UNIVERSITÀ
di VERONA

On different variants of the Burrows-Wheeler-Transform of string collections

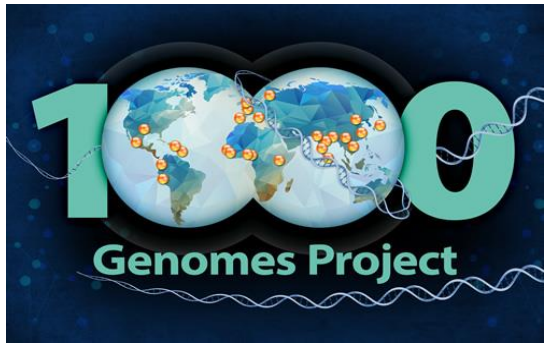
Davide Cenzato and Zsuzsanna Lipták

University of Verona, Department of Computer Science.

DCC 2022, March 22 – 25, 2022 - Snowbird, Utah, US

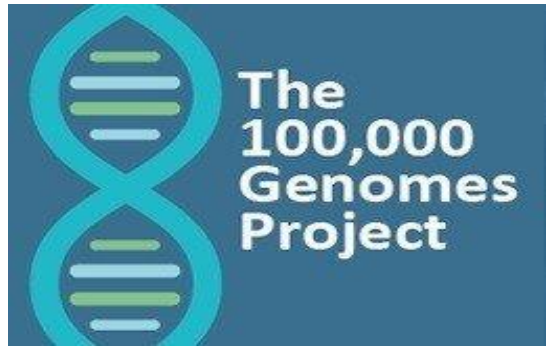
Large string collections are highly abundant

- number of sequenced genomes is growing at unprecedented pace
- focus has moved from single strings to string collections



2008

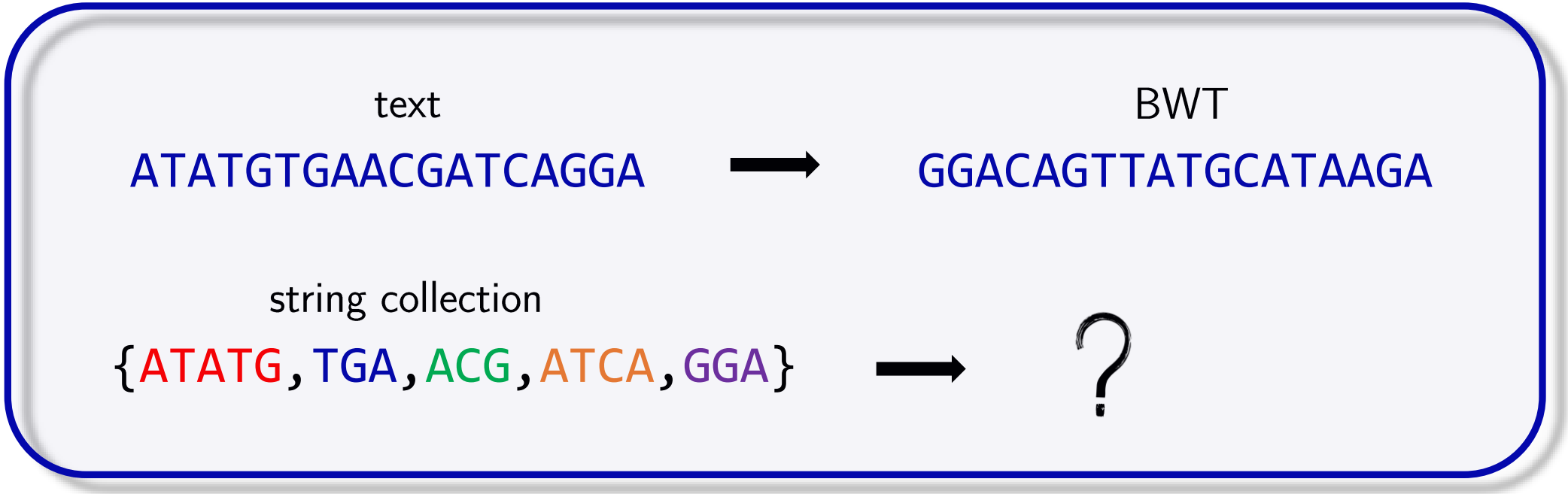
2009



2013

The Burrows-Wheeler-Transform for string collections

- originally defined on single strings
- not all tools compute the same transform



BWT variants in the literature

We identified **5 different BWT variants** computed by different tools available.

		e.g. tool	
separator based			
dolEBWT($\mathcal{M}$)	GGAAACGG$$$TTACTGTAAA	G2BWT	
mdolBWT($\mathcal{M}$)	GAGAAGCG$$$TTATCTGAAA	BCR	(do use dollars)
concBWT($\mathcal{M}$)	AAGAGGGC$$$TTACTGTAAA	Big-BWT	
colexBWT($\mathcal{M}$)	AAAGGCGG$$$TTACTGTAAA	ropebwt2	
non-sep. based			
eBWT($\mathcal{M}$) (Mantaci et al.)	CGGGATGTACGTTAAAAA	pfpebwt	(does not use dollars)

all BWT variants of the string collection
$\mathcal{M} = \{\texttt{ATATG}, \texttt{TGA}, \texttt{ACG}, \texttt{ATCA}, \texttt{GGA}\}$

Interesting intervals

Each variant uses a **different order** for the characters in the interesting intervals.

BWT variant	**criteria** for characters ordering
eBWT($\mathcal{M}$)	omega-order of strings
dolEBWT($\mathcal{M}$)	lexicographic order of strings
mdolBWT($\mathcal{M}$)	input order of strings
concBWT($\mathcal{M}$)	lexicographic order of subsequent strings in input
colexBWT($\mathcal{M}$)	colexicographic order ("reverse lex. order")

- depends on the **input order:**

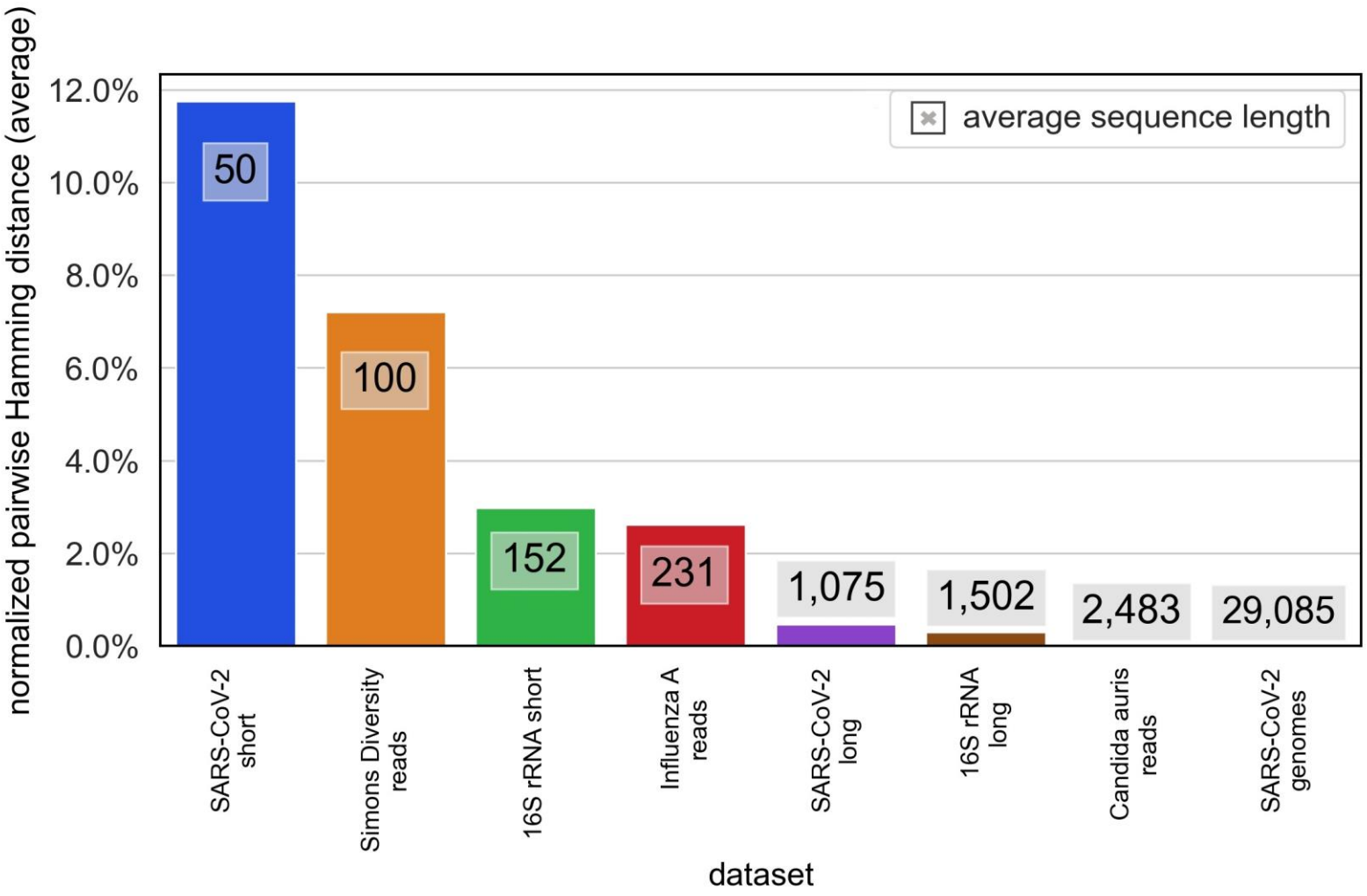
$\text{mdolBWT}(\mathcal{M}_1) =$ GAGAAGCG$$$TTATCTGAAA $\quad \mathcal{M}_1 = \{\texttt{ATATG}, \texttt{TGA}, \texttt{ACG}, \texttt{ATCA}, \texttt{GGA}\}$

$\text{mdolBWT}(\mathcal{M}_2) =$ GGAAAGGC$$$TTACTGTAAA $\quad \mathcal{M}_2 = \{\texttt{ACG}, \texttt{ATATG}, \texttt{GGA}, \texttt{TGA}, \texttt{ATCA}\}$

Experimental results 1: Hamming distance

We conducted experiments on 8 real-life datasets to determine how much these differences matter.

- strongly depends on **sequence length**
- SARS-Cov-2 short: **500,000** sequences of length **50**
- on average **almost 12%** different positions



Experimental results 2: number of runs

These differences extend to the number of runs (r) of all BWT variants.

- **number of runs** is more variable on short sequence datasets
- **average runlength** (n/r) is an equivalent measure

no. runs **SARSCov2short** dataset		
	r	n/r
eBWT	1,902,148	13.143
doIEBWT	1,868,581	13.647
mdoIBWT	3,113,818	8.189
concBWT	**3,402,513**	7.494
colexBWT	**808,906**	31.524

Summary

Different tools compute different BWT variants

- identified **five** different BWT variants computed by these tools
- some of these are sensitive to **input order**

Differences between BWT variants are not negligible

- more relevant on **short sequences**
- extends to the **number of runs** of the BWT variants

UNIVERSITÀ
di VERONA

Thank you for your attention

contact: davide.cenzato@univr.it

GitHub: https://github.com/davidecenzato/BWT-variants-of-string-collections

arXiv: http://arxiv.org/abs/2202.13235