
CAGTATGGATA	CAGATGGAGA	TATC	ATCGAGTAGGGC	ACTGTAC	CCAGAG	
A G	T	T C	C	G	T	C

- ▶ original sequences align \iff reduced sequences align
 - ▶ This property is very useful

Assembling using MSR sketches

Read 1
Read 2
Read 3

ATGCATCGAGTAGGGGCACTGTACC
GAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
CAGATGGAGAATGCATCGAGTAGG

↓
sketch the reads

Read 1
Read 2
Read 3

TTGGCC
GGCCGGGGT
GTGATTGG

↓
assemble the sketches

GTGATTGGCCGGGGT
Read 3 ——————
Read 1 ——————
Read 2 ——————

↓
deduce final assembly

CAGATGGAGAATGCATCGAGTAGGGGCACTGTACCAGAGCCAGTAGCAT



MSRs keep SNPs

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

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

sequence1

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1

A

sequence2

CAGTATGGATACAGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG

sketch2

A

MSRs keep SNPs

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

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

sequence1

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1

A

sequence2

CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2

A

MSRs keep SNPs

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

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

sequence1	CA <u>GTATGGATA</u> CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G
sequence2	CA <u>GTATGGATA</u> CAGATGGAGATAT <u>G</u> ATCGAGTAGGGGCACTGTACCAGAG
sketch2	A G

MSRs keep SNPs

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

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

sequence1	CAGTATGGATAACAG	ATGGAGATAT	CATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T		
sequence2	CAGTATGGATAACAG	ATGGAGATATG	CATCGAGTAGGGGCACTGTACCAGAG
sketch2	A G T		

MSRs keep SNPs

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

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

sequence1	CAGTATGGATAACAGA	TGGAGATATC	ATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T		T
sequence2	CAGTATGGATAACAGA	TGGAGATATG	ATCGAGTAGGGGCACTGTACCAGAG
sketch2	A G T		

MSRs keep SNPs

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

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

sequence1	CAGTATGGATAACAGAT	GGAGATATCA	TCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T		T
sequence2	CAGTATGGATAACAGAT	GGAGATATGA	TCGAGTAGGGGCACTGTACCAGAG
sketch2	A G T		G

MSRs keep SNPs

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

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

sequence1	CAGTATGGATAACAGATG	GAGATATCAT	CGAGTAGGGCACTGTACCAGAG
sketch1	A G T		T C
sequence2	CAGTATGGATAACAGATG	GAGATATGAT	CGAGTAGGGCACTGTACCAGAG
sketch2	A G T		G

MSRs keep SNPs

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

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

sequence1	CAGTATGGATAACAGATGG	AGATATC ATC	GAGTAGGGCACTGTACCAGAG
sketch1	A G T	T C	
sequence2	CAGTATGGATAACAGATGG	AGATATG ATC	GAGTAGGGCACTGTACCAGAG
sketch2	A G T	G	

MSRs keep SNPs

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

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

sequence1

CAGTATGGATAACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

sketch1

A G T T C C G T C

sequence2

CAGTATGGATAACAGATGGAGATAT**G**ATCGAGTAGGGGCAC**TGTACCAGAG**

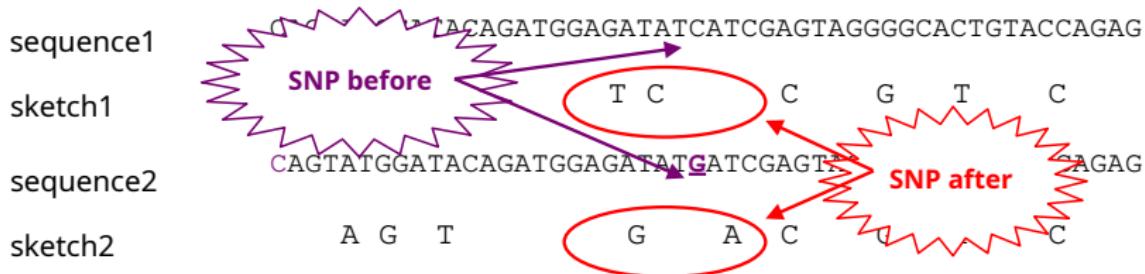
sketch2

A G T G A C G T C

MSRs keep and amplify SNPs

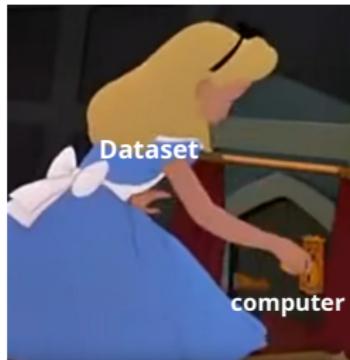
$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

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



The Alice assembler: assembling with MSR

Input dataset is too big **1. sketch reads** **2. assemble sketches** **3. inflate assembly** **Result**



Credits: Alice in Wonderland, Lewis, Disney

- ▶ Any assembler for step 2., by default BCALM2+tip-clipping
- ▶ github.com/rolandfaure/alice-asm



The Alice assembler: results

The Alice assembler: results

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

	Genome fraction (%)	
	metamdbq	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 Alice assembler: results

- ▶ Assembling a human gut metagenome (HiFi sequencing)

Fly
4d, 256G RAM



hifiasm_meta
11d, 454G RAM



metaMDBG
19h, 10G RAM



The Alice assembler: results

- ▶ Assembling a human gut metagenome (HiFi sequencing)

Flye
4d, 256G RAM



hifiasm_meta
11d, 454G RAM



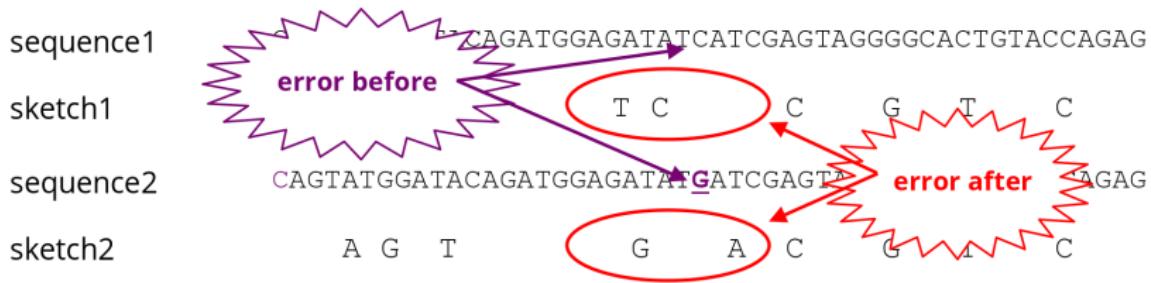
metaMDBG
19h, 10G RAM



Alice
5h, 10G RAM



The dark side of MSR: errors



- ▶ Errors are amplified: Alice only works on highly accurate reads
- ▶ New error rate \approx Original error rate / compression rate c

MSR sketching: take-home messages



Metagenomic datasets in the future

MSR sketching: take-home messages



- ▶ Alice HiFi assembler (github.com/rolandfaure/alice-asm) is **fast** and **strain-aware**

Metagenomic datasets in the future

MSR sketching: take-home messages



- ▶ Alice HiFi assembler (github.com/rolandfaure/alice-asm) is **fast** and **strain-aware**
- ▶ MSR sketches **keep & amplify** differences between sequences

Metagenomic datasets in the future

MSR sketching: take-home messages



Metagenomic datasets in the future

- ▶ Alice HiFi assembler (github.com/rolandfaure/alice-asm) is **fast** and **strain-aware**
- ▶ MSR sketches **keep & amplify** differences between sequences
- ▶ MSR sketches are sequences and can be manipulated as such

Perspectives

- ▶ Other uses for MSR sketching

Perspectives

- ▶ Other uses for MSR sketching

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

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

sequence

CAGTATGGATAACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

$$\text{hash}(\textcolor{red}{TGTACCAGAG}) = 0.06$$
$$f(\textcolor{red}{TGTACCAGAG}) = C$$

sketch

A G T T C C G T C

- ▶ Changing f , l , c