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

Roland Faure[1,2,3], Jean-François Flot[1], Dominique Lavenier[2]

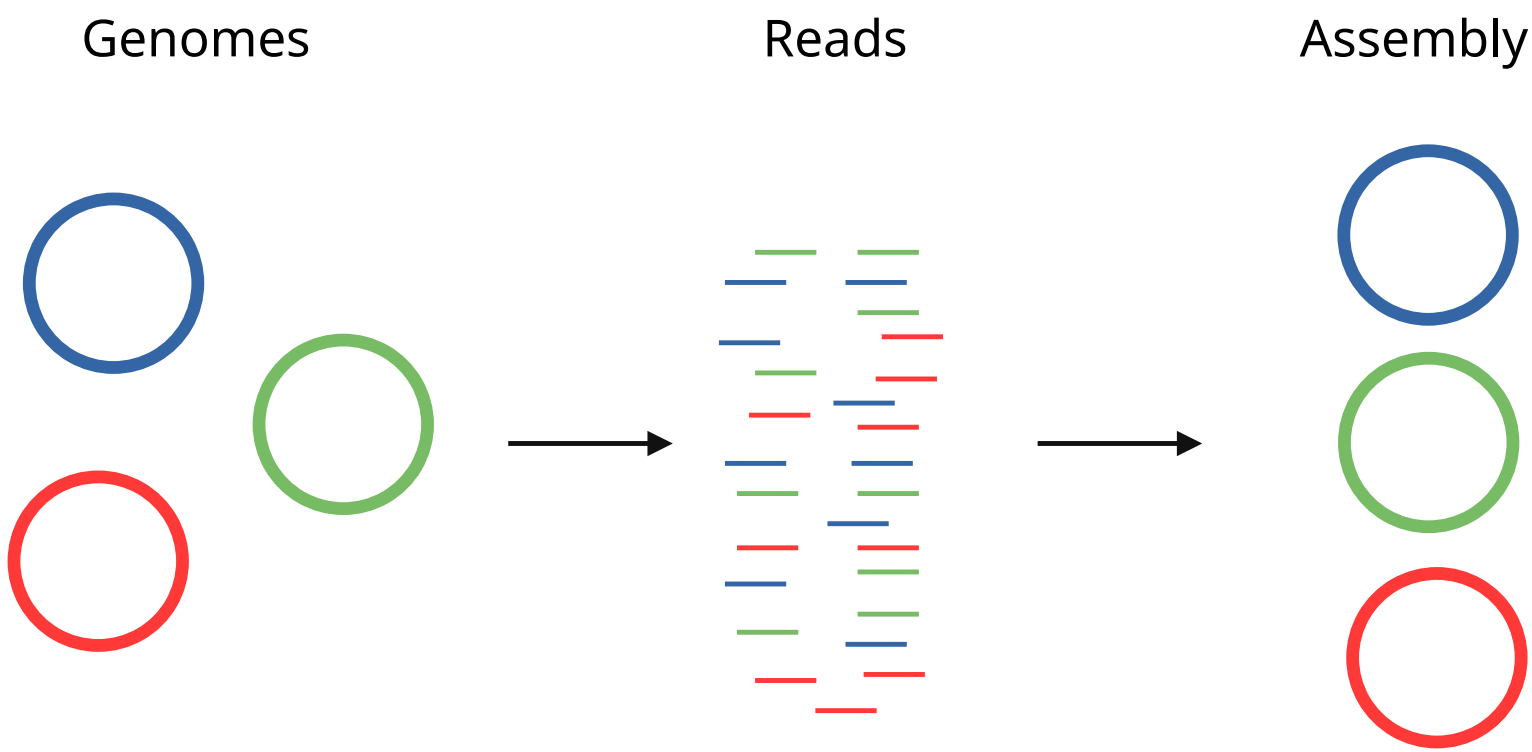
[1]Université libre de Bruxelles (ULB) - Belgium

[2]Université de Rennes, IRISA - France

[3]Insitut 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

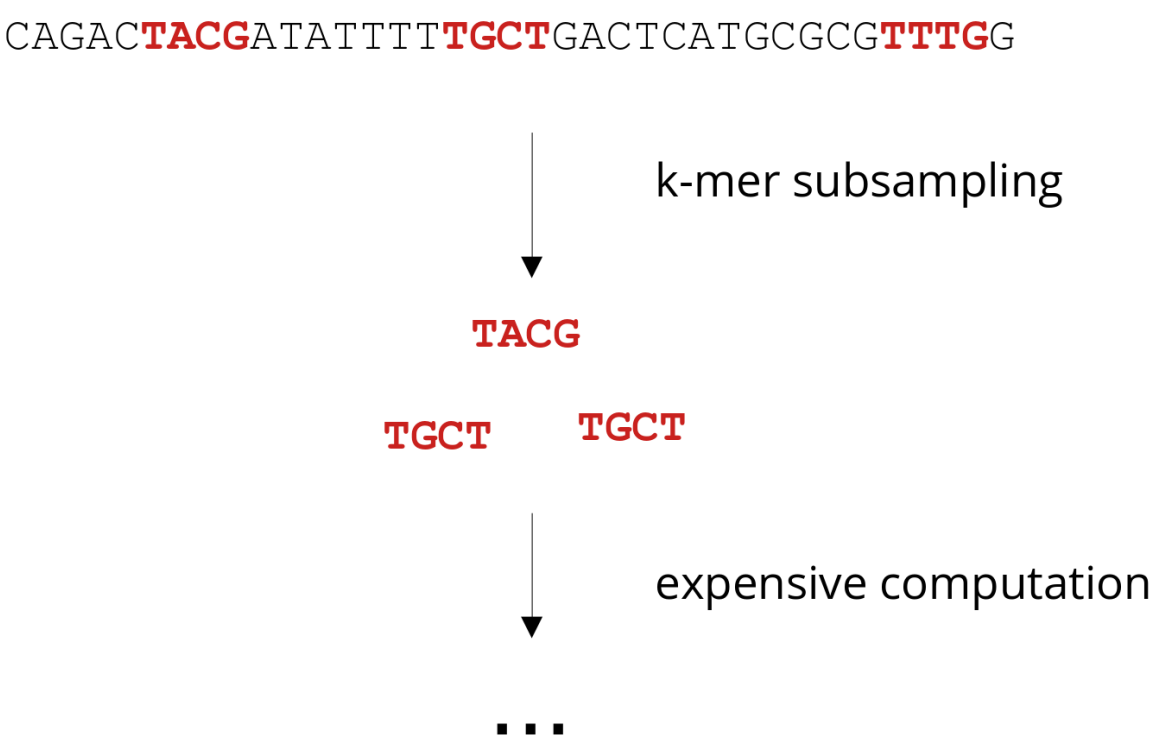
- Assembling a human gut metagenome (HiFi, 250Gpb)

metaFlye
4 days, 256GB RAM

hifiasm_meta
11 days, 454GB RAM

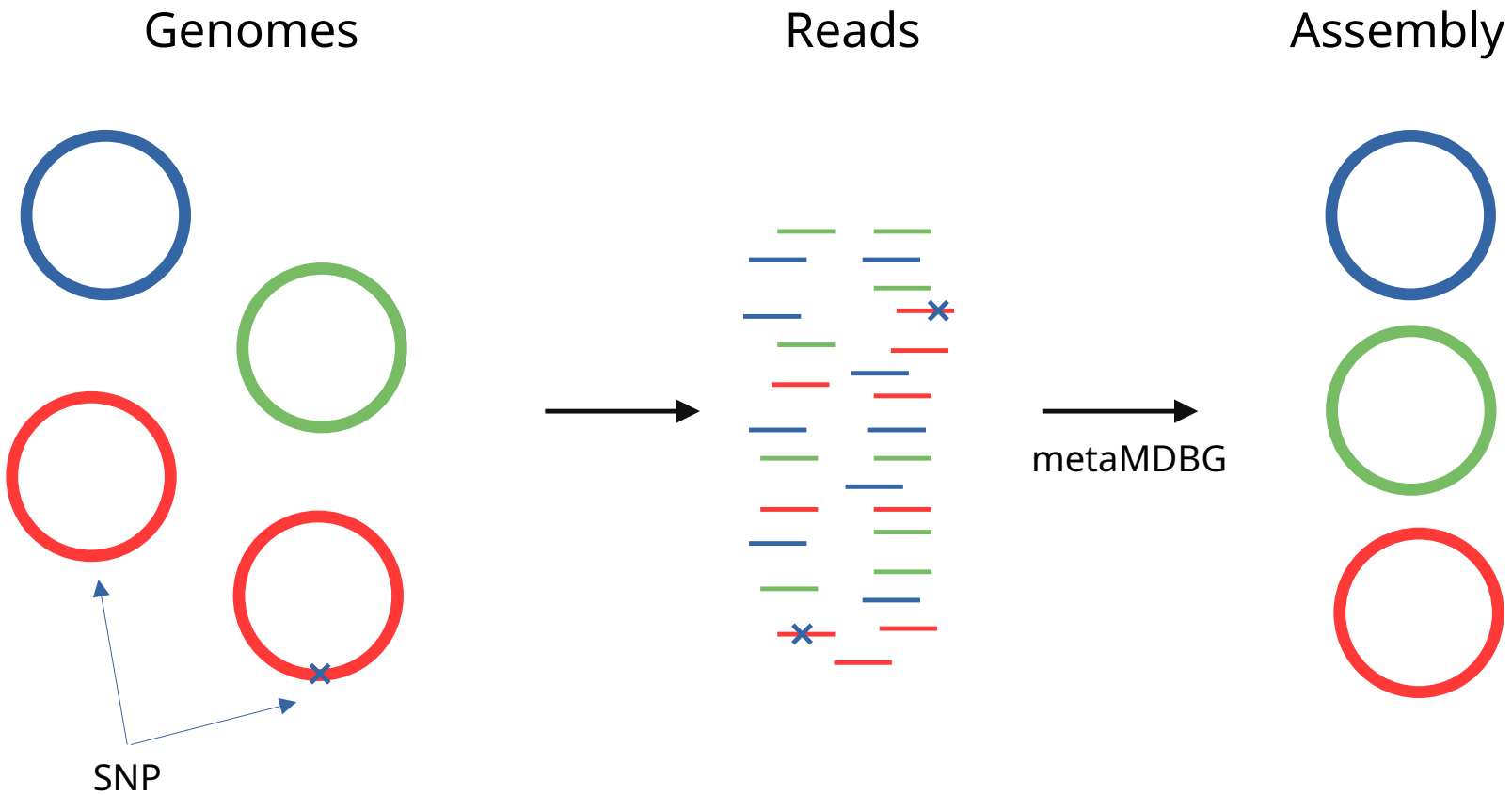
metaMDBG
19h, 10G RAM

metaMDBG: the trick is sketching input reads



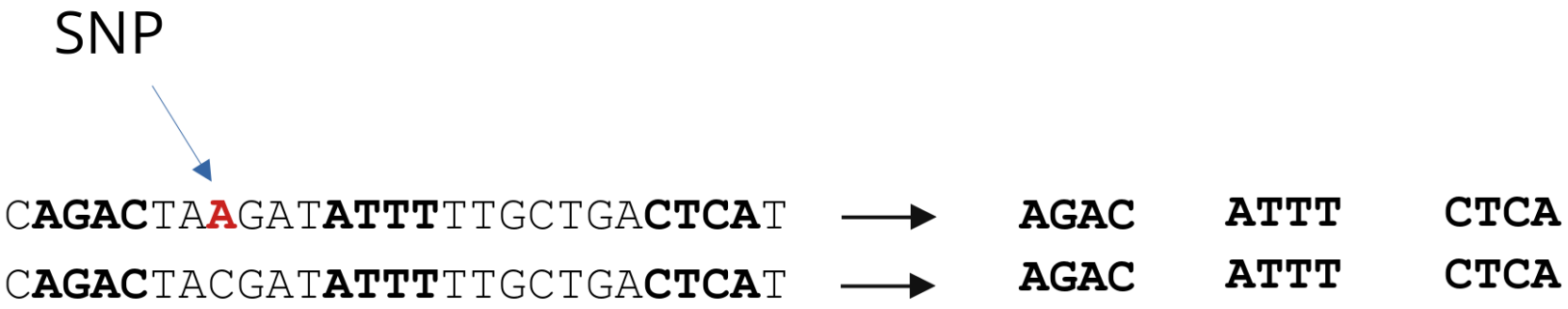
- minimizers, FracMinHash, seed-chain, strobemers...
- minimap2, Mash, BLAST, **metaMDBG**...

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.

Mapping-friendly Sequence Reductions

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

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

Mapping-friendly Sequence Reductions

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

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

sequence `CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG`

Mapping-friendly Sequence Reductions

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

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

sequence **CAGTATGGAT**ACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**CAGTATGGAT**)= 0.0023

f(**CAGTATGGAT**)= A

sketch A

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(AGTATGGATA)= 0.624

f(AGTATGGATA)= $\emptyset$

sketch A

Mapping-friendly Sequence Reductions

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

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

sequence CA**GTATGGATAC**AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GTATGGATAC**)= 0.124

f(**GTATGGATAC**)= G

sketch A G

Mapping-friendly Sequence Reductions

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

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

sequence CAG**TATGGATACA**GATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TATGGATACA**)= 0.88

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

sketch A G

Mapping-friendly Sequence Reductions

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

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

sequence CAGT**ATGGATACAG**ATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATGGATACAG**)= 0.32

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

sketch A G

Mapping-friendly Sequence Reductions

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

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

sequence CAGTA**TGGATACAGA**TGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TGGATACAGA**)= 0.19

f(**TGGATACAGA**)= T

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTAT**GGATACAGAT**GGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GGATACAGAT**)= 0.214

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

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATG**GATACAGATG**GAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GATACAGATG**)= 0.678

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

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATGG**ATACAGATGG**AGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATACAGATGG**)= 0.669

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

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

hash(**TGTACCAGAG**)= 0.06

f(**TGTACCAGAG**)= C

sketch A G T T C C G T C

Mapping-friendly Sequence Reductions

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

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

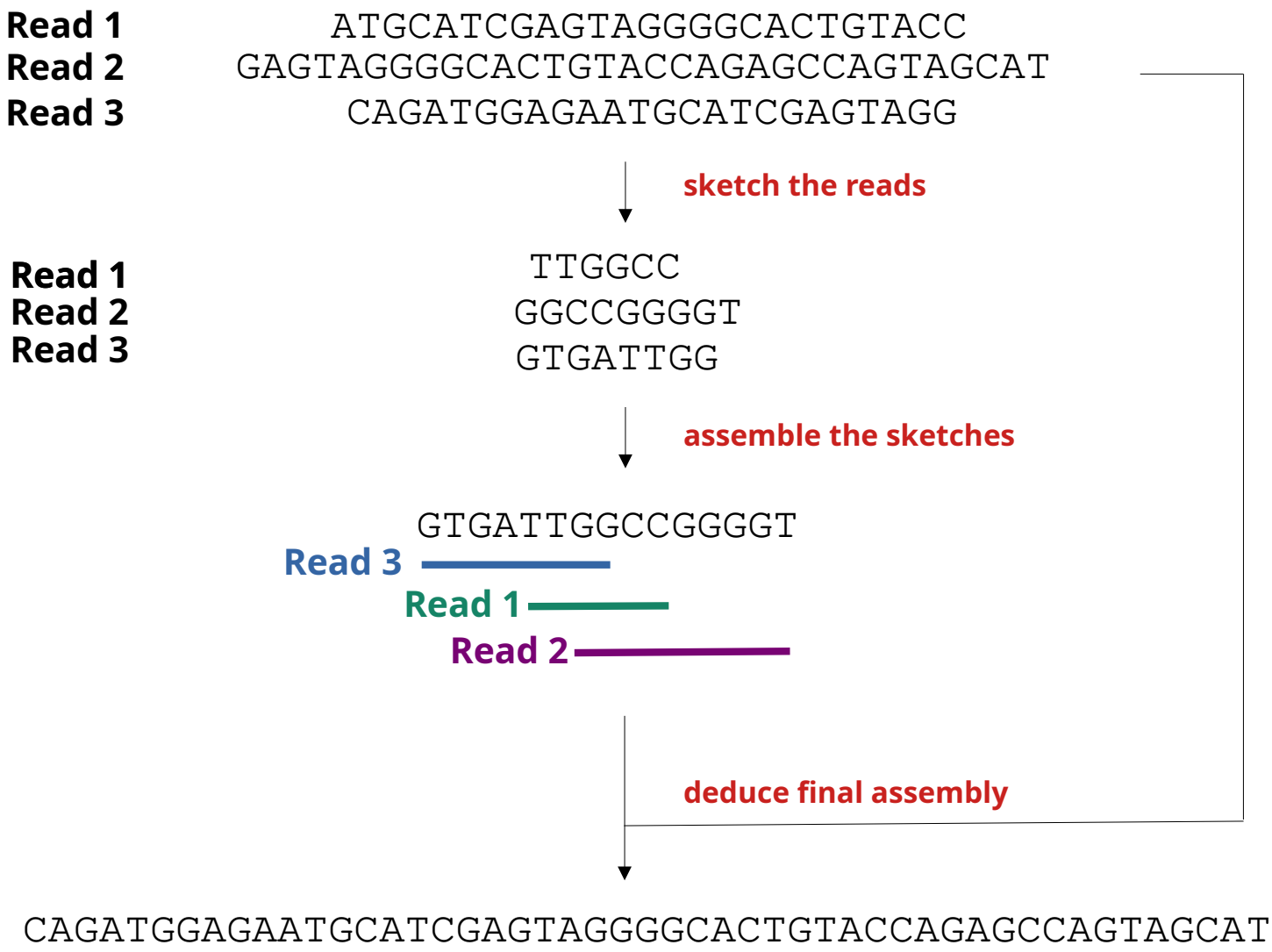
- original sequences align $\iff$ reduced sequences align

MSRs=Mapping-friendly Sequence Reductions

- MSR reductions are mapping-friendly

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

Assembling using MSR sketches



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	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

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

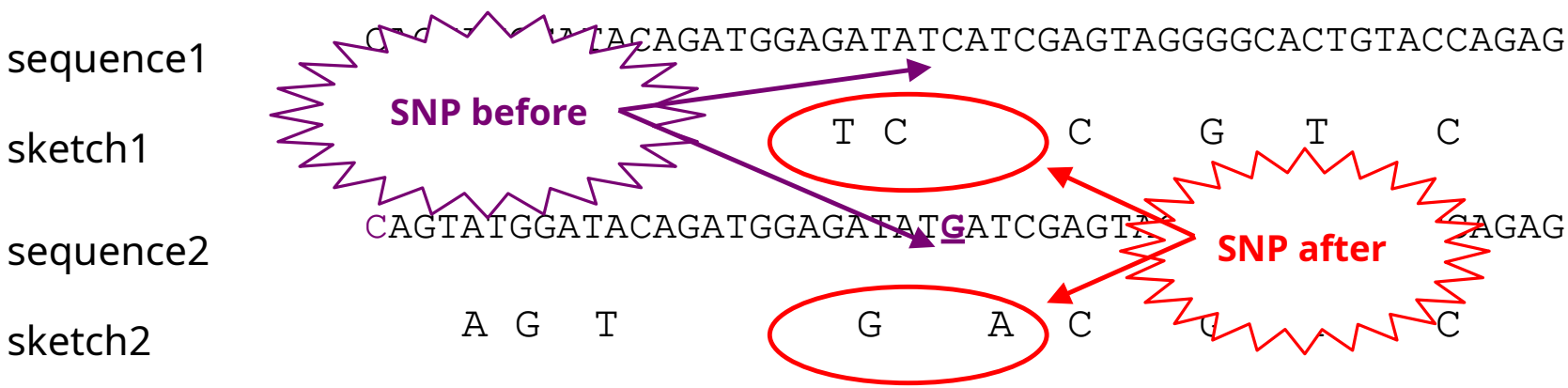
$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]$$
$$f(10-mer) \rightarrow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

sequence1	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

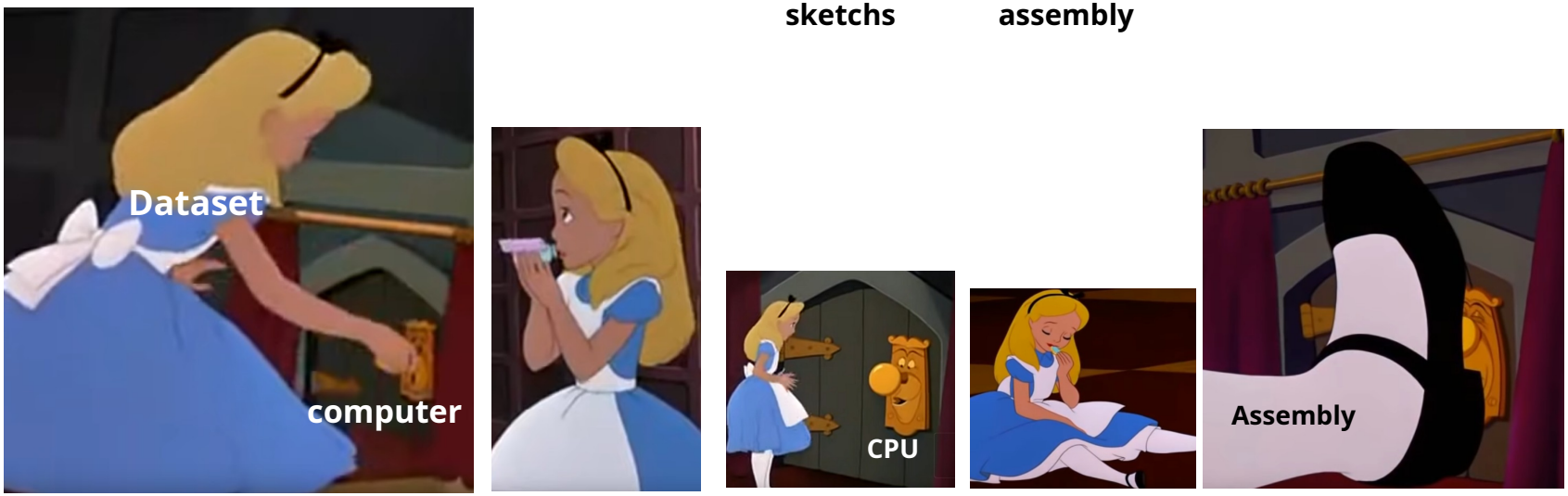
$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

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



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

- 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 (%)

	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

- 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

Metagenomic datasets in the future

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

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 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

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

hash(**TGTACCAGAG**)= 0.06
f(**TGTACCAGAG**)= C

sketch A G T T C C G T C

- Changing f, l, c