Travis Gagie
Dalhousie University
et al.

Data Compression Conference March 23rd, 2023

## Travis Gagie et al.

## DNA alignment

reference bias EDI pangenomes graphs

WGS

## theorem (sketch)

data structure counting locating MEM-finding

## scaling up

timeline RLCSA Toehold Lemma r-index

data structure examples

#### 100

future work thanks to...

# **DNA** alignment

Travis Gagie et al.

## DNA alignment

WGS reference bias

referenc

pangenomes graphs

theorem (sketch)

#### M-index

data structure counting locating MEM-finding

## scaling up

timeline

RLCSA Toehold Lemma

Toehold Lemr r-index

MONI

#### ΜΔΡΙΔ

data structure examples

theorem

#### conclusion

future work thanks to...

questions?

## WGS

genome	G	Α	T	Α	C	Α	T
read 1 read 2	G	Α	T	Α			
read 2		Α	T	Α	С		
read 3				Α	С	Α	T

Travis Gagie et al.

## DNA alignment

WGS reference bias

reterence b

pangenomes graphs theorem (sketch)

#### EM index

data structure counting locating

MEM-finding

scaling up

timeline

RLCSA Toehold Lemma

r-index

## MONI MARIA

data structure examples

theorem

conclusion

future work

augstions

WGS

reference	G	Α	T	T	Α	C	Α	T
read 1		Α	T	-	Α			
read 2		Α	-	T	Α	C		
read 3					Α	C	Α	T
output	G	Α	T	-	Α	C	Α	T

Travis Gagie et al.

## DNA alignment

WGS

reference bias

EDI

pangenomes graphs theorem (sketch)

#### M index

data structure counting locating

MEM-finding

MEM-finding

scaling up

RLCSA

Toehold Lemn r-index

## data structure

examples theorem

future work

questions?

## reference bias

reference	G	Α	T	T	Α	C	Α	T
read 1		Α	T	-	Α			
read 2		A	_	Τ	Α	G		
read 3					Α	G	Α	Τ
output	G	A	T	-	Α	С	A	T

Travis Gagie et al.

## reference bias

wgs reference bias

pangenomes graphs theorem (sketch)

data structure

locating MEM-findin

scaling up timeline RLCSA Toehold Lemma

MARIA

data structure examples

conclusio

future work thanks to... To the scientists' puzzlement, however, the boy's sequence showed no sign of the mutation in the gene known to cause Baratela Scott, called XYLT1. Nor did the DNA of the next boy with the disorder, or the next. As they tried to compare the boys' DNA sequences to the reference genome, it was like trying to check a spelling in a Webster's from which a prankster had torn handfuls of pages. Many pieces of the boys' genomes, called short reads, "weren't in the reference genome at all," ... There was no way to check them for disease-causing misspellings.

— Sharon Begley, Stat News, March 11th, 2019

Travis Gagie et al.

DNA alignmen

reference bias

pangenomes grap

theorem (sketch)

FM-indexe

counting locating MEM-findin

MEM-finding scaling up

timeline RLCSA Toehold Lemm

r-index

MARIA

examples

conclusio

future work thanks to...

## reference bias

This bias limits the kind of genetic variation that can be detected, leaving some patients without diagnoses and potentially without proper treatment. What is more, people who share less ancestry with the man from Buffalo will probably benefit less from the incoming era of precision medicine, which promises to tailor healthcare to individuals.

[O]ur understanding of diversity within populations of European descent is now so good that we can start to use it for precision medicine. But for other populations, "We do not have the same kind of data ... [This] is going to increase healthcare disparities above and beyond what they are today." ... [A] huge new project is offering a different solution with the aim to represent global diversity: a human pangenome.

— Ida Emilie Steinmark, Guardian, January 29th, 2023

## Travis Gagie et al.

## DNA alignment

reference hise

### EDI

pangenomes graphs theorem (sketch)

## theorem (sketch)

data structure counting

locating MEM-finding

### scaling up timeline RLCSA

Toehold Lemi r-index

### MARIA

data structure examples

#### . . . . . . . . . . . .

future work thanks to...



— webpage of the *Human Pangenome*Reference Consortium

Travis Gagie et al.

DNA alignmen

WGS

reference bias

pangenomes graphs

theorem (sketch)

data structure

locating MEM-findin

scaling up timeline RLCSA Toehold Lemma

r-index MONI

WAKIA

examples theorem

conclusion

future work thanks to...

## **EDI**

Around 38,000 cancer patients in England and approximately 154,000 patients in the US are initiated on fluoropyrimidine-based [chemotherapy] treatments every year [but] between 20% and 30% of the people who receive these drugs require lower doses, because their bodies struggle to process them. If given the standard dose, they experience reactions which can vary from severe to fatal.

In recent years, genetic-sequencing studies have started to get to the bottom of why some people react so badly ... The only problem is that these studies were done entirely on white people ... "Ethnic minority patients will usually be given conventional doses of the drugs ... Some of these patients will carry other ethnic-specific variants which also affect their ability to metabolise these drugs, but we do not currently genotype for those, largely because we do not know."

— David Cox, BBC, February 28th, 2023

Travis Gagie et al.

WGS reference hise

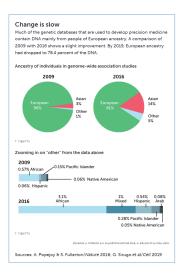
FDI

pangenomes graphs theorem (sketch)

### scaling up

r-index

## **FDI**



— Tina Hesman Saey, Science News, March 21st, 2021

Travis Gagie et al.

DNA alignmen

WGS reference hiss

reference b

pangenomes graphs

theorem (sketch)

FM-indexe

data structure counting

locating MEM-finding

scaling up

Toehold Lemn

MARIA

data structure examples theorem

conclusion

future work thanks to...

## pangenome graphs

[T]he project is not just about sequencing more diverse data. "We need to come up with a better data structure to encode that information," ... That data structure is called a genome graph. In contrast to the current reference, which is just a long string of letters, the genome graph shows variation between genomes as detours on an otherwise shared path. That will enable researchers and doctors to map short reads to the version of the path that best fits their sample.

The natural question is: how does one choose who gets to represent the world? ... 350 people might do a better job of representing the world than one person, but "[the consortium] have made some choices about groups ... Who did they sample? Who did they not sample?" As long as the reference contains only a subset, arguably someone will not make the cut.

— Ida Emilie Steinmark, Guardian, January 29th, 2023

Travis Gagie et al.

### DNA alignmen

WGS reference bias

pangenomes graphs

## theorem (sketch)

FM-indexe

data struct counting locating

locating MEM-finding

MEM-finding scaling up

timeline RLCSA

Toehold Lemm r-index

## MARIA

examples theorem

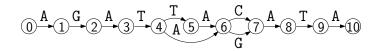
### conclusion

future work

## pangenome graphs

## dataset:

- GATTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA



Travis Gagie et al.

WGS

reference hise

pangenomes graphs

theorem (sketch)

MEM-finding

scaling up

r-index

## pangenome graphs

## dataset:

- **GATTACAT**
- AGATACAT
- GATACAT
- **GATTAGAT**
- GATTAGATA





Travis Gagie et al.

WGS

reference hise

## pangenomes graphs

theorem (sketch)

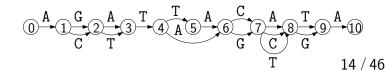
MEM-finding

## scaling up

## pangenome graphs

## dataset:

- GATTACAT
- AGATACAT
  - GATACAT
- **GATTAGAT**
- GATTAGATA
- CATTACAT
- GTTAGAT
- GATTCCATA
- GATTACAGA



Travis Gagie et al.

WGS reference hise

## pangenomes graphs

theorem (sketch)

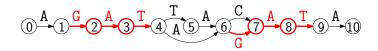
MEM-finding scaling up

r-index

## pangenome graphs

## dataset:

- **GATTACAT**
- AGATACAT
- **GATACAT**
- **GATTAGAT**
- GATTAGATA



Travis Gagie et al.

WGS reference hise

## pangenomes graphs

theorem (sketch)

MEM-finding

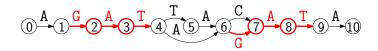
scaling up

r-index

## pangenome graphs

## dataset:

- **GATTACAT**
- AGATACAT
- GATACAT
- **GATTAGAT**
- GATTAGATA



Travis Gagie et al.

WGS reference hise

theorem (sketch)

scaling up

## Theorem (sketch)

Given a pangenomic dataset of total length n and a pangenome graph for it, we can store the dataset in reasonable space — that is, much less than n — such that when given a read of length m, we can

- 1) find good seeds in the read with respect to the dataset in  $O(m \log n)$  time
- 2 for each seed, list all the distinct vertices in the graph where we start processing occurrences in the dataset of that seed, in constant time per vertex listed.

## Pangenomic FM-indexes Travis Gagie et al.

WGS reference bias

pangenomes graphs

## theorem (sketch)

## FM-indexes

data structure counting

MEM-finding

### scaling up

RLCSA

r-index

theorem

# FM-indexes

Travis Gagie et al.

### DNA alignment

WGS reference bias

EDI

pangenomes graphs theorem (sketch)

#### M-indexe

data structure

counting

locating

MEM-finding

## scaling up

timeline

RLCSA

Toehold Lemi r-index

MONI

#### MARIA

data structure

......

#### conclusion

future work

questions?

## data structure

SGATTACAT
ACATSGATTAC
ATTACATSG
CATSGATTA
GATTACATS
TSGATTACA
TACATSGAT
TTACATSGA

Travis Gagie et al.

### DNA alignment

WGS reference bias

EDI

pangenomes graphs theorem (sketch)

#### M-index

data structure
counting
locating

MEM-finding

scaling up

RLCSA

Toehold Lemn

r-index

#### MARIA

data structure

theore

#### conclusio

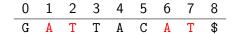
future work

questions

## counting



SGATTACAT
ACATSGATTAC
ATTACATSG
CATSGATTA
GATTACATS
TSGATTACA
TACATSGAT
TTACATSGAT



Travis Gagie et al.

### DNA alignment

WGS reference bias

reference

pangenomes graphs theorem (sketch)

#### FM index

data structur counting

locating

MEM-finding

## scaling up

timeline

RLCSA

Toehold Lemi

r-index

#### MARIA

data structure examples

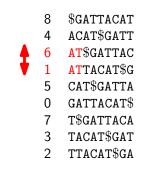
theore

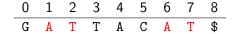
#### conclusion

future work

questions

## locating





Travis Gagie et al.

### DNA alignmen

WGS
reference bias
EDI

theorem (sketch)

FM-indexe

data struct counting

### MEM-finding

scaling up timeline RLCSA

Toehold Lemm r-index MONI

### MARIA

data structur examples

conclusion

future work thanks to... questions?

## **MEM-finding**

A substring P[i..j] of P is a maximal exact match (MEM) between P and T if

- P[i..j] occurs in T
- neither P[i-1..j] nor P[i..j+1] occurs in T.

The MEMs of GATAGAT with respect to GATTACAT are GAT (twice) and TA, but with respect to

 $\label{thm:cats} {\tt GATTACAT\$AGATACAT\$GATTAGAT\$GATTAGATA\$}$  they are GATA and TAGAT.

pangenomes graphs theorem (sketch)

data structure counting locating

locating MEM-finding

scaling up timeline RLCSA Toehold Lemma

MARIA

data stru examples theorem

conclusio

future work thanks to... questions? We can store a text T[1..n] over  $\{A, C, G, T\}$  in 2n + o(n) bits such that when given a pattern P[1..m], we can

MEM-finding

- **1** find the MEMs of P with respect to T in  $O(m \log^{1+\epsilon} n)$  time
- 2 list their occurrences' positions in T in  $O(\log^{1+\epsilon} n)$  time per occurrence.

## Why not just scale this up for pangenomic alignment?

- 1 2n + o(n) is way too big when n is huge (even just o(n) may be too big!)
- 2 the MEMs could occur too often in the dataset

## Travis Gagie et al.

### DNA alignment

WGS reference bias

reference b

pangenomes graphs theorem (sketch)

#### =M\_index

data structure counting locating MEM-finding

### scaling up

timeline

RLCSA

r-index

#### . . . . . . .

data structure

examples

#### conclusion

future work thanks to...

questions

# scaling up

Pangenomic FM-indexes					
Travis Gagie et al.		fast	fast	fast MEM-	
DNA alignment		counting	locating	finding	scalable
WGS					
reference bias EDI pangenomes graphs theorem (sketch)	FM-index ( <i>JACM</i> , 2005)	yes	yes	yes	no
FM-indexes  data structure  counting  locating  MEM-finding	RLCSA ( <i>JCB</i> , 2010)	yes	no	no	yes
scaling up timeline RLCSA	Toehold Lemma ( <i>Algorithmica</i> , 2018)	yes	1	no	yes
Toehold Lemma r-index MONI MARIA	r-index ( <i>JACM</i> , 2020)	yes	yes	no	yes
data structure examples theorem	MONI ( <i>JCB</i> , 2022)	no	yes	yes	yes
future work thanks to questions?					/ -
					25 / 46

Pangenomic FM-indexes					
Travis Gagie et al.		fast	fast	fast MEM-	
DNA alignment		counting	locating	finding	scalable
WGS	-				
reference bias					
EDI .	$FM ext{-index}$	yes	yes	yes	no
pangenomes graphs theorem (sketch)		yes	yes	yes	110
	( <i>JACM</i> , 2005)				
FM-indexes	DI CCA				
data structure	RLCSA	yes	no	no	yes
counting	( <i>JCB</i> , 2010)				
MEM-finding	(362, 2010)				
scaling up	Toehold Lemma	yes	1	no	yes
timeline	(Algorithmica, 2018)				
RLCSA	( "g" )				
Toehold Lemma	r-index	yes	yes	no	yes
r-index MONI		yes	yes	110	yes
	( <i>JACM</i> , 2020)				
MARIA					
data structure	MONI	no	yes	yes	yes
examples	( ICP 2022)		•	,	•
theorem	( <i>JCB</i> , 2022)				
conclusion					
future work					
thanks to					
questions?					
					26 / 46
					= = / . •

Pangenomic

### DNA alignment

WGS reference bias

pangenomes graphs

theorem (sketch)

data structure

locating MEM-finding

scaling up

RLCSA

Toehold Lemi r-index

MONI MARIA

examples

conclusio

future work thanks to... questions? GATTACAT\$ AGATACAT\$ GATACAT\$ GATTAGAT\$ GATTAGATA\$

BWT	context	BWT	context	BWT	context
T	\$AGATACA	G	AT\$GATTA	-\$	GATTAGAT
T	\$GATACA	G	ATA\$GATTA	\$	GATTAGATA
T	\$GATTACA	G	ATACAT\$A	Α	T\$AGATAC
T	\$GATTAGA	G	ATACAT\$	Α	T\$GATAC
Α	\$GATTAGAT	G	ATTACAT\$	Α	T\$GATTAC
T	A\$GATTAGA	G	ATTAGAT\$	Α	T\$GATTAG
T	ACAT\$AGA	G	ATTAGATA\$	Α	TA\$GATTAG
T	ACAT\$GA	Α	CAT\$AGAT	Α	TACAT\$AG
T	ACAT\$GAT	Α	CAT\$GAT	Α	TACAT\$G
T	AGAT\$GAT	Α	CAT\$GATT	T	TACAT\$GA
T	AGATA\$GAT	Α	GAT\$GATT	T	TAGAT\$GA
\$	AGATACAT	Α	GATA\$GATT	T	TAGATA\$GA
C	AT\$AGATA	Α	GATACAT\$	Α	TTACAT\$G
C	AT\$GATA	\$	GATACAT	Α	TTAGAT\$G
C	AT\$GATTA	\$	GATTACAT	Α	${\tt TTAGATA\$G}$

### DNA alignment

WGS reference bias

pangenomes graphs

theorem (sketch)

data structure

locating MEM-finding

MEM-finding scaling up

RLCSA

Toehold Lem

r-index MONI

MARIA

examples

conclusio

future work thanks to... GATTACAT\$ AGATACAT\$ GATACAT\$ GATTAGAT\$ GATTAGATA\$

BWT	context	BWT	context	BWT	context
Т	\$AGATACA	G	AT\$GATTA	\$	GATTAGAT
T	\$GATACA	G	ATA\$GATTA	\$	GATTAGATA
T	\$GATTACA	G	ATACAT\$A	Α	T\$AGATAC
T	\$GATTAGA	G	ATACAT\$	Α	T\$GATAC
Α	\$GATTAGAT	G	ATTACAT\$	Α	T\$GATTAC
T	A\$GATTAGA	G	ATTAGAT\$	Α	T\$GATTAG
T	ACAT\$AGA	G	ATTAGATA\$	A A	TA\$GATTAG
T	ACAT\$GA	Α	CAT\$AGAT	Α	TACAT\$AG
T	ACAT\$GAT	Α	CAT\$GAT	Α	TACAT\$G
T	AGAT\$GAT	Α	CAT\$GATT	T	TACAT\$GA
T	AGATA\$GAT	Α	GAT\$GATT	T	TAGAT\$GA
\$	AGATACAT	Α	GATA\$GATT	T	TAGATA\$GA
C	AT\$AGATA	Α	GATACAT\$	Α	TTACAT\$G
C	AT\$GATA	\$	GATACAT	Α	TTAGAT\$G
C	AT\$GATTA	\$	GATTACAT	Α	TTAGATA\$G

GATTACATS AGATACATS GATACATS GATTAGATS GATTAGATAS

tag	context	tag	context	tag	g context
9	\$AGATACA	7	AT\$GATTA	1	GATTAGAT
9	\$GATACA	7	ATA\$GATTA	1	GATTAGATA
9	\$GATTACA	2	ATACAT\$A	8	T\$AGATAC
9	\$GATTAGA	2	ATACAT\$	8	T\$GATAC
10	\$GATTAGAT	2	ATTACAT\$	8	T\$GATTAC
9	A\$GATTAGA	2	ATTAGAT\$	8	T\$GATTAG
4	ACAT\$AGA	2	ATTAGATA\$	<b>A</b> 8	TA\$GATTAG
4	ACAT\$GA	6	CAT\$AGAT	3	TACAT\$AG
5	ACAT\$GAT	6	CAT\$GAT	3	TACAT\$G
5	AGAT\$GAT	6	CAT\$GATT	4	TACAT\$GA
5	AGATA\$GAT	6	GAT\$GATT	4	TAGAT\$GA
0	AGATACAT	6	GATA\$GATT	4	TAGATA\$GA
7	AT\$AGATA	1	GATACAT\$	3	TTACAT\$G
7	AT\$GATA	1	GATACAT	3	TTAGAT\$G
7	AT\$GATTA	1	GATTACAT	3	TTAGATA\$G



## DNA alignment

WGS reference bias

pangenomes graphs

theorem (sketch)

FM-indexe

data structure counting

MEM-finding

MEM-findin

scaling up

RLCSA

Toehold Lemn r-index

MONI MARIA

data structure

theorem

conclusion

future work thanks to...

questions?

## DNA alignment

WGS reference bias

EDI pangenomes graphs

theorem (sketch)

data structur counting locating

scaling up

RLCSA

Toehold Lemma r-index

### MARIA

examples theorem

conclusion

future work thanks to...

## Theorem

We can store a text T[1..n] over  $\{A, C, G, T, \$\}$  in O(r) space, where r is the number of runs in BWT, such that when given a pattern P[1..m] we can count the occurrences of P in T in O(m) time.

## Corollary

We can store T in O(r+t) space, where t is the numbers of runs in tag, such that when given P we can list all the distinct vertices where we start processing occurrences of P in T, in O(m) time plus constant time per vertex listed.

## ...but no MEMs!

Travis Gagie et al

reference hise

pangenomes graphs

theorem (sketch)

RI CSA

The Past Century Self-Indexing: The Dawn of a New Era A New Challenge: Fully Compressed Self-Indexes

## A New Challenge: Full Compression



## Application Scenario: Computational Biology

- Sequencing genomes is becoming cheap and fast.
- We are not far from the day where we will have databases of thousands or millions of genomes.
- The applications of such a database are unimaginable, BUT
- ▶ 1 million uncompressed genomes ⇒ about 3 petabytes
- ▶ a classical suffix tree ⇒ 30 petabytes
- compressed with current self-indexes \imp 750 terabytes
- ▶ just the sublinear part we mentioned ⇒ 200 terabytes
- Overall, the best we can do requires close to 1 petabyte.

G Navarro

Indexing LZ77

Travis Gagie et al.

DNA alignment

WGS

reference bias

pangenomes graphs

pangenomes grap theorem (sketch)

FM-indexe

1 IVI IIIGCX

data structi counting

MEM-finding

scaling up

RI CSA

RECSA

r-index

#### MARIA

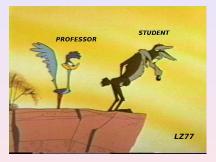
data structure

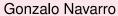
theoren

conclusio

future work thanks to... questions?

# Indexing LZ77: The Next Step in Self-Indexing





Department of Computer Science, University of Chile gnavarro@dcc.uchile.cl





### Travis Gagie et al.

## DNA alignmen

WGS

reference bia

EDI pangenomes graphs

theorem (sketch)

data structure counting locating

MEM-finding scaling up

RLCSA

Toehold Lemm r-index

## MONI

data structure examples

conclusio

future work thanks to... questions?



FM-indexes		I.			
Travis Gagie et al.		fast	fast	fast MEM-	
DNA alignment		counting	locating	finding	scalable
WGS					
reference bias					
EDI .	FM-index	yes	yes	yes	no
pangenomes graphs		yes	yes	yes	110
theorem (sketch)	( <i>JACM</i> , 2005)				
FM-indexes	,				
data structure	RLCSA	yes	no	no	yes
counting	( ICB 2010)	,			,
locating	( <i>JCB</i> , 2010)				
MEM-finding			_		
scaling up	Toehold Lemma	yes	1	no	yes
timeline	(Algorithmica, 2018)	_			
RLCSA	(Algoritimica, 2010)				
Toehold Lemma	, :				
r-index	r-index	yes	yes	no	yes
MONI	( <i>JACM</i> , 2020)				
MARIA	(3/10/11, 2020)				
data structure	MONI	no	yes	yes	yes
examples		110	ycs	ycs	ycs
theorem	( <i>JCB</i> , 2022)				
conclusion	,	I			
future work					
thanks to					
questions?					
					32 / 46
					J2 / TO

Pangenomic

Travis Gagie et al.

## DNA alignmer

WGS reference bias

pangenomes graphs

theorem (sketch)

FM-indexe

data struc counting

MEM-finding

scaling up

timeline RLCSA

Toehold Lemma

r-index

### MARIA

data struct

theore

### conclusion

future work

questions?

## Toehold Lemma

GATTACATS GATACATS GATTAGATS GATTAGATS RW/T context RW/T context **BWT** context Т \$AGATACA G AT\$GATTA \$ GATTAGAT ATA\$GATTA \$ \$GATACA GATTAGATA G \$GATTACA ATACAT\$A T\$AGATAC Α \$GATTAGA G ATACAT\$ Α T\$GATAC SGATTAGAT. G ATTACAT\$ T\$GATTAC Α **A**\$GATTAGA ATTAGAT\$ T\$GATTAG Α ACAT\$AGA ATTAGATA\$ TA\$GATTAG Α CAT\$AGAT TACAT\$AG ACAT\$GA Α ACAT\$GAT CAT\$GAT TACAT\$G AGAT\$GAT CAT\$GATT Т TACAT\$GA AGATA\$GAT GAT\$GATT Т TAGAT\$GA AGATACAT **GATA**\$GATT TAGATA\$GA AT\$AGATA GATACAT\$ Α TTACAT\$G \$ AT\$GATA TTAGAT\$G GATACAT Α Ar\$gatta GATTACAT TTAGATA\$G Α

Travis Gagie et al.

WGS reference hise

pangenomes graphs

theorem (sketch)

MEM-finding

scaling up

Toehold Lemma

## Toehold Lemma

GATTACATS GATACATS GATTAGATS GATTAGATS RW/T context RW/T context **BWT** context Т \$AGATACA G AT\$GATTA \$ GATTAGAT ATA\$GATTA \$ \$GATACA GATTAGATA G \$GATTACA ATACAT\$A T\$AGATAC Α \$GATTAGA G ATACAT\$ Α T\$GATAC SGATTAGAT. G ATTACAT\$ T\$GATTAC Α **A**\$GATTAGA ATTAGAT\$ T\$GATTAG Α ACAT\$AGA ATTAGATA\$ TA\$GATTAG Α CAT\$AGAT TACAT\$AG ACAT\$GA Α ACAT\$GAT CAT\$GAT TACAT\$G AGAT\$GAT CAT\$GATT Т TACAT\$GA AGATA\$GAT GAT\$GATT Т TAGAT\$GA AGATACAT **GATA**\$GATT TAGATA\$GA AT\$AGATA GATACAT\$ Α TTACAT\$G \$ AT\$GATA TTAGAT\$G GATACAT Α Ar\$gatta GATTACAT TTAGATA\$G

Travis Gagie et al.

WGS reference hise

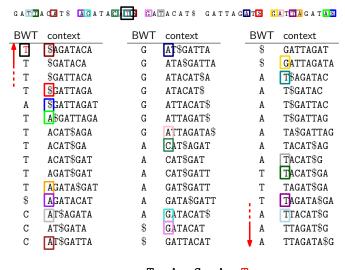
pangenomes graphs

theorem (sketch)

MEM-finding

scaling up

Toehold Lemma



Travis Gagie et al.

WGS reference hise

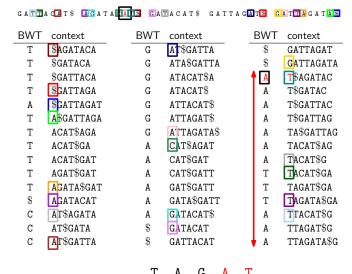
pangenomes graphs

theorem (sketch)

MEM-finding

scaling up

Toehold Lemma



Travis Gagie et al.

## DNA alignmen

WGS reference bias

EDI pangenomes graphs

theorem (sketch)

data struct

locating

MEM-finding scaling up

timeline RLCSA

r-index

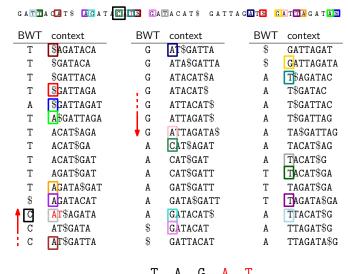
#### MARIA

data structu

LIICOICIII

conclusio

future work thanks to...



Travis Gagie et al.

## DNA alignment

WGS reference bias

EDI pangenomes graphs

theorem (sketch)

FM-indexe

data struct counting

MEM-finding

scaling up

RLCSA
Toehold Lemma

r-index

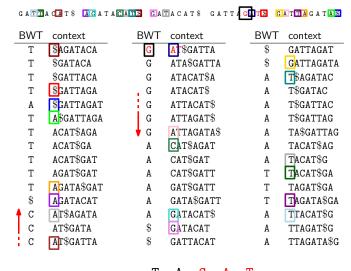
#### MARIA

data structu

theorer

#### conclusio

thanks to...



Travis Gagie et al.

## DNA alignment

WGS reference bias

pangenomes graphs

theorem (sketch)

FM-indexe

data struct counting locating

MEM-finding

timeline RLCSA

Toehold Lemma

MONI

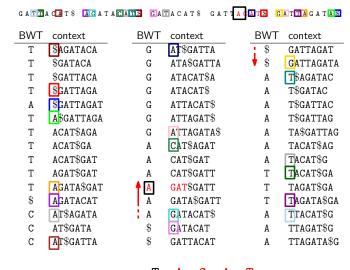
## MARIA

data structu

theorem

conclusio

future work thanks to...



Travis Gagie et al.

## DNA alignment

WGS reference bias

pangenomes graphs

theorem (sketch)

data struct

locating MEM-finding

scaling up

timeline

Toehold Lemma

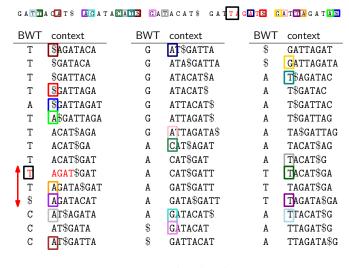
r-index

#### MARIA

data structu examples

conclusio

future work



Travis Gagie et al.

WGS reference hise

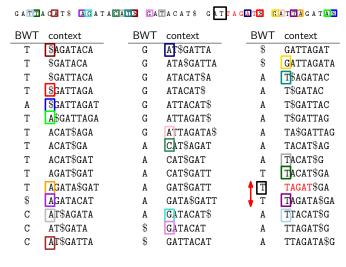
pangenomes graphs

theorem (sketch)

MEM-finding

scaling up

Toehold Lemma



FM-indexes		•		_	
Travis Gagie et al.		fast	fast	fast MEM-	ccalabla
DNA alignment		counting	locating	finding	scalable
WGS reference bias					
EDI pangenomes graphs theorem (sketch)	FM-index ( <i>JACM</i> , 2005)	yes	yes	yes	no
FM-indexes  data structure  counting  locating  MEM-finding	RLCSA ( <i>JCB</i> , 2010)	yes	no	no	yes
scaling up timeline RLCSA	Toehold Lemma ( <i>Algorithmica</i> , 2018)	yes	1	no	yes
Toehold Lemma r-index MONI	r-index ( <i>JACM</i> , 2020)	yes	yes	no	yes
MARIA data structure examples theorem	MONI ( <i>JCB</i> , 2022)	no	yes	yes	yes
conclusion future work thanks to questions?					
					21/16

Pangenomic

Travis Gagie et al.

WGS reference hise

pangenomes graphs

theorem (sketch)

scaling up

r-index

## r-index

## Theorem

We can store a text T[1..n] over  $\{A, C, G, T, \$\}$  in O(r) space such that when given a pattern P[1..m], we can

- **1)** count P's occurrences in T in O(m) time
- 2 list those occurrences' positions in T in constant time per occurrence.

...but still no MEMs!

FM-indexes		ı		_	
Travis Gagie				fast	
et al.		fast	fast	MEM-	
DNA alignment		counting	locating	finding	scalable
WGS					
reference bias					
EDI	FM-index	yes	yes	yes	no
pangenomes graphs		ycs	ycs	ycs	110
theorem (sketch)	( <i>JACM</i> , 2005)				
FM-indexes	(= - ,)				
data structure	RLCSA	yes	no	no	yes
counting		yes	110	110	ycs
locating	( <i>JCB</i> , 2010)				
MEM-finding	( , ,				
scaling up	Toehold Lemma	yes	1	no	yes
timeline		,	_		,
RLCSA	(Algorithmica, 2018)				
Toehold Lemma					
r-index	r-index	yes	yes	no	yes
MONI	( IACM 2020)		,		J
MARIA	( <i>JACM</i> , 2020)				
data structure	MONI	no	yes	yes	yes
examples			<i>y</i> <b>55</b>	700	700
theorem	( <i>JCB</i> , 2022)				
conclusion		Į.			
future work					
thanks to					
questions?					
					36 / 46

Pangenomic

Travis Gagie et al.

## DNA alignmen

WGS reference bias

pangenomes graphs

theorem (sketch)

FM-indexe

data structure

MEM-finding

scaling up

timeline RLCSA

RLCSA Toehold Lei

r-index

MONI

## MARIA

data struct examples

theorem

future work thanks to... MONI
GATTACATS GATACATS GATTAGATS GATTAGATS

BWT	context	BWT	context	BWT	context
T	\$AGATACA	G	AT\$GATTA	-\$	GATTAGAT
T	\$GATACA	G	ATA\$GATTA	\$	GATTAGATA
T	\$GATTACA	G	ATACAT\$A	Α	T\$AGATAC
T	\$GATTAGA	G	ATACAT\$	Α	T\$GATAC
Α	\$GATTAGAT	G	ATTACAT\$	Α	T\$GATTAC
T	A\$GATTAGA	G	ATTAGAT\$	Α	T\$GATTAG
T	ACAT\$AGA	G	A'TTAGATA\$	Α	TA\$GATTAG
T	ACAT\$GA	Α	CAT\$AGAT	Α	TACAT\$AG
T	ACAT\$GAT	Α	CAT\$GAT	Α	TACAT\$G
T	AGAT\$GAT	Α	CAT\$GATT	T	TACAT\$GA
T	AGATA\$GAT	Α	GAT\$GATT	T	TAGAT\$GA
\$	AGATACAT	Α	GATA\$GATT	T	TAGATA\$GA
C	A T\$AGATA	Α	GATACAT\$	Α	T TACAT\$G
C	AT\$GATA	\$	GATACAT	Α	TTAGAT\$G
C	AT\$GATTA	\$	GATTACAT	Α	${\tt TTAGATA\$G}$

Travis Gagie et al.

## DNA alignmen

WGS reference bias

pangenomes graphs

theorem (sketch)

FM-indexe

data structure counting

locating MEM-finding

MEM-findin

scaling up timeline RLCSA

Toehold Lemma r-index

MONI

### MARI

data structure

theorem

conclusion

future work thanks to...

# MONI

GATT	A O T \$	AGATACA	TS GA	TACAT\$	GATTAG	ATS G	ATTAGAT <mark>AS</mark>
BWT	conte	ĸt	BWT	context		BWT	context
T	\$AGA7	ACA	G	AT\$GAT	ГА	\$	GATTAGAT
T	\$GATA	CA	G	ATA\$GA	ГТА	\$	GATTAGATA
T	\$GATT	CACA	G	ATACAT	\$A	Α	T\$AGATAC
T	\$GAT7	AGA	G	ATACAT	\$	Α	T\$GATAC
Α	\$GAT7	AGAT	G	ATTACA	Т\$	Α	T\$GATTAC
T	A\$GA7	TAGA	G	ATTAGA	Г\$	Α	T\$GATTAG
T	ACATS	BAGA	G	A'TTAGA'	ΓΑ\$	Α	TA\$GATTAG
T	ACATS	GA	Α	CAT\$AG	AT	Α	TACAT\$AG
Т	ACATS	GAT	Α	CAT\$GA	Γ	Α	TACAT\$G
T	AGATS	GAT	Α	CAT\$GA	ГТ	T	TACAT\$GA
Т	AGATA	\$GAT	Α	GAT\$GA	ГТ	T	TAGAT\$GA
\$	AGATA	CAT	Α	GATA\$G	ATT	T	TAGATA\$GA
C	АГ\$АС	ATA	Α	GATACA'	Г\$	Α	TTACAT\$G
C	AT\$GA	TA	\$	GATACA'	Γ	Α	TTAGAT\$G
C	AT\$GA	ATTA	\$	GATTAC	AT	Α	TTAGATA\$G
	_						

G A T A G A T

Travis Gagie et al.

WGS reference hise

pangenomes graphs

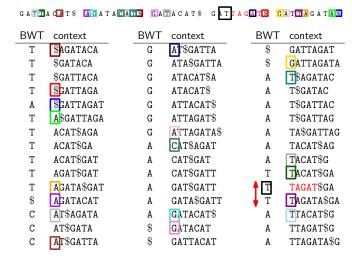
theorem (sketch)

MEM-finding

scaling up

### MONI

# MONI



Travis Gagie et al.

## DNA alignmer

WGS reference bias

pangenomes graphs

theorem (sketch)

FM-indexe

data struct

MEM-finding

scaling up

timeline RLCSA

Toehold Lem

MONI

## WARIA

data struct examples

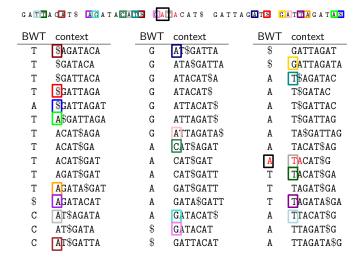
theore

conclusio

future work thanks to...

questions

# MONI



Travis Gagie et al.

## DNA alignmen

WGS reference bias

pangenomes graphs

theorem (sketch)

FM-indexe

data struct

locating MEM-finding

scaling up

timeline

RLCSA

r-index

### MONI

## data struct

examples

theorer

conclusio

future work thanks to...

questions

# MONI

GATTACATS GATTAGATS GATTAGATS GATTAGATS RW/T context RW/T context **BWT** context Т \$AGATACA G AT\$GATTA \$ GATTAGAT ATA\$GATTA \$ \$GATACA G GATTAGATA \$GATTACA ATACAT\$A T\$AGATAC Α \$GATTAGA G ATACAT\$ Α T\$GATAC SGATTAGAT ATTACAT\$ T\$GATTAC Α **A**\$GATTAGA ATTAGAT\$ T\$GATTAG Α ACAT\$AGA ATTAGATA\$ TA\$GATTAG Α ACAT\$GA

CAT\$AGAT CAT\$GAT CAT\$GATT GAT\$GATT **GATA**\$GATT GATACAT\$ \$ GATACAT GATTACAT

ACAT\$GAT

AGAT\$GAT

AGATA\$GAT

AGATACAT

AT\$AGATA

AT\$GATA

Ar\$gatta

A TSGATAC
A TSGATTAC
A TSGATTAG
A TASGATTAG
A TACATSAG
A TACATSG
T TACATSG
T TAGATSGA
T TAGATSGA
A TTAGATSG
A TTAGATSG
A TTAGATSG
A TTAGATSG
A TTAGATSG

A T A G A T 3 5 4 3 2 1

Travis Gagie et al.

## DNA alignment

WGS reference bias

EDI

pangenomes graphs theorem (sketch)

theorem (sketch

data structure

locating MEM-finding

MEM-finding scaling up

timeline

Toehold Lem

MONI

### MARI

data structi examples

theore

COTICIUSIO

future work thanks to...

questions'

# MONI

GATTA CATS AGATACATS GATACATS GATTAGATS CATTAGATAS

BWT	context	BWT	context	BWT	context
T	\$AGATACA	G	AT\$GATTA	\$	GATTAGAT
T	\$GATACA	G	ATA\$GATTA	\$	GATTAGATA
T	\$GATTACA	G	ATACAT\$A	Α	T\$AGATAC
T	\$GATTAGA	G	ATACAT\$	Α	T\$GATAC
Α	\$GATTAGAT	G	ATTACAT\$	Α	T\$GATTAC
T	<b>A</b> \$GATTAGA	G	ATTAGAT\$	Α	T\$GATTAG
T	ACAT\$AGA	G	A'TTAGATA\$	Α	TA\$GATTAG
T	ACAT\$GA	Α	CAT\$AGAT	Α	TACAT\$AG
T	ACAT\$GAT	Α	CAT\$GAT	Α	TACAT\$G
T	AGAT\$GAT	Α	CAT\$GATT	T	TACAT\$GA
T	AGATA\$GAT	Α	GAT\$GATT	T	TAGAT\$GA
\$	AGATACAT	Α	GATA\$GATT	T	TAGATA\$GA
C	AT\$AGATA	Α	GATACAT\$	Α	TTACAT\$G
C	AT\$GATA	\$	GATACAT	Α	TTAGAT\$G
C	Ar\$gatta	\$	GATTACAT	Α	TTAGATA\$G
	_				

4 3 5 4 3 2 3

## DNA alignment

WGS reference bias EDI pangenomes graphs

theorem (sketch)

FM-indexes

data structur counting locating

MEM-finding Scaling up

timeline RLCSA

r-index MONI

## MARI

data structu examples theorem

#### conclusion

future work thanks to...

## Theorem

We can store a text T[1..n] over  $\{A, C, G, T, \$\}$  in O(r + |LCE|) space, where |LCE| is the space for an LCE data structure, such that when given a pattern P[1..m] we can

- **1** find the MEMs of P with respect to T in  $O(m \log n)$  time
- 2 list their occurrences' positions in T in constant time per occurrence.

# Corollary

We can store T in  $O(r + t + |\mathsf{LCE}|)$  space such that when given P, we can list all the distinct **vertices** where we start processing occurrences of P in T, in  $O(m \log n)$  time plus constant time per **occurrence**.

Can we find the interval directly? DO WE NEED IT?

Travis Gagie et al.

WGS reference hise

pangenomes graphs theorem (sketch)

data structure MEM-finding

### scaling up

RLCSA.

r-index

### MARIA

# **MARIA**

(unpublished joint work with Andrej Baláž, Adrián Goga and Alessia Petescia)

Travis Gagie et al.

#### DNA alignmen

WGS reference bias

EDI pangenomes graphs

theorem (sketch)

data structure counting

locating MEM-finding

MEM-finding scaling up

timeline RLCSA

Toehold Lemr

r-index

#### MARIA

#### data structure

example

a a made at a

future work thanks to...

thanks to... questions?

# data structure

CATTACATS AGATACATS GATACATS GATTAGATS GATTAGATAS

tag	context	:	tag	context		tag	context	:
9	\$AGATA	CA	7	AT\$GATTA		1	GATTAC	FAT
9	\$GATAC	CA	7	AFA\$GATTA	L	1	GATTAC	ATA
9	\$GATTA	CA	2	A FACAT\$A		8	T\$AGAT	CAC
9	\$GATTA	IGA	2	ATACAT\$		8	T\$GATA	C
10	\$GATTA	GAT	2	ATTACAT\$		8	T\$GATT	CAC
9	A GATT	AGA	2	ATTAGAT\$		8	T\$GATT	AG
4	ACAT\$	AGA	2	A TTAGATAS	;	8	TA\$GAT	TAG
4	ACAT\$C	ξA	6	CAT\$AGAT		3	TACATS	GAG
5	ACAT\$C	TA	6	CAT\$GAT		3	TACATS	3G
5	AGAT\$0	TA	6	CAT\$GATT		4	TACATS	GA
5	AGATAS	GAT	6	GAT\$GATT		4	TAGATS	GA
0	AGATAC	CAT	6	GATA\$GATT	•	4	TAGATA	A\$GA
7	A'T\$AGA	ATA	1	GATACAT\$		3	TTACAT	Γ\$G
7	AT\$GAT	<b>.</b> A	1	GATACAT		3	TTAGAT	T\$G
7	AT\$GAT	TA	1	GATTACAT		3	TTAGAT	TA\$G
٨	C	٨	т	T ,	С	٨	т	٨

Travis Gagie et al.

WGS reference hise pangenomes graphs

theorem (sketch)

data structure

MEM-finding

scaling up

RLCSA.

r-index

examples

# examples

CATTACATS AGATACATS GATTAGATS GATTAGATS

tag	contex	t	tag	g co	ontext		tag	contex	t
9	\$AGAT.	ACA	7	A'	Γ\$GATTA		1	GATTA	GAT
9	\$GATA	CA	7	Α	ΓA\$GATT	A	1	GATTA	GATA
9	\$GATT.	ACA	2	Α	ГАСАТ\$А		8	T\$AGA	TAC
9	\$GATT.	AGA	2	A.	ГАСАТ\$		8	T\$GAT.	AC
10	\$GATT.	AGAT	2	A.	TTACAT\$		8	T\$GAT	TAC
9	A GAT	ΓAGA	2	A.	TTAGAT\$		8	T\$GAT	TAG
4	ACAT\$.	AGA	2	Α	TTAGATA	\$	8	TA\$GA	TTAG
4	ACAT\$	GA	6	C	AT\$AGAT		3	TACAT	\$AG
5	ACAT\$	GAT	6	C	AT\$GAT		3	TACAT	\$G
5	AGAT\$	GAT	6	C	AT\$GATT		4	TACAT	\$GA
5	AGATA	\$GAT	6	G	AT\$GATT		4	TAGAT	\$GA
0	AGATA	CAT	6	G	ATA\$GAT	Τ	4	TAGAT.	A\$GA
7	AT\$AG	ATA	1	G	ATACAT\$		3	TTACA	T\$G
7	AT\$GA	ΓΑ	1	G.	TACAT		3	TTAGA	T\$G
7	AT\$GA	ГТА	1	G	ATTACAT		3	TTAGA	TA\$G
٨	C	۸	т	Τ	٨	С	٨	т	٨

Travis Gagie et al.

WGS reference hise

pangenomes graphs

theorem (sketch)

MEM-finding

scaling up

examples

# examples



ATA\$GATTA \$GATACA GATTAGATA 9 \$GATTACA A LACALSA 8 T\$AGATAC g \$GATTAGA ATACAT\$ 8 T\$GATAC 10 SGATTAGAT ATTACAT\$ T\$GATTAC g **A**GATTAGA ATTAGAT\$ 8 T\$GATTAG ACAT\$AGA ATTAGATA\$ TA\$GATTAG TACAT\$AG ACAT\$GA 6 CAT\$AGAT 3 4

ACAT\$GAT CAT\$GAT AGAT\$GAT CAT\$GATT

AGATASGAT GAT\$GATT AGATACAT

AT\$AGATA AT\$GATA

5

AT\$GATTA

GATA\$GATT

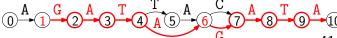
GATACATS

**GATACAT** GATTACAT TAGAT\$GA TAGATA\$GA TTACAT\$G

3 TTAGAT\$G TTAGATA\$G

TACAT\$G

TACAT\$GA



Travis Gagie et al.

WGS reference hise

pangenomes graphs

theorem (sketch)

data structure

MEM-finding

scaling up

RLCSA.

r-index

examples

# examples



tag	contex	t	tag	context		tag	context	t
9	\$AGAT	ACA	7	AT\$GATTA	_	1	GATTAC	GAT
9	\$GATA	CA	7	AFA\$GATTA		1	GATTA(	GATA
9	\$GATT	ACA	2	A FACAT\$A		8	T\$AGA7	TAC
9	\$GATT	AGA	2	ATACAT\$		8	T\$GATA	AC.
10	\$GATT.	AGAT	2	ATTACAT\$		8	T\$GAT7	TAC
9	A <sup>B</sup> GAT	TAGA	2	ATTAGAT\$		8	T\$GAT7	TAG
4	ACAT\$	AGA	2	Artagata\$		8	TA\$GA7	TAG
4	ACAT\$	GA	6	CAT\$AGAT		3	TACATS	BAG
5	ACAT\$	GAT	6	CAT\$GAT		3	TACAT	3G
5	AGAT\$	GAT	6	CAT\$GATT		4	TACATS	6GA
5	AGATA	\$GAT	6	GAT\$GATT		4	TAGAT	6GA
0	AGATA	CAT	6	GATA\$GATT		4	TAGATA	A\$GA
7	A'Γ\$AG	ATA	1	GATACAT\$		3	TTACAT	Γ\$G
7	AT\$GA	TA	1	GATACAT		3	TTAGAT	Γ\$G
7	AT\$GA	TTA	1	GATTACAT		3	TTAGAT	ΓA\$G
۸	α	٨	T	Т	C	٨	T	۸

Travis Gagie et al.

WGS reference hise

pangenomes graphs theorem (sketch)

MEM-finding scaling up

RLCSA.

r-index

examples

# examples

CATTACATS AGATACATS GATACATS GATTAGATAS

	tag	context		tag	cor	ntext		tag	context	t
_	9	\$AGATAC	A	7	AT	\$GATTA	_	1	GATTA	GAT
	9	\$GATACA		7	АΓ	A\$GATT <i>A</i>	A	1	GATTA	GATA
	9	\$GATTACA	A	2	АΓ	ACAT\$A		8	T\$AGAT	ΓAC
	9	\$GATTAGA	A	2	AT.	ACAT\$		8	T\$GATA	AC
	10	\$GATTAGA	AΤ	2	AT'	TACAT\$		8	T\$GAT	ΓAC
	9	A GATTA	GA	2	AT'	TAGAT\$		8	T\$GAT	ΓAG
	4	ACAT\$AGA	A	2	АΓ	TAGATAS	\$	8	TA\$GAT	ΓTAG
	4	ACAT\$GA		6	CA'	T\$AGAT		3	TACATS	BAG
	5	ACAT\$GAT	Γ	6	CA'	T\$GAT		3	TACAT	₿G
	5	AGAT\$GAT	Γ	6	CA'	T\$GATT		4	TACATS	BGA
	5	AGATA\$GA	AΤ	6	GA'	T\$GATT		4	TAGAT	BGA
	0	AGATACA:	Γ	6	GA'	TA\$GAT7	Γ	4	TAGATA	A\$GA
	7	AT\$AGAT	A	1	GA'	TACAT\$		3	TTACAT	Г\$G
	7	AT\$GATA		1	GA'	TACAT		3	TTAGA	Г\$G
	7	AT\$GATT	A	1	GA'	TTACAT		3	TTAGA	ΓA\$G
		~		_	Т		C		_	

Travis Gagie et al.

WGS reference hise

pangenomes graphs theorem (sketch)

data structure

MEM-finding

scaling up

RLCSA.

r-index

examples

# examples

CATTACATS AGATACATS GATACATS GATTAGATAS

tag	context		tag	context		tag	context	
9	\$AGATA	CA	7	AT\$GATTA		1	GATTAG	AT
9	\$GATAC	A	7	AFA\$GATTA		1	GATTAG	ATA
9	\$GATTA	CA	2	A FACAT\$A		8	T\$AGAT	AC
9	\$GATTA	GA	2	ATACAT\$		8	T\$GATA	C
10	\$GATTA	GAT	2	ATTACAT\$		8	T\$GATT	AC
9	ABGATT.	AGA	2	ATTAGAT\$		8	T\$GATT	AG
4	ACAT\$A	GA	2	ATTAGATA\$		8	TA\$GAT	TAG
4	ACAT\$G	A	6	CAT\$AGAT		3	TACAT\$	AG
5	ACAT\$G	AT	6	CAT\$GAT		3	TACAT\$	G
5	AGAT\$G	AT	6	CAT\$GATT		4	TACAT\$	GA
5	AGATA\$	GAT	6	GAT\$GATT		4	TAGAT\$	GA
0	AGATAC.	AT	6	GATA\$GATT		4	TAGATA	\$GA
7	A'T\$AGA'	ΓΑ	1	GATACAT\$		3	TTACAT	\$G
7	AT\$GAT	A	1	GATACAT		3	TTAGAT	\$G
7	AT\$GAT	ΓΑ	1	GATTACAT		3	TTAGAT	A\$G
	а	٨	T	Т	C		T	۸

Travis Gagie et al.

WGS reference hise

pangenomes graphs

theorem (sketch)

data structure

MEM-finding

scaling up

RLCSA.

r-index

examples

# examples

CATTACATS AGATACATS GATACATS GATTAGATAS

	tag	context	tag	context		tag	context
-	9	\$AGATACA	7	AT\$GATTA	_	1	GATTAGAT
	9	\$GATACA	7	AFA\$GATTA		1	GATTAGATA
	9	\$GATTACA	2	A FACAT\$A		8	T\$AGATAC
	9	\$GATTAGA	2	ATACAT\$		8	T\$GATAC
	10	\$GATTAGAT	2	ATTACAT\$		8	T\$GATTAC
	9	A BGATTAGA	. 2	ATTAGAT\$		8	T\$GATTAG
	4	ACAT\$AGA	2	A TTAGATA\$		8	TA\$GATTAG
	4	ACAT\$GA	6	CAT\$AGAT		3	TACAT\$AG
	5	ACAT\$GAT	6	CAT\$GAT		3	TACAT\$G
	5	AGAT\$GAT	6	CAT\$GATT		4	TACAT\$GA
	5	AGATA\$GAT	6	GAT\$GATT		4	TAGAT\$GA
	0	AGATACAT	6	GATA\$GATT		4	TAGATA\$GA
	7	A'T\$AGATA	1	GATACAT\$		3	TTACAT\$G
	7	AT\$GATA	1	GATACAT		3	TTAGAT\$G
	7	AT\$GATTA	1	GATTACAT		3	TTAGATA\$G
		~ .	_	т.	C		



## WGS alignment

reference bias EDI pangenomes graphs theorem (sketch)

data structur

locating MEM-finding

timeline RLCSA Toehold Lemma

### MARIA

examples

theorem

future work thanks to...

## **Theorem**

We can store a text T[1..n] in  $O(r + t + |\mathsf{LCE}|)$  space such that when given a pattern P[1..m], we can

- **1** find the MEMs of P with respect to T in  $O(m \log n)$  time
- 2 list all the distinct vertices where we start processing occurrences in T of each MEM, in constant time per vertex listed.

## Pangenomic FM-indexes Travis Gagie

# et al.

#### DNA alignment

WGS reference bias

pangenomes graphs

## theorem (sketch)

#### FM-index

data structure counting locating MEM-finding

#### scaling up

timeline

RLCSA

r-index

data structure

theorem

#### conclusion

future work thanks to...

# conclusion

reference hise

theorem (sketch)

scaling up

future work

With MONI and MARIA, we have a pipeline for our goal:

- 1 index the dataset as a set of strings losslessly without adding or excluding any variations
- 2 find good seeds in the reads exactly matching substrings of the dataset
- 3 map those seeds **directly** onto walks in the graph\* in time independent of the number of matches in the dataset (which we don't even compute).

When MARIA is finished, someone (else!) should still extend the MEMs to approximate matches for the reads, build a consensus sequence, etc.

<sup>\*</sup>The seeds needn't match the walks exactly! We can simplify the graph — as a coordinate system — without weakening the index.

Travis Gagie et al.

### DNA alignment

WGS reference bias

pangenomes graphs

theorem (sketch)

data struct

locating
MEM-finding

scaling up

timeline

Toehold Lemn

r-index

#### MARIA

data structu examples

conclucio

future work thanks to...

thanks to... questions?

Omar Ahmed Jarno Alanko Andrej Baláž Hideo Bannai Paola Bonizzoni Christina Boucher Nate Brown Nicola Cotumaccio Adrián Goga Simon Heumos Tomohiro I Dominik Köppl Alan Kuhnle Ben Langmead Veli Mäkinen Giovanni Manzini Taher Mun Gonzalo Navarro Alessia Petescia Nicola Prezza Massimiliano Rossi Jouni Sirén

# thanks to...













## Pangenomic FM-indexes Travis Gagie

# et al.

#### DNA alignment

WGS reference bias

reference b

pangenomes graphs

## theorem (sketch)

#### FM-index

data structure counting locating

## MEM-finding

scaling up

RLCSA

Toehold Lemn

r-index

#### MADIA

data structure

examples

#### conclusion

future work

questions?

# questions?