

Fast estimation of pangenome openness using prefix-free parsing

MIGGS - Novel Pangenome Paradigms
JOBIM 2025, July 10, Bordeaux

Igor MARTAYAN

 @imartayan

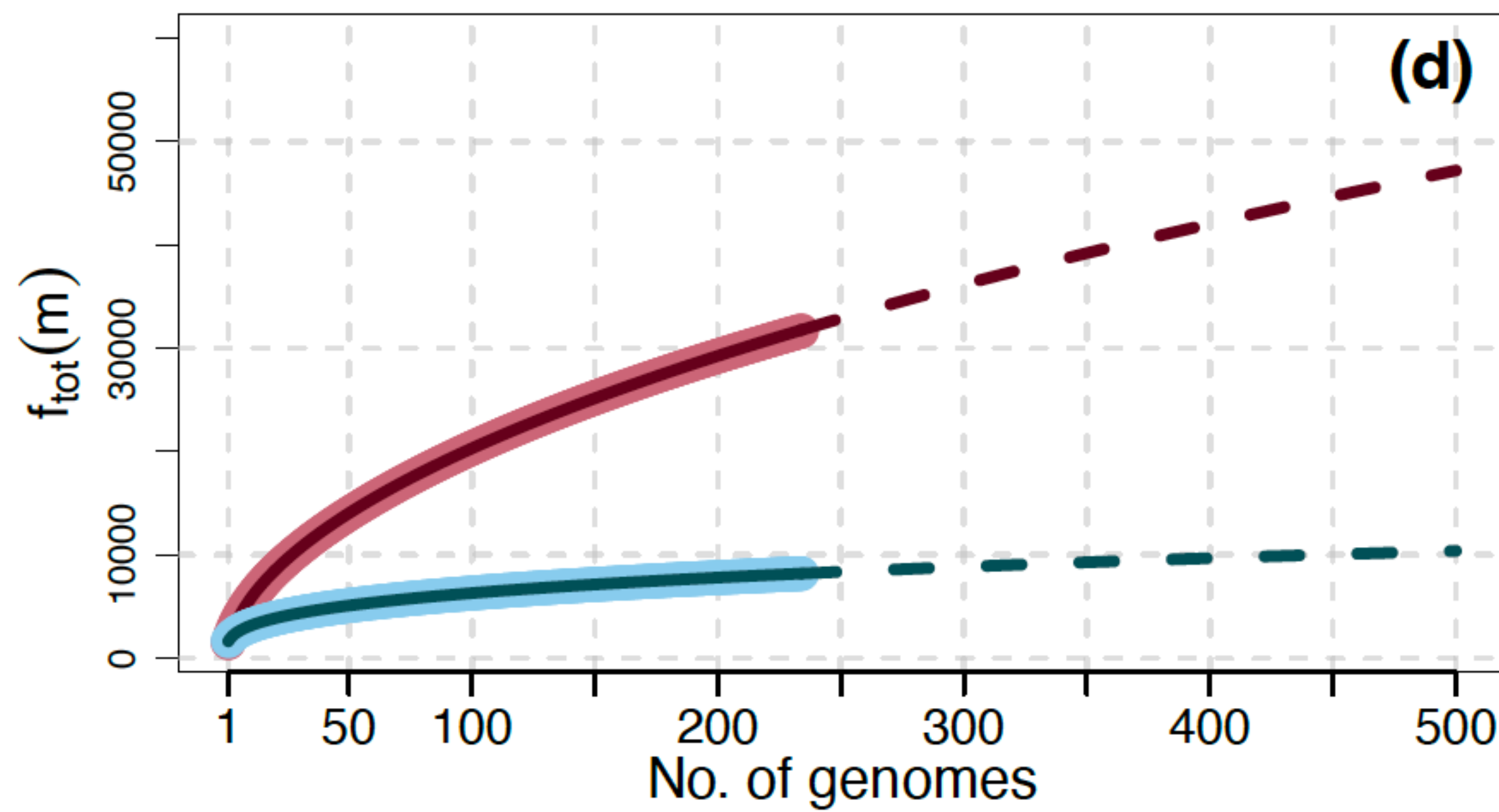


Pangenome Openness

$f(n)$ = # new genes added by the n^{th} genome $\sim k n^{-a}$ (Heap's law)

$a < 1 \Rightarrow$ open

$a > 1 \Rightarrow$ closed



Estimated # genes of *H. pylori* and *C. jejuni*,
from [Parmigiani Wittler Stoye, 2024]

Common approach expensive to compute
(every genome ordering) [Tettelin et al., 2005]

Faster histogram-based approach for k -mers
in $O(n^2)$ [Parmigiani Wittler Stoye, 2024]

**Genome analysis of multiple pathogenic
isolates of *Streptococcus agalactiae*:
Implications for the microbial “pan-genome”**

Hervé Tettelin, Vega Massignani, Michael J. Cieslewicz, [+42](#), and Claire M. Fraser [Authors Info & Affiliations](#)

PNAS, 2005

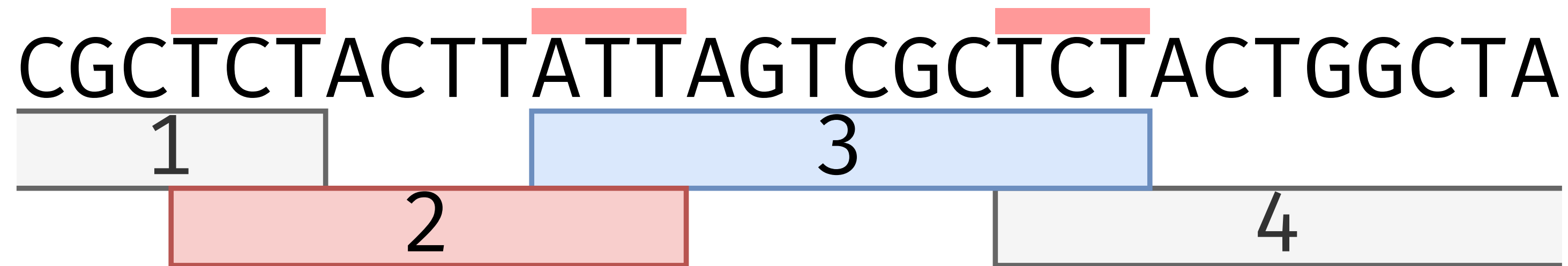
Revisiting pangenome openness with k -mers

Parmigiani, Luca^{1, 2, 3} [ID](#) ; Wittler, Roland^{1, 2} [ID](#) ; Stoye, Jens^{1, 2} [ID](#) [Home](#)

PCJ, 2024

Prefix-Free Parsing (PFP)

- Select windows with $hash \bmod p = 0$ (p acts as a sampling parameter)
- Selected windows partition the text into *phrases*



Local consistency:
same context $\Rightarrow$ same selection $\Rightarrow$ same phrases

The # phrases is a good metric for repetitiveness

Prefix-free parsing for building big BWTs

Christina Boucher [✉](#), Travis Gagie, Alan Kuhnle, Ben Langmead, Giovanni Manzini & Taher Mun

AMB, May 2019

Pfp-fm: an accelerated FM-index

Aaron Hong¹, Marco Oliva¹, Dominik Köppl^{2,3}, Hideo Bannai³, Christina Boucher^{1*} and Travis Gagie⁴

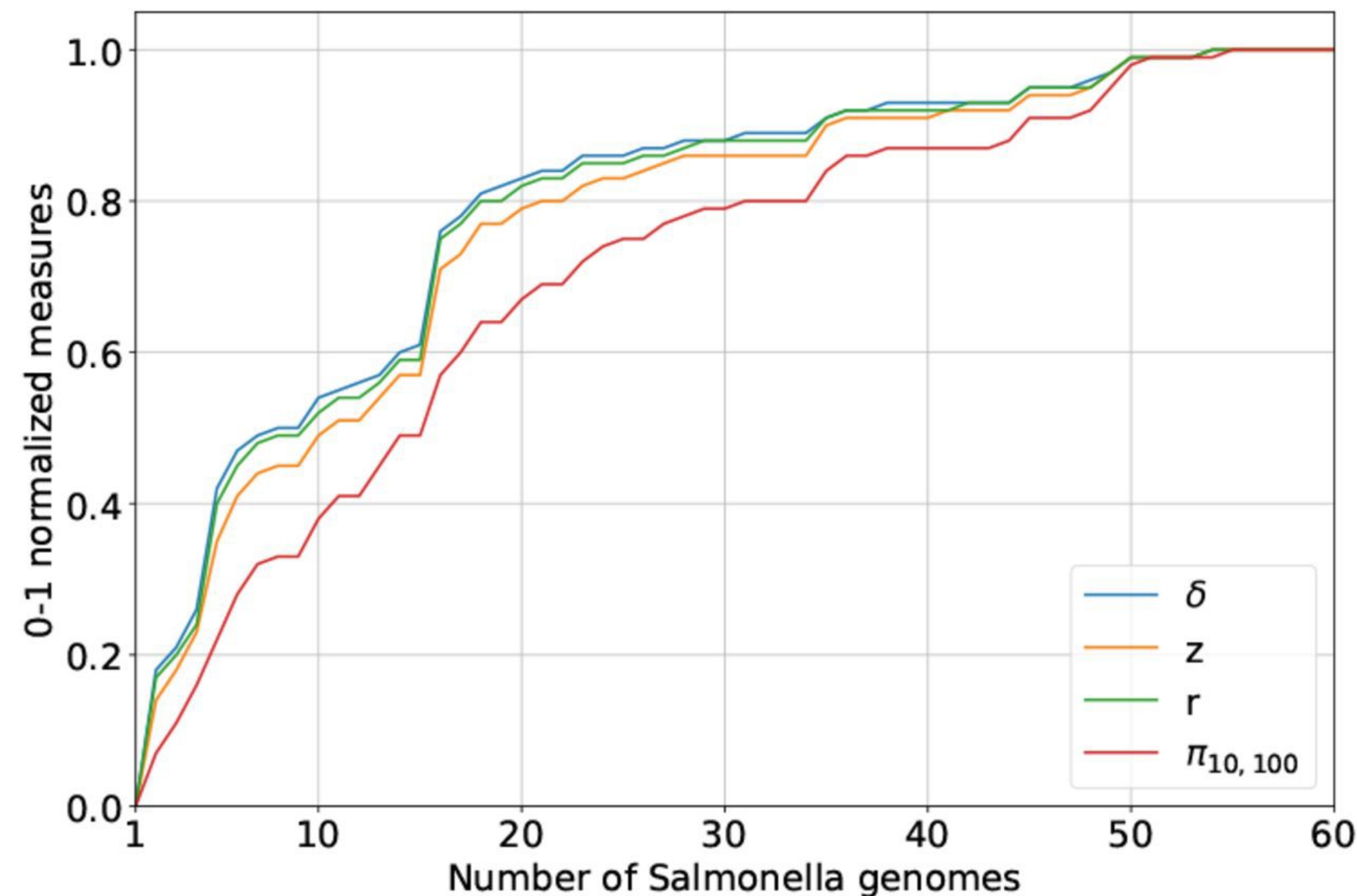
AMB, April 2024

Prefix-free parsing for merging big BWTs

Diego Diaz-Dominguez, Travis Gagie, Veronica Guerrini, Ben Langmead, Zsuzsanna Liptak, Giovanni Manzini, Francesco Masillo, Vikram Shivakumar

Preprint, June 2025

Estimating Openness with PFP



Normalized growth pattern for
 k -mers / LZ factors / BWT runs / PFP,
from [Lipták Lucà Masillo, 2025]

Openness can be estimated using
PFP phrases
[Lipták Lucà Masillo, 2025]

Benefit: PFP space is much
sparser than k -mer space

Measuring Genomic Data with PFP

Zsuzsanna Lipták, Simone Lucà, Francesco Masillo

Preprint, Feb. 2025

github.com/simolucaa/piPFP (original)

github.com/imartayan/pfp (WIP, 5-10x faster)

Take-home messages

- PFP is a simple and high-throughput parsing method
- Provides a good estimation of pangenome openness (a such that # new genes $f(n) \sim k n^{-a}$)
- Can be reused for lightweight indexing (ongoing work)

Thank you!

