A new way to handle large data: MSR sketching

Roland Faure[1,2,3]

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

[2]Université de Rennes, IRISA - France

[3]Insitut Pasteur, Paris - France

Penn State, September 2025

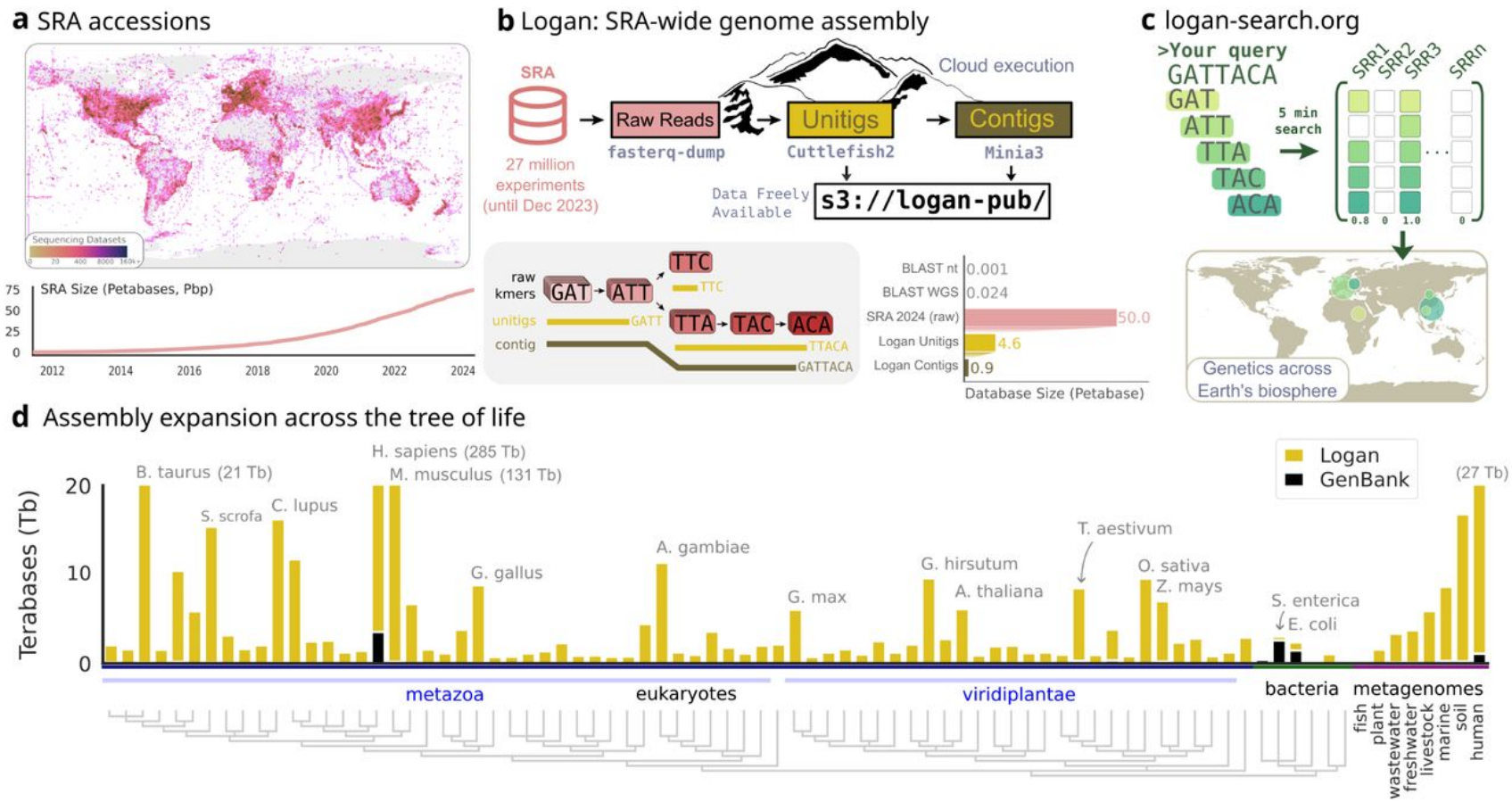
About me: postdoc, since Feb. 2025

Institut Pasteur, Paris

Rayan Chikhi
Institut Pasteur, Paris
Focus: massive genomics

My postdoc : Index & Search the Logan database

The Logan project

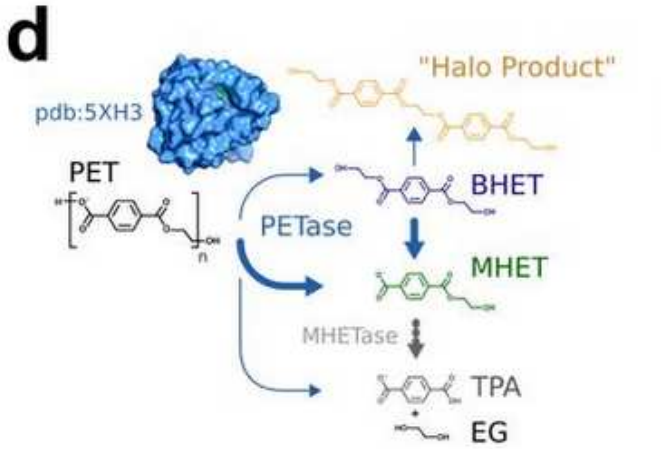


Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity, *bioRxiv*, 2025

The Logan project: exciting example

PAST SUCCESS - STARTUP

Plastivores: Plastic-Degrading Super-Microbes and Enzymes

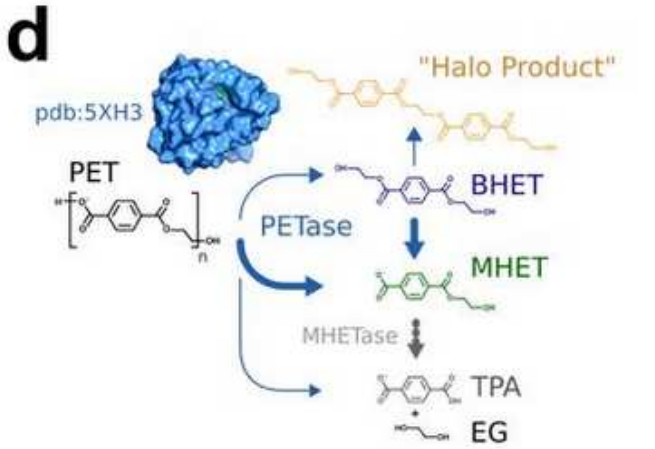


- 213 known PETases

The Logan project: exciting example

PAST SUCCESS - STARTUP

Plastivores: Plastic-Degrading Super-Microbes and Enzymes

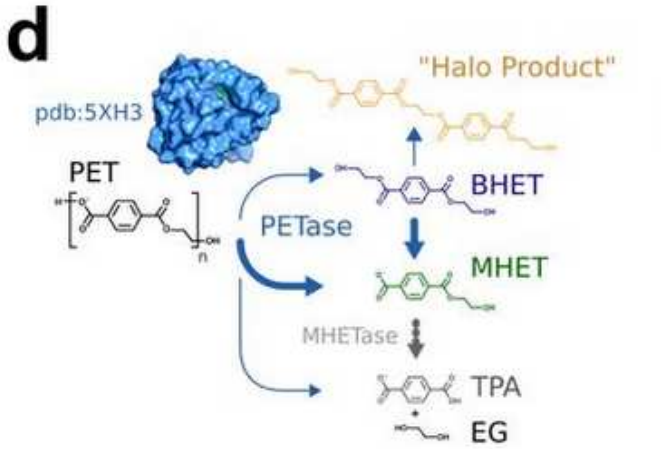


- 213 known PETases
- Let's look in Logan for homologs

The Logan project: exciting example

PAST SUCCESS - STARTUP

Plastivores: Plastic-Degrading Super-Microbes and Enzymes

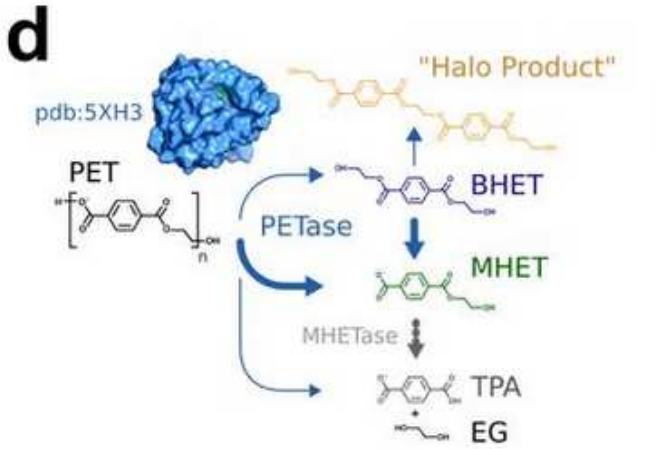


- 213 known PETases
- Let's look in Logan for homologs
- Result: 215M distinct sequences

The Logan project: exciting example

PAST SUCCESS - STARTUP

Plastivores: Plastic-Degrading Super-Microbes and Enzymes

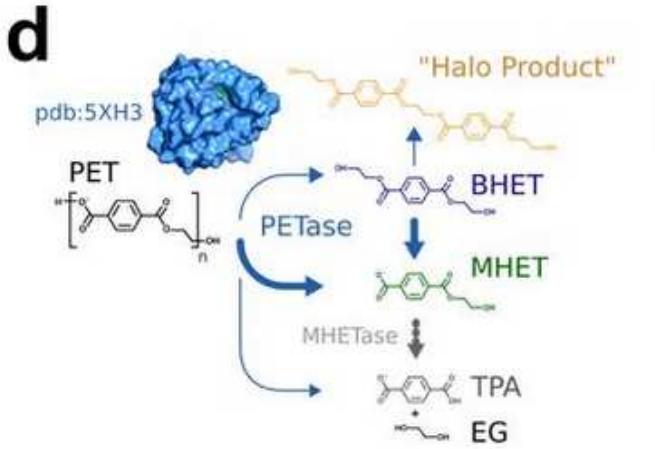


- 213 known PETases
- Let's look in Logan for homologs
- Result: 215M distinct sequences
- Some of them best than previously known PETases

The Logan project: exciting example

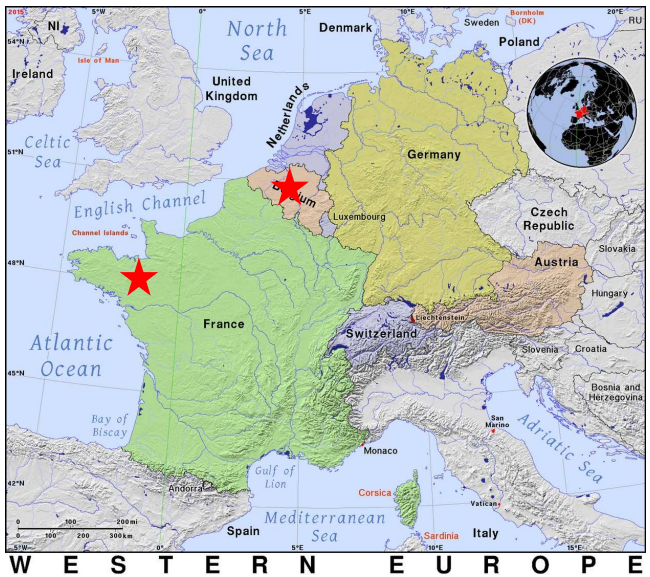
PAST SUCCESS - STARTUP

Plastivores: Plastic-Degrading Super-Microbes and Enzymes



- 213 known PETases
- Let's look in Logan for homologs
- Result: 215M distinct sequences
- Some of them best than previously known PETases
- My job: improving speed/cost of the search

About me: Ph.D., 2021-2024

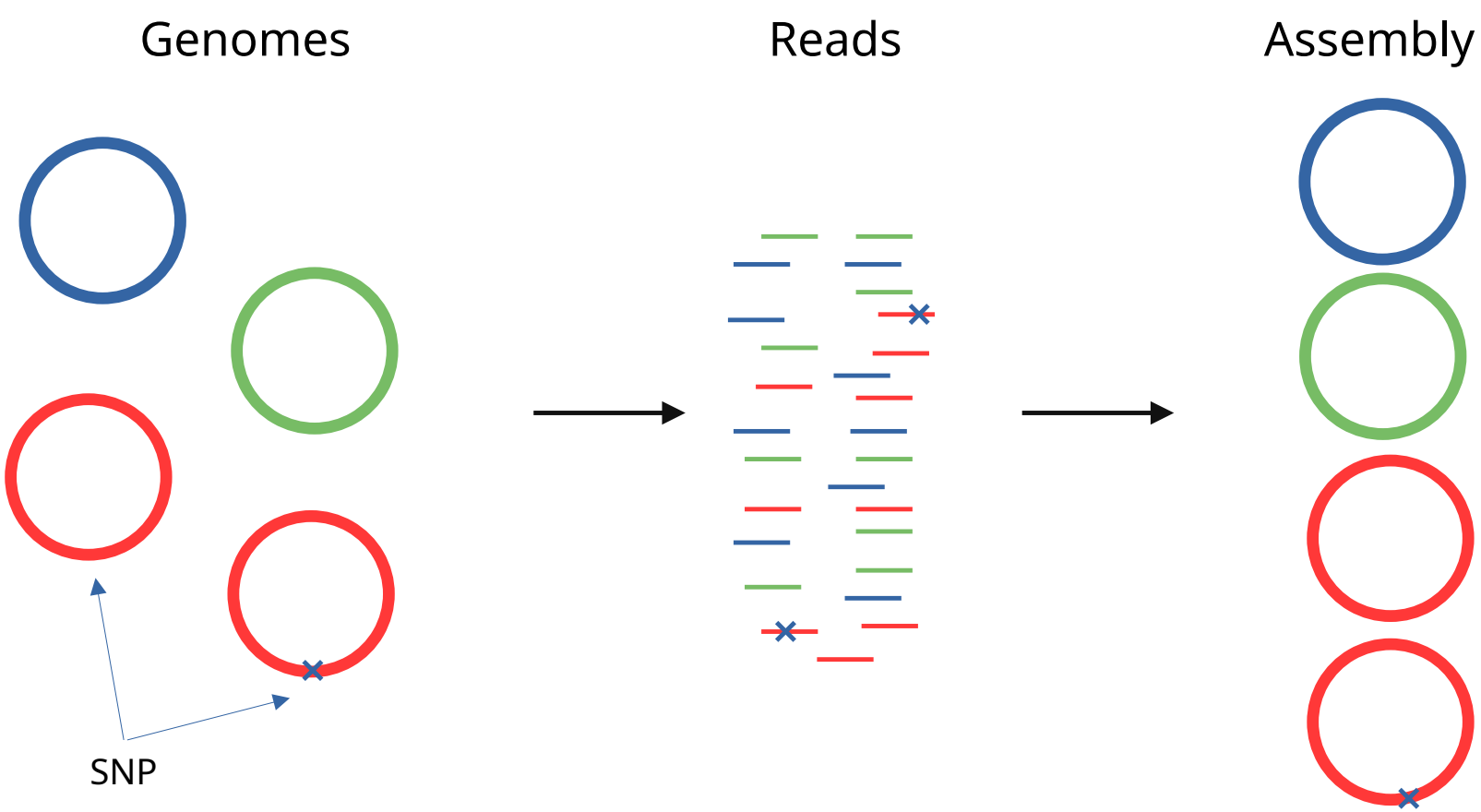


Jean-François Flot
Université Libre de Bruxelles
Focus: assembling wild genomes

Dominique Laveneir
Université de Rennes
Focus: computational methods

My Ph.D. : Haplotype assembly from long reads

Focus of my Ph.D.: 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)

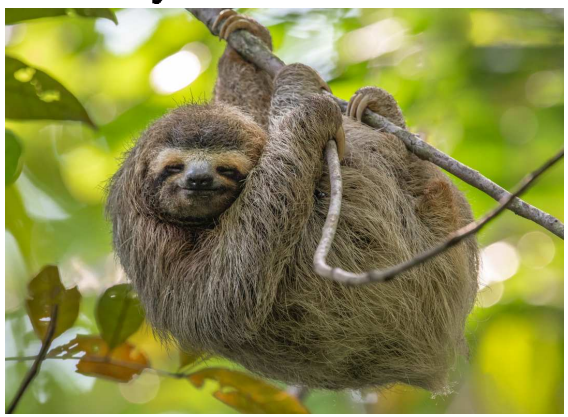
metaFlye
4 days, 256GB RAM

(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

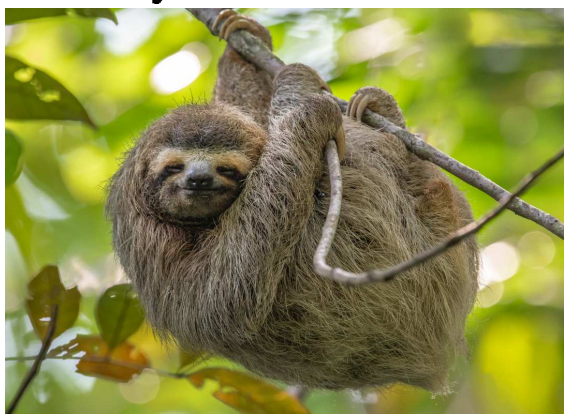


(Meta)genome assembly is a big computation

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

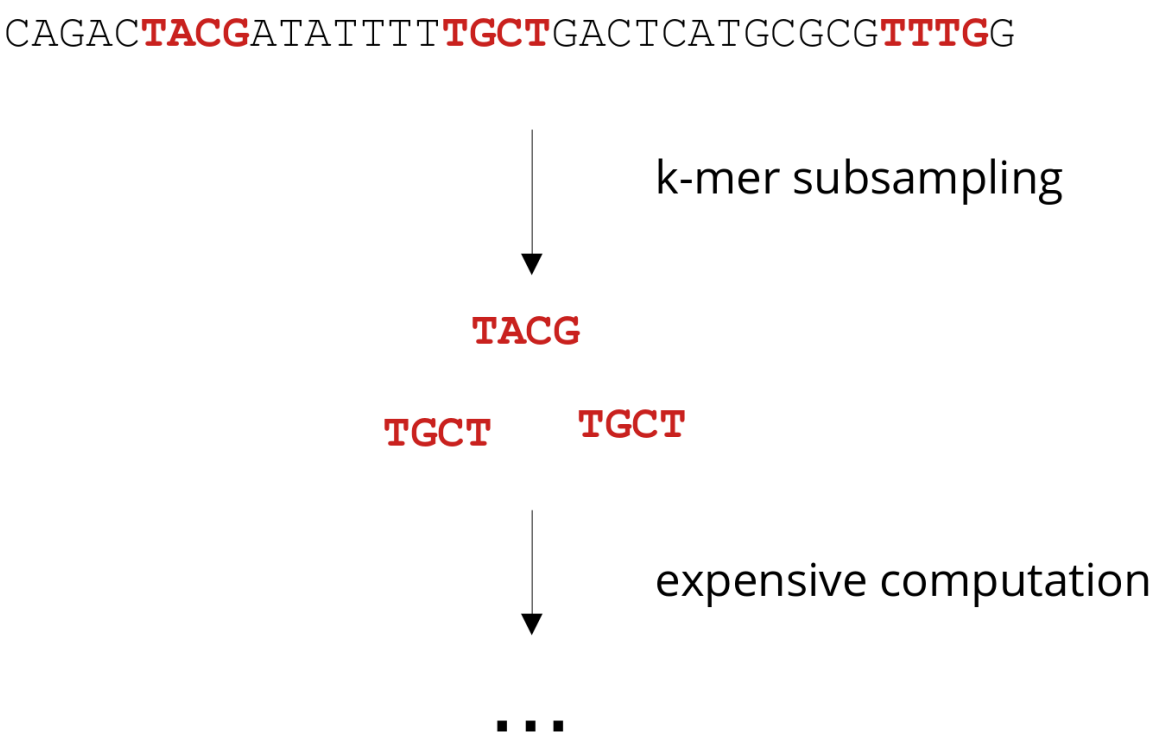
metaFlye
4 days, 256GB RAM

hifiasm_meta
11 days, 454GB RAM



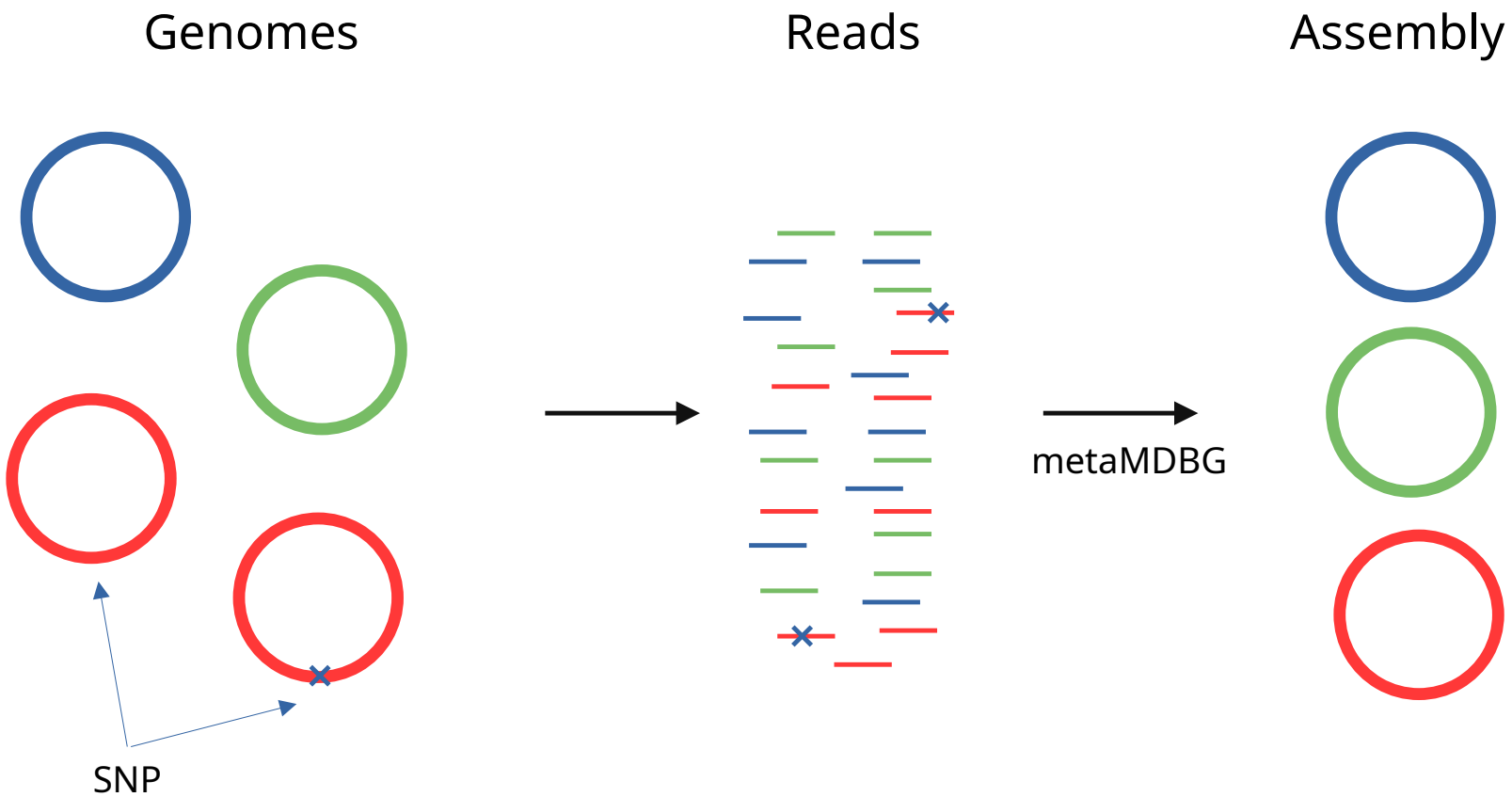
metaMDBG
19h, 10G RAM

metaMDBG: the trick is sketching input reads



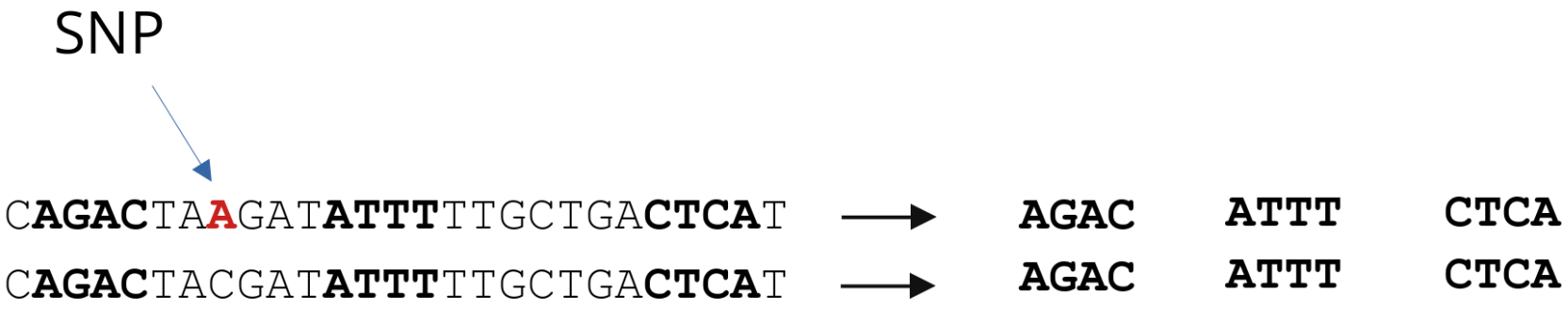
- minimizers, FracMinHash, seed-chain, strobemers...
- minimap2, Mash, BLAST, **metaMDBG**...

metaMDBG loses strain diversity



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

k-mer sketching loses SNPs



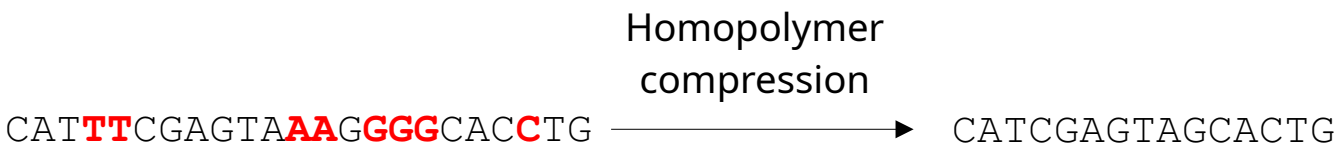
k-mer sketching loses SNPs

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

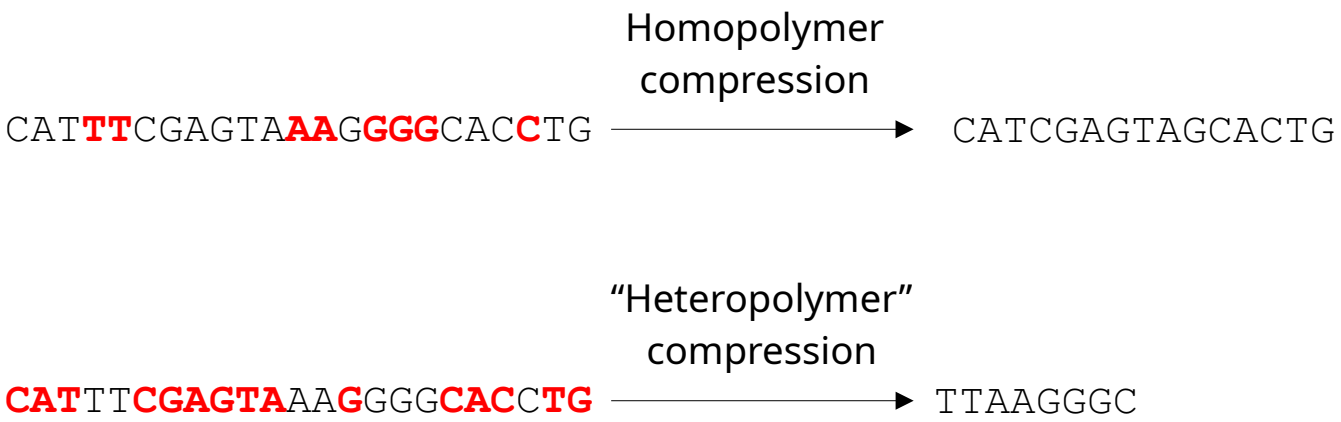
k-mer sketching loses SNPs

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

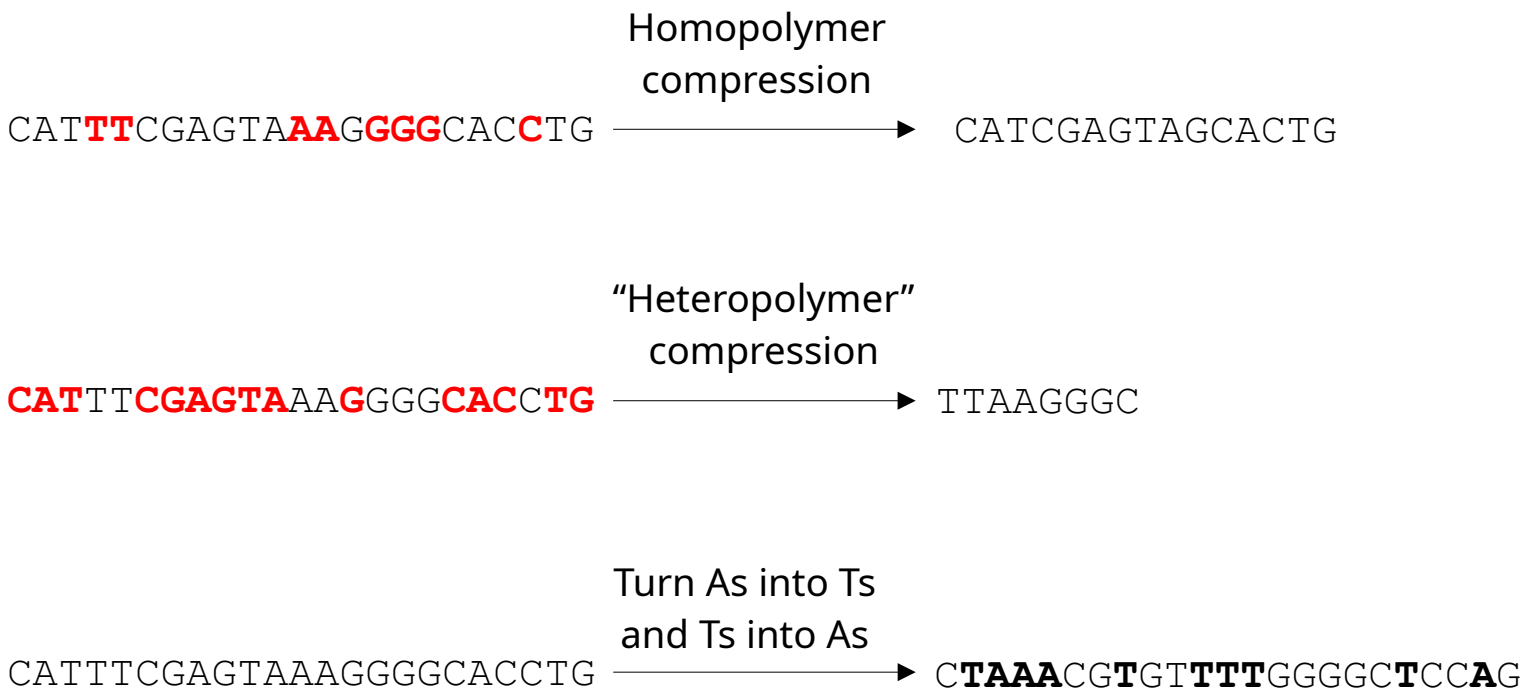
Generalizing Homopolymer Compression



Generalizing Homopolymer Compression



Generalizing Homopolymer Compression



Generalizing Homopolymer Compression

			compression
CATTTCGAGTAAAGGGGCACCTG	Homopolymer compression	CATCGAGTAGCACTG	0.75
CATTTCGAGTAAAGGGGCACCTG	"Heteropolymer" compression	TTAAGGGC	0.25
CATTTCGAGTAAAGGGGCACCTG	Turn As into Ts and Ts into As	CTAAACGTGTTTGGGGCTCCAG	1.0

Mapping-friendly Sequence Reductions: an example

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

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

Mapping-friendly Sequence Reductions: an example

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

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

sequence `CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG`

Mapping-friendly Sequence Reductions: an example

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

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

sequence **CAGTATGGAT**ACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**CAGTATGGAT**)= 0.0023

f(**CAGTATGGAT**)= A

sketch A

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(AGTATGGATA)= 0.624

f(AGTATGGATA)= $\emptyset$

sketch A

Mapping-friendly Sequence Reductions: an example

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

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

sequence CA**GTATGGATAC**AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GTATGGATAC**)= 0.124

f(**GTATGGATAC**)= G

sketch A G

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAG**TATGGATACA**GATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TATGGATACA**)= 0.88

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

sketch A G

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAGT**ATGGATACAG**ATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATGGATACAG**)= 0.32

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

sketch A G

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAGTA**TGGATACAGA**TGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TGGATACAGA**)= 0.19

f(**TGGATACAGA**)= T

sketch A G T

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAGTAT**GGATACAGAT**GGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GGATACAGAT**)= 0.214

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

sketch A G T

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAGTATG**GATACAGATG**GAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GATACAGATG**)= 0.678

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

sketch A G T

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAGTATGG**ATACAGATGG**AGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATACAGATGG**)= 0.669

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

sketch A G T

Mapping-friendly Sequence Reductions: an example

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

hash(**TGTACCAGAG**)= 0.06

f(**TGTACCAGAG**)= C

sketch A G T T C C G T C

Mapping-friendly Sequence Reductions: homopolymer compression

$$f : \{A, C, G, T\}^2 \rightarrow \{A, C, G, T, \emptyset\}$$

$f(AA) \rightarrow \emptyset$	$f(CA) \rightarrow A$	$f(GA) \rightarrow A$	$f(TA) \rightarrow A$
$f(AC) \rightarrow C$	$f(CC) \rightarrow \emptyset$	$f(GC) \rightarrow C$	$f(TC) \rightarrow C$
$f(AG) \rightarrow G$	$f(CG) \rightarrow G$	$f(GG) \rightarrow \emptyset$	$f(TG) \rightarrow G$
$f(AT) \rightarrow T$	$f(CT) \rightarrow T$	$f(GT) \rightarrow T$	$f(TT) \rightarrow \emptyset$

Mapping-friendly Sequence Reductions: homopolymer compression

$$f:\{A,C,G,T\}^2 \rightarrow \{A,C,G,T,\emptyset\}$$

$f(AA)\rightarrow\emptyset$	$f(CA)\rightarrow A$	$f(GA)\rightarrow A$	$f(TA)\rightarrow A$
$f(AC)\rightarrow C$	$f(CC)\rightarrow\emptyset$	$f(GC)\rightarrow C$	$f(TC)\rightarrow C$
$f(AG)\rightarrow G$	$f(CG)\rightarrow G$	$f(GG)\rightarrow\emptyset$	$f(TG)\rightarrow G$
$f(AT)\rightarrow T$	$f(CT)\rightarrow T$	$f(GT)\rightarrow T$	$f(TT)\rightarrow\emptyset$

sequence ACGTTG

sketch

Mapping-friendly Sequence Reductions: homopolymer compression

$$f:\{A,C,G,T\}^2 \rightarrow \{A,C,G,T,\emptyset\}$$

$f(AA)\rightarrow\emptyset$	$f(CA)\rightarrow A$	$f(GA)\rightarrow A$	$f(TA)\rightarrow A$
$f(AC)\rightarrow C$	$f(CC)\rightarrow\emptyset$	$f(GC)\rightarrow C$	$f(TC)\rightarrow C$
$f(AG)\rightarrow G$	$f(CG)\rightarrow G$	$f(GG)\rightarrow\emptyset$	$f(TG)\rightarrow G$
$f(AT)\rightarrow T$	$f(CT)\rightarrow T$	$f(GT)\rightarrow T$	$f(TT)\rightarrow\emptyset$

sequence **AC**GTTG

sketch C

Mapping-friendly Sequence Reductions: homopolymer compression

$$f:\{A,C,G,T\}^2 \rightarrow \{A,C,G,T,\emptyset\}$$

$f(AA)\rightarrow\emptyset$	$f(CA)\rightarrow A$	$f(GA)\rightarrow A$	$f(TA)\rightarrow A$
$f(AC)\rightarrow C$	$f(CC)\rightarrow\emptyset$	$f(GC)\rightarrow C$	$f(TC)\rightarrow C$
$f(AG)\rightarrow G$	$f(CG)\rightarrow G$	$f(GG)\rightarrow\emptyset$	$f(TG)\rightarrow G$
$f(AT)\rightarrow T$	$f(CT)\rightarrow T$	$f(GT)\rightarrow T$	$f(TT)\rightarrow\emptyset$

sequence ACGTTG

sketch CG

Mapping-friendly Sequence Reductions: homopolymer compression

$$f:\{A,C,G,T\}^2 \rightarrow \{A,C,G,T,\emptyset\}$$

$f(AA)\rightarrow\emptyset$	$f(CA)\rightarrow A$	$f(GA)\rightarrow A$	$f(TA)\rightarrow A$
$f(AC)\rightarrow C$	$f(CC)\rightarrow\emptyset$	$f(GC)\rightarrow C$	$f(TC)\rightarrow C$
$f(AG)\rightarrow G$	$f(CG)\rightarrow G$	$f(GG)\rightarrow\emptyset$	$f(TG)\rightarrow G$
$f(AT)\rightarrow T$	$f(CT)\rightarrow T$	$f(GT)\rightarrow T$	$f(TT)\rightarrow\emptyset$

sequence ACGTTG

sketch CGT

Mapping-friendly Sequence Reductions: homopolymer compression

$$f:\{A,C,G,T\}^2 \rightarrow \{A,C,G,T,\emptyset\}$$

$f(AA)\rightarrow\emptyset$	$f(CA)\rightarrow A$	$f(GA)\rightarrow A$	$f(TA)\rightarrow A$
$f(AC)\rightarrow C$	$f(CC)\rightarrow\emptyset$	$f(GC)\rightarrow C$	$f(TC)\rightarrow C$
$f(AG)\rightarrow G$	$f(CG)\rightarrow G$	$f(GG)\rightarrow\emptyset$	$f(TG)\rightarrow G$
$f(AT)\rightarrow T$	$f(CT)\rightarrow T$	$f(GT)\rightarrow T$	$f(TT)\rightarrow\emptyset$

sequence	ACG**TT**G
sketch	CGT

Mapping-friendly Sequence Reductions: homopolymer compression

$$f:\{A,C,G,T\}^2 \rightarrow \{A,C,G,T,\emptyset\}$$

$f(AA)\rightarrow\emptyset$	$f(CA)\rightarrow A$	$f(GA)\rightarrow A$	$f(TA)\rightarrow A$
$f(AC)\rightarrow C$	$f(CC)\rightarrow\emptyset$	$f(GC)\rightarrow C$	$f(TC)\rightarrow C$
$f(AG)\rightarrow G$	$f(CG)\rightarrow G$	$f(GG)\rightarrow\emptyset$	$f(TG)\rightarrow G$
$f(AT)\rightarrow T$	$f(CT)\rightarrow T$	$f(GT)\rightarrow T$	$f(TT)\rightarrow\emptyset$

sequence ACGT**TG**

sketch CGT G

Mapping-friendly Sequence Reductions: definition

$$l \in \mathbb{N}$$

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

$$\forall\, kmer \in \{A, C, G, T\}^l, rc(f(kmer)) = f(rc(kmer))$$

Mapping-friendly Sequence Reductions: key parameters

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

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=Mapping-friendly Sequence Reductions

- MSR reductions are mapping-friendly

- original sequences align $\iff$ reduced sequences align

MSRs=Mapping-friendly Sequence Reductions

- MSR reductions are mapping-friendly

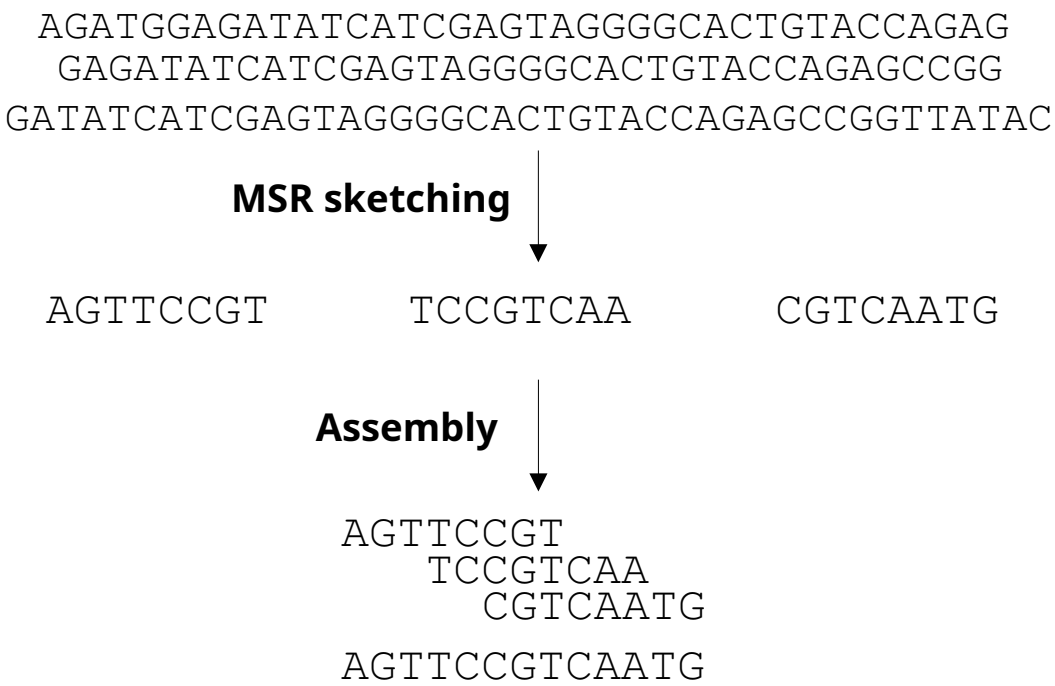
- original sequences align $\iff$ reduced sequences align
- Key property of MSRs

MSRs=Mapping-friendly Sequence Reductions

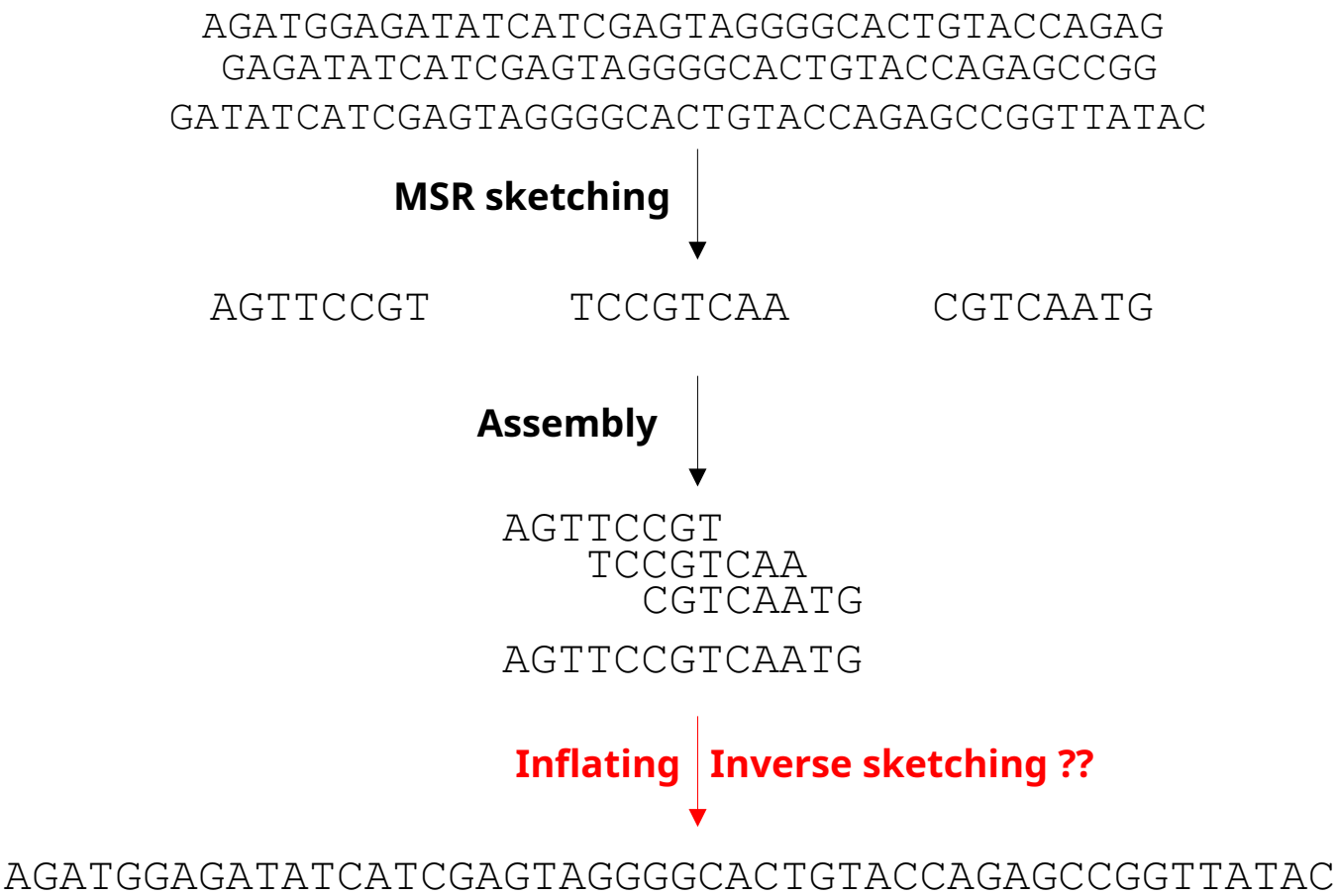
- MSR reductions are mapping-friendly

- original sequences align $\iff$ reduced sequences align
- Key property of MSRs
- Let's compress massively and assemble!

Assembling using MSR sketches



Assembling using MSR sketches



Inflating a reduced assembly

- Keep a record while compressing

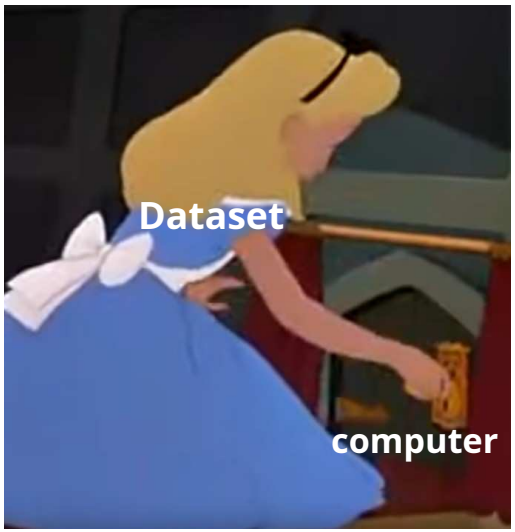
sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch _____**AG**_____**A**_____**G**_____**C**_____**A**_

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

The Alice assembler: assembling with MSR

Input dataset is too big



1. sketch reads

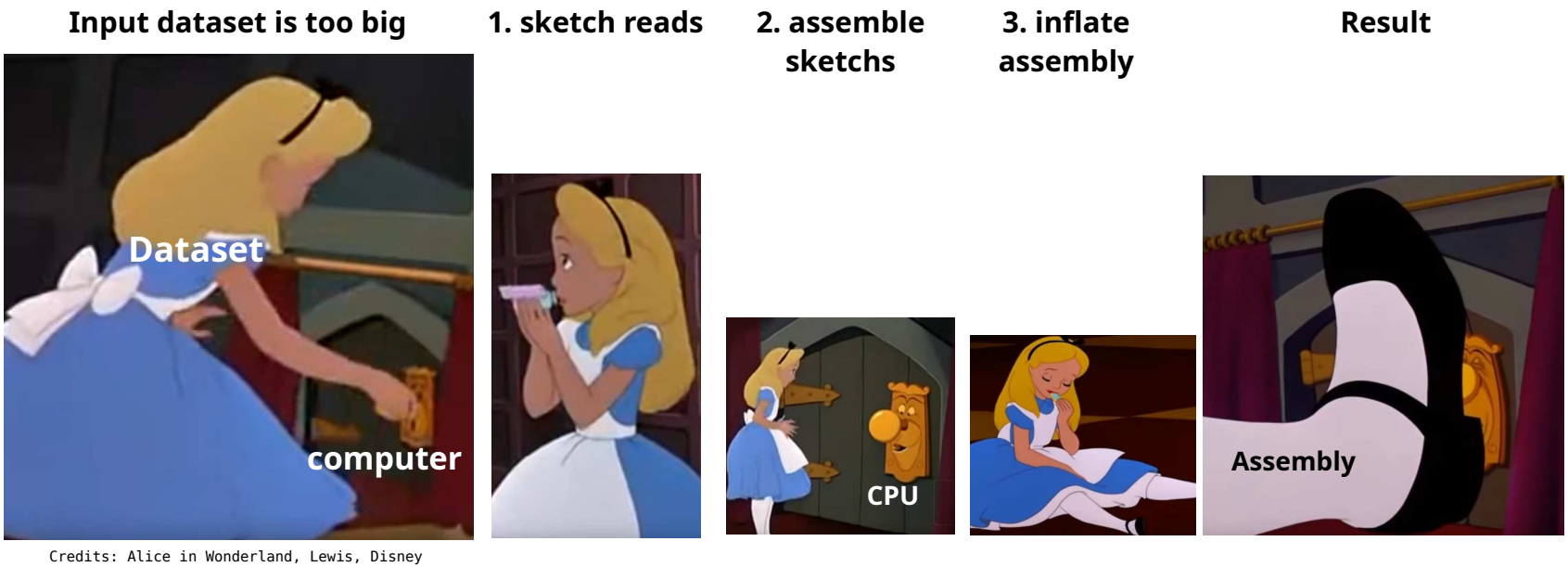
2. assemble sketchs

3. inflate assembly

Result

Credits: Alice in Wonderland, Lewis, Disney

The Alice assembler: assembling with MSR



Credits: Alice in Wonderland, Lewis, Disney

- Function f: chaotic hash function
- Any assembler for step 2., by default BCALM2+tip-clipping
- `github.com/rolandfaure/alice-asm`

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 CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1 A

sequence2 CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2 A

MSRs keep SNPs

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

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

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	CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T T
sequence2	CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG
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

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

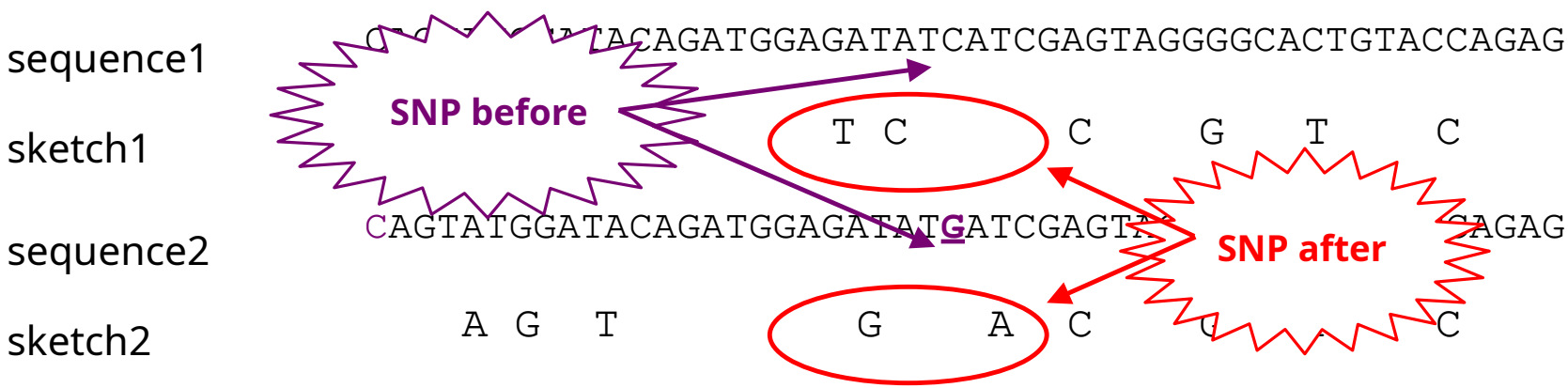
$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \emptyset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

sequence1	CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
sketch1	A G T T C C G T C
sequence2	CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG
sketch2	A G T G A C G T C

MSRs keep and amplify SNPs

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

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



MSRs keep and amplify SNPs

- A SNP affects l l-mers
- Each l-mer outputs a base with probability c
- Probability that a SNP disappears in the sketch:

$$\sum_{i=0}^{l}\left(\binom{l}{i} c^i \cdot (1-c)^{l-i}\right)^2 \cdot \frac{1}{4^i} \approx (1-c)^{2l}$$

MSRs keep and amplify SNPs

- A SNP affects l l-mers
- Each l-mer outputs a base with probability c
- Probability that a SNP disappears in the sketch:

$$\sum_{i=0}^{l}\left(\binom{l}{i} c^i \cdot (1-c)^{l-i}\right)^2 \cdot \frac{1}{4^i} \approx (1-c)^{2l}$$

	k-mer subsampling	MSR
c=0.1	0.81	10^{-10}
c=0.01	0.98	0.13

Table: Probability that a SNP disappears in sketch, using l=101

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

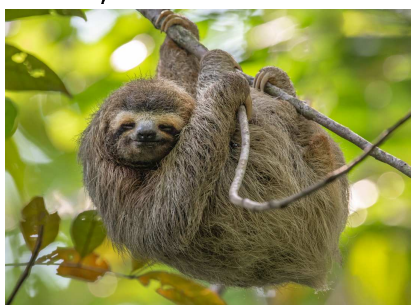
- Strains are not collapsed

The Alice assembler: results

- Assembling a human gut metagenome (HiFi sequencing)

Flye
4d, 256G RAM

hifiasm_meta
11d, 454G RAM



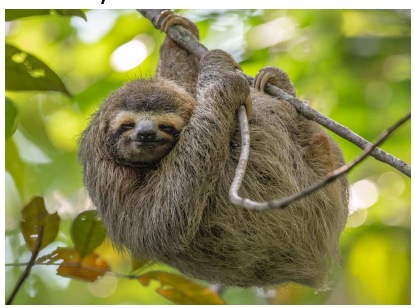
metaMDBG
19h, 10G RAM

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

- Distance between errors $\approx$ Original distance * compression ratio c
- original sequences align $\iff$ reduced sequences align **not completely true**

Potential applications and future MSRs

- MSRs are wild and unexplored

Potential applications and future MSRs

- MSRs are wild and unexplored
- Alignment? SNP calling? Indexing? Whole genome operations (e.g. pangenome graph building)?

Potential applications and future MSRs

- MSRs are wild and unexplored
- Alignment? SNP calling? Indexing? Whole genome operations (e.g. pangenome graph building)?
- Changing the MSR itself: error rate? Biology-informed MSR?