A new sketching method for genomic data

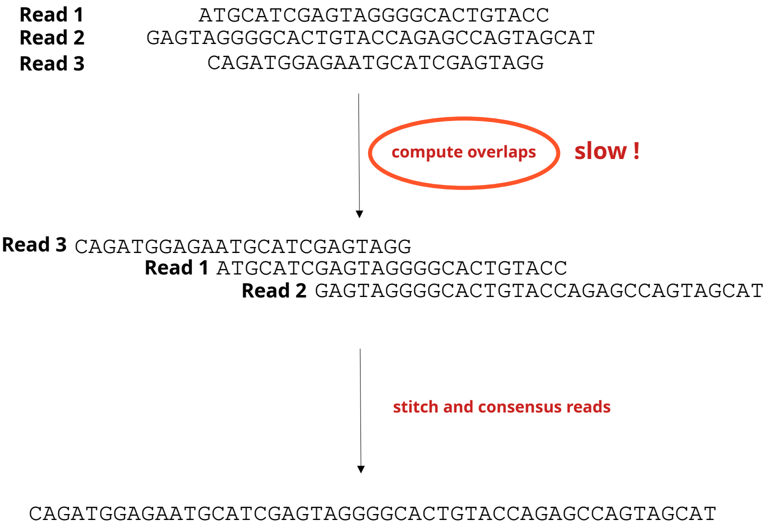
Roland Faure[1,2], Baptiste Hilaire[2], Jean-François Flot[1], Dominique Lavenier[2]

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

[2]Université de Rennes, IRISA - France

DSB 2025

Genome assembly is poooooo slooooooow



```
Read 1      ATGCATCGAGTAGGGGCACTGTACC
Read 2      GAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
Read 3      CAGATGGAGAATGCATCGAGTAGG
```

compute overlaps

slow !

```
Read 3 CAGATGGAGAATGCATCGAGTAGG
                Read 1 ATGCATCGAGTAGGGGCACTGTACC
                       Read 2 GAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
```

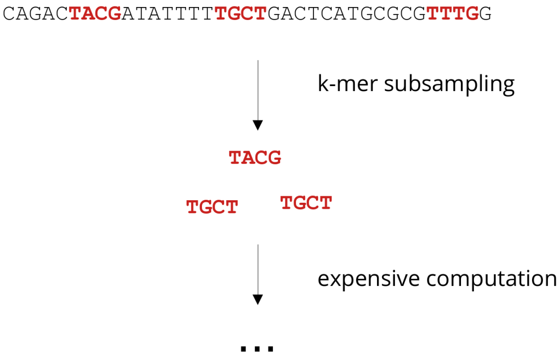
stitch and consensus reads

```
CAGATGGAGAATGCATCGAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
```

Sequencing data is too big: what can I do?

Sequencing data is too big: what can I do?

Sketching with sequence subsampling



CAGAC**TACG**ATATTTT**TGCT**GACTCATGCGCG**TTTG**G

k-mer subsampling

TACG

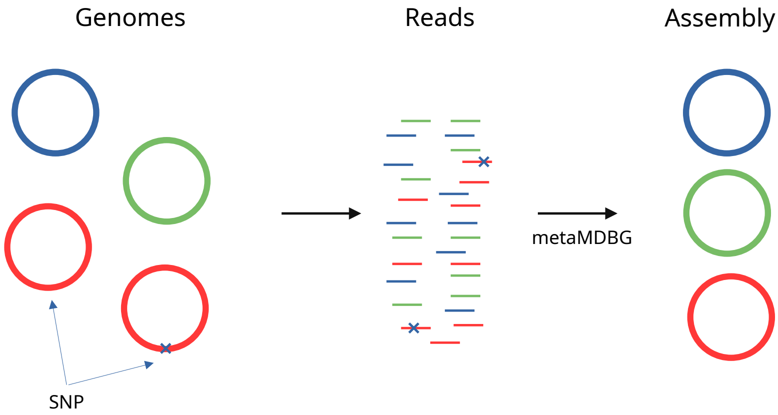
TGCT **TGCT**

expensive computation

...

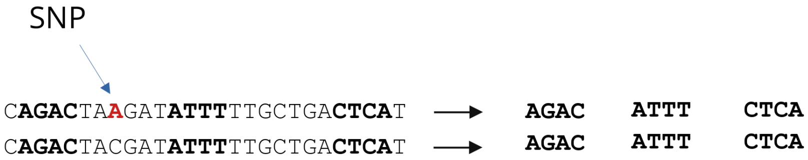
- minimizers, FracMinHash, seed-chain, strobemers...
- minimap2, Mash, BLAST, **metaMDBG**...

Sketched metagenome assembly



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

Sequence subsampling does not preserve SNPs

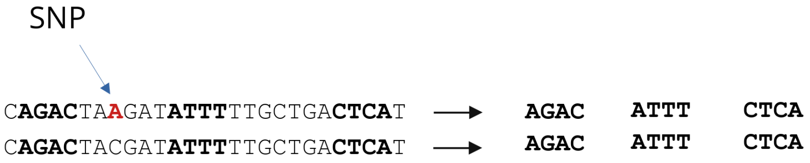


SNP

C**AGAC**TA**A**GAT**ATTT**TTGCTGA**CTCA**T → **AGAC** **ATTT** **CTCA**

C**AGAC**TACGAT**ATTT**TTGCTGA**CTCA**T → **AGAC** **ATTT** **CTCA**

Sequence subsampling does not preserve SNPs



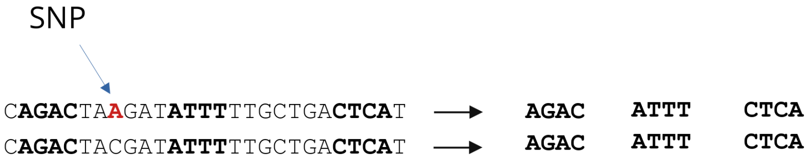
SNP

C**AGAC**TA**A**GAT**ATTT**TTGCTGA**CTCA**T → **AGAC** **ATTT** **CTCA**

C**AGAC**TACGAT**ATTT**TTGCTGA**CTCA**T → **AGAC** **ATTT** **CTCA**

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

Sequence subsampling does not preserve SNPs



SNP

C**AGAC**TA**A**GAT**ATTT**TTGCTGA**CTCA**T ⟶ **AGAC** **ATTT** **CTCA**

C**AGAC**TACGAT**ATTT**TTGCTGA**CTCA**T ⟶ **AGAC** **ATTT** **CTCA**

- 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

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

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

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

Mapping-friendly Sequence Reductions

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

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

sequence **CAGTATGGAT**ACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**CAGTATGGAT**)= 0.0023
f(**CAGTATGGAT**)= A

sketch A

Mapping-friendly Sequence Reductions

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

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

sequence C**AGTATGGATA**CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**AGTATGGATA**)= 0.624

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

sketch A

Mapping-friendly Sequence Reductions

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

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

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

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

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

sequence CAGT**ATGGATACAG**ATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATGGATACAG**)= 0.32
f(**ATGGATACAG**)= $\emptyset$

sketch A G

Mapping-friendly Sequence Reductions

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

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

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

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

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

sequence CAGTATG**GATACAGATG**GAGATATCATCGAGTAGGGCACTGTACCAGAG

hash(**GATACAGATG**)= 0.678

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

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATGG**ATACAGATGG**AGATATCATCGAGTAGGGCACTGTACCAGAG

hash(**ATACAGATGG**)= 0.669

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

sketch A G T

Mapping-friendly Sequence Reductions

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

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

sketch A G T T C C G T C

Mapping-friendly Sequence Reductions

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

order (l)

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

compression ratio (c)

```
sequence    CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch          A G  T        T C      C    G   T    C
```

Mapping-friendly Sequence Reductions (MSR)

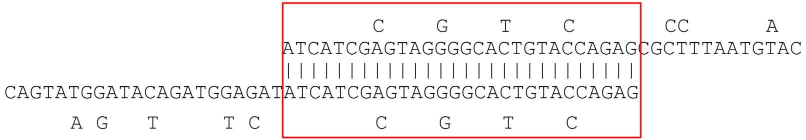
- MSR transform sequences into smaller sequences
- Computed in a streaming fashion
- Reverse-complement property on f: ∀ seq,
  f(reverse_comp(seq)) = reverse_comp(f(seq))

Mapping-friendly Sequence Reductions (MSR)

- MSR transform sequences into smaller sequences
- Computed in a streaming fashion
- Reverse-complement property on f: $\forall$ seq, f(reverse_comp(seq)) = reverse_comp(f(seq))
- Can they be used as a sketching method?

MSRs=Mapping-friendly Sequence Reductions

- MSR reductions are mapping-friendly

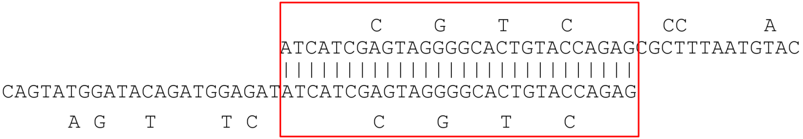


```
                             C    G    T    C       CC      A
                      ATCATCGAGTAGGGGCACTGTACCAGAGCGCTTTAATGTAC
                      ||||||||||||||||||||||||||||
CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
     A G   T     T C         C    G    T    C
```

- original sequences align $\iff$ reduced sequences align

MSRs=Mapping-friendly Sequence Reductions

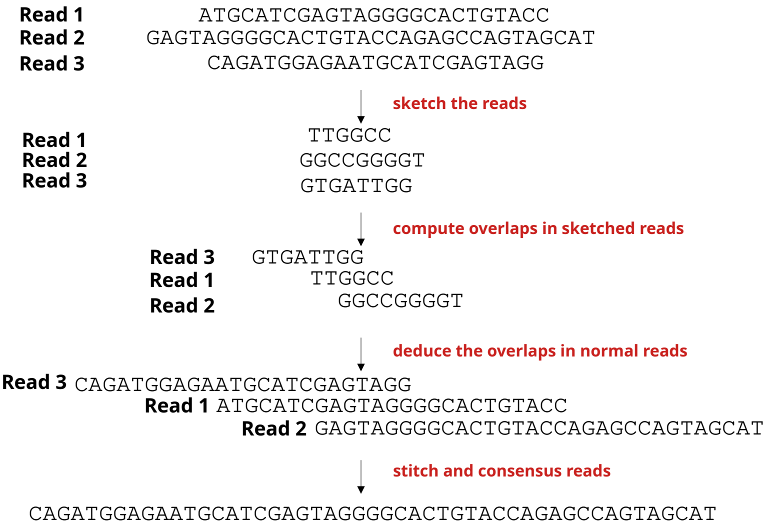
- MSR reductions are mapping-friendly



```
                             C    G    T    C       CC      A
                      ATCATCGAGTAGGGGCACTGTACCAGAGCGCTTTAATGTAC
                      ||||||||||||||||||||||||||||
CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
     A G   T     T C         C    G    T    C
```

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

MSRs can be used as sketches



```
Read 1     ATGCATCGAGTAGGGGCACTGTACC
Read 2 GAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
Read 3      CAGATGGAGAATGCATCGAGTAGG
```

sketch the reads

```
Read 1 TTGGCC
Read 2 GGCCGGGGT
Read 3 GTGATTGG
```

compute overlaps in sketched reads

```
Read 3 GTGATTGG
Read 1     TTGGCC
Read 2       GGCCGGGGT
```

deduce the overlaps in normal reads

```
Read 3 CAGATGGAGAATGCATCGAGTAGG
Read 1           ATGCATCGAGTAGGGGCACTGTACC
Read 2                  GAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
```

stitch and consensus reads

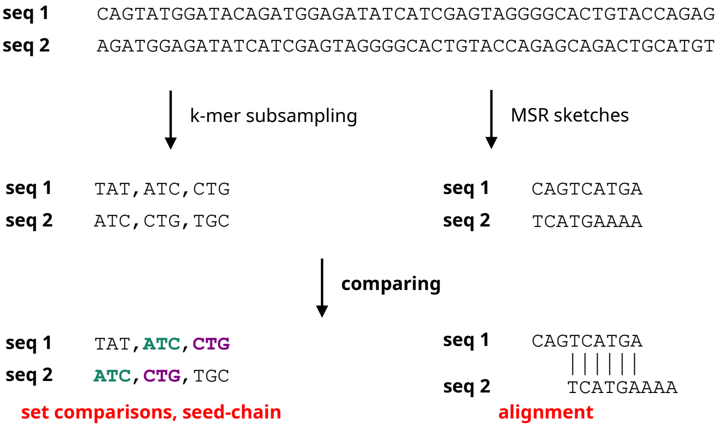
```
CAGATGGAGAATGCATCGAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
```

MSR sketches vs k-mer subsampling

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

	MSR sketching	**k-mer subsampling**
	AGTTCCGTC	TAT,CGA,CCA
Storing & compression	Fasta, BWT	Bloom filters

Comparing two sequences with MSR



seq 1 CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

seq 2 AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAGCAGACTGCATGT

k-mer subsampling

MSR sketches

seq 1 TAT,ATC,CTG

seq 2 ATC,CTG,TGC

seq 1 CAGTCATGA

seq 2 TCATGAAAA

comparing

seq 1 TAT,**ATC**,**CTG**

seq 2 **ATC**,**CTG**,TGC

set comparisons, seed-chain

seq 1 CAGTCATGA

seq 2 TCATGAAAA

alignment

MSR sketches vs k-mer subsampling

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

	MSR sketching	**k-mer subsampling**
	AGTTCCGTC	TAT,CGA,CCA
Storing & compression	Fasta, BWT	Bloom filters
Seq comparison	Alignment	Set comparisons, seed-chain

MSRs keep SNPs

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

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

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

MSRs keep SNPs

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

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

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

MSRs keep SNPs

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

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

sequence1 CA**GTATGGATAC**AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1 A G

sequence2 CA**GTATGGATAC**AGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG

sketch2 A G

MSRs keep SNPs

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

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

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

MSRs keep SNPs

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

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

```
sequence1   CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1         A G   T         T

sequence2   CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2         A G   T
```

MSRs keep SNPs

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

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

```
sequence1  CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1        A G   T         T

sequence2  CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2        A G   T          G
```

MSRs keep SNPs

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

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

sequence1 CAGTATGGATACAGATG**GAGATATCAT**CGAGTAGGGGCACTGTACCAGAG

sketch1 A G T T C

sequence2 CAGTATGGATACAGATG**GAGATATGAT**CGAGTAGGGGCACTGTACCAGAG

sketch2 A G T G

MSRs keep SNPs

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

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

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

MSRs keep SNPs

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

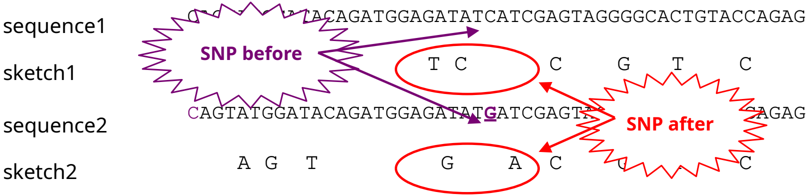
$$\begin{aligned}
&f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]\\
&f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]\\
&f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]\\
&f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]\\
&f(10-mer)\rightarrow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

```
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

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

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



sequence1 TACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1 T C C G T C

sequence2 CAGTATGGATACAGATGGAGATATGATCGAGTA AGAG

sketch2 A G T G A C C

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:

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

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:

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

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

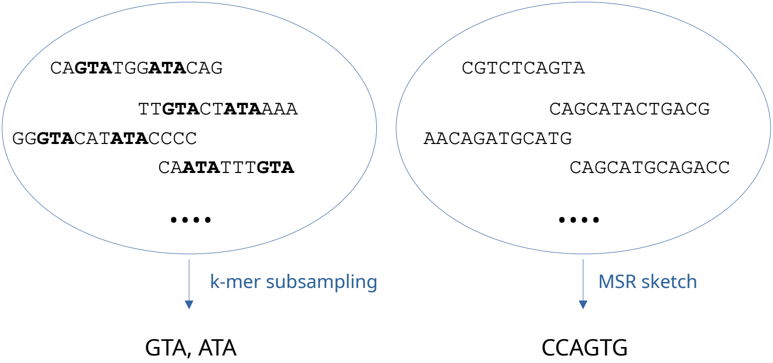
Table: Probability that a SNP disappears in sketch, using l=101

MSRs keep and amplify SNPs

- Yet MSR do not keep more information than k-mers

MSRs keep and amplify SNPs

- Yet MSR do not keep more information than k-mers



CA**GTA**TGG**ATA**CAG

TT**GTA**CT**ATA**AAA

GG**GTA**CAT**ATA**CCCC

CA**ATA**TTT**GTA**

....

k-mer subsampling

GTA, ATA

CGTCTCAGTA

CAGCATACTGACG

AACAGATGCATG

CAGCATGCAGACC

....

MSR sketch

CCAGTG

- MSR sketching is more chaotic than k-mer subsampling (tunable with l)

MSR sketches vs k-mer subsampling

`CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG`

	MSR sketching	**k-mer subsampling**
	`AGTTCCGTC`	`TAT,CGA,CCA`
Storing & compression	Fasta, BWT	Bloom filters
Seq comparison	Alignment	Set comparisons, seed-chain
SNPs	Mostly kept	Mostly discarded

The Alice assembler: assembling with MSR

**1.
sketch the reads**

**2.
assemble the sketchs**

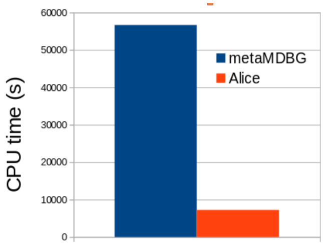
**3.
inflate the assembly**

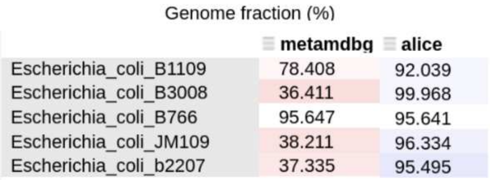
`Credits: Alice in Wonderland, Lewis, Disney`

- 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*



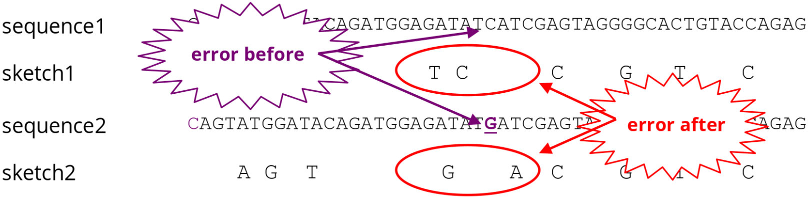

Genome fraction (%)

	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
- New error rate ≈ Original error rate / compression rate c

MSR sketches vs k-mer subsampling

CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

	MSR sketching	**k-mer subsampling**
	AGTTCCGTC	TAT,CGA,CCA
Storing & compression	Fasta, BWT	Bloom filters
Seq comparison	Alignment	Set comparisons, seed-chain
SNPs and errors	Mostly kept	Mostly discarded

MSR sketching: take-home messages

MSR sketching: take-home messages

- MSR sketches are sequences and can be manipulated as such

MSR sketching: take-home messages

- MSR sketches are sequences and can be manipulated as such
- MSR sketches keep & amplify differences between sequences

MSR sketching: take-home messages

- MSR sketches are sequences and can be manipulated as such
- MSR sketches keep & amplify differences between sequences
- Alice HiFi assembler out soon (`github.com/rolandfaure/alice-asm`)

MSR sketching: take-home messages

- MSR sketches are sequences and can be manipulated as such
- MSR sketches keep & amplify differences between sequences
- Alice HiFi assembler out soon (`github.com/rolandfaure/alice-asm`)
- There are other things than k-mer subsampling in life

Perspectives

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

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

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCAC**TGTACCAGAG**

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

sketch A G T T C C G T C

- Changing f, l, c