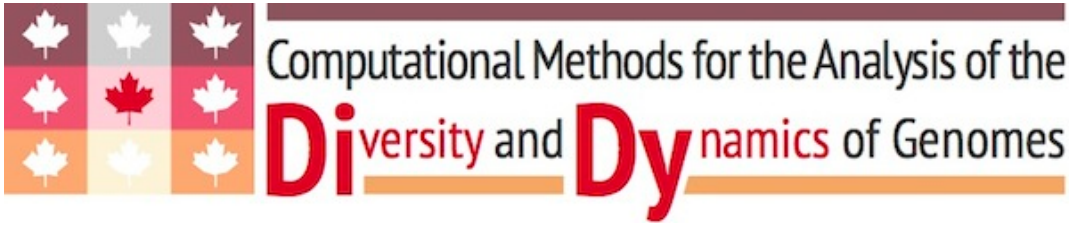


DARRC

DYNAMIC ALIGNMENT-FREE AND REFERENCE-FREE READ COMPRESSION

Guillaume Holley, Roland Wittler, Jens Stoye [1]
Faraz Hach [2]

[1] Genome Informatics group, Faculty of Technology
Bielefeld University
Bielefeld, Germany

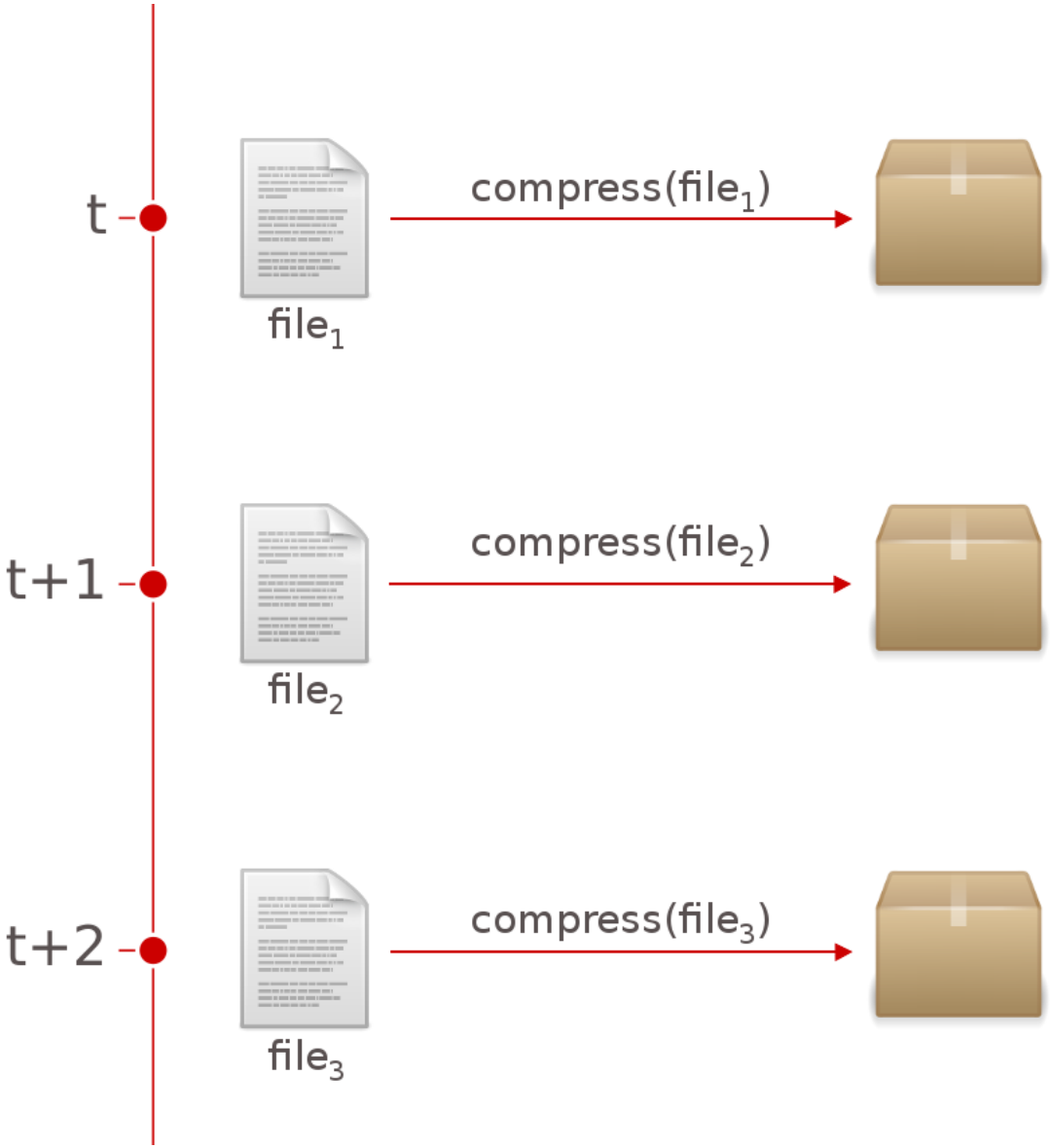
[2] School of Computing Science
Simon Fraser University
Burnaby, Canada

Motivations

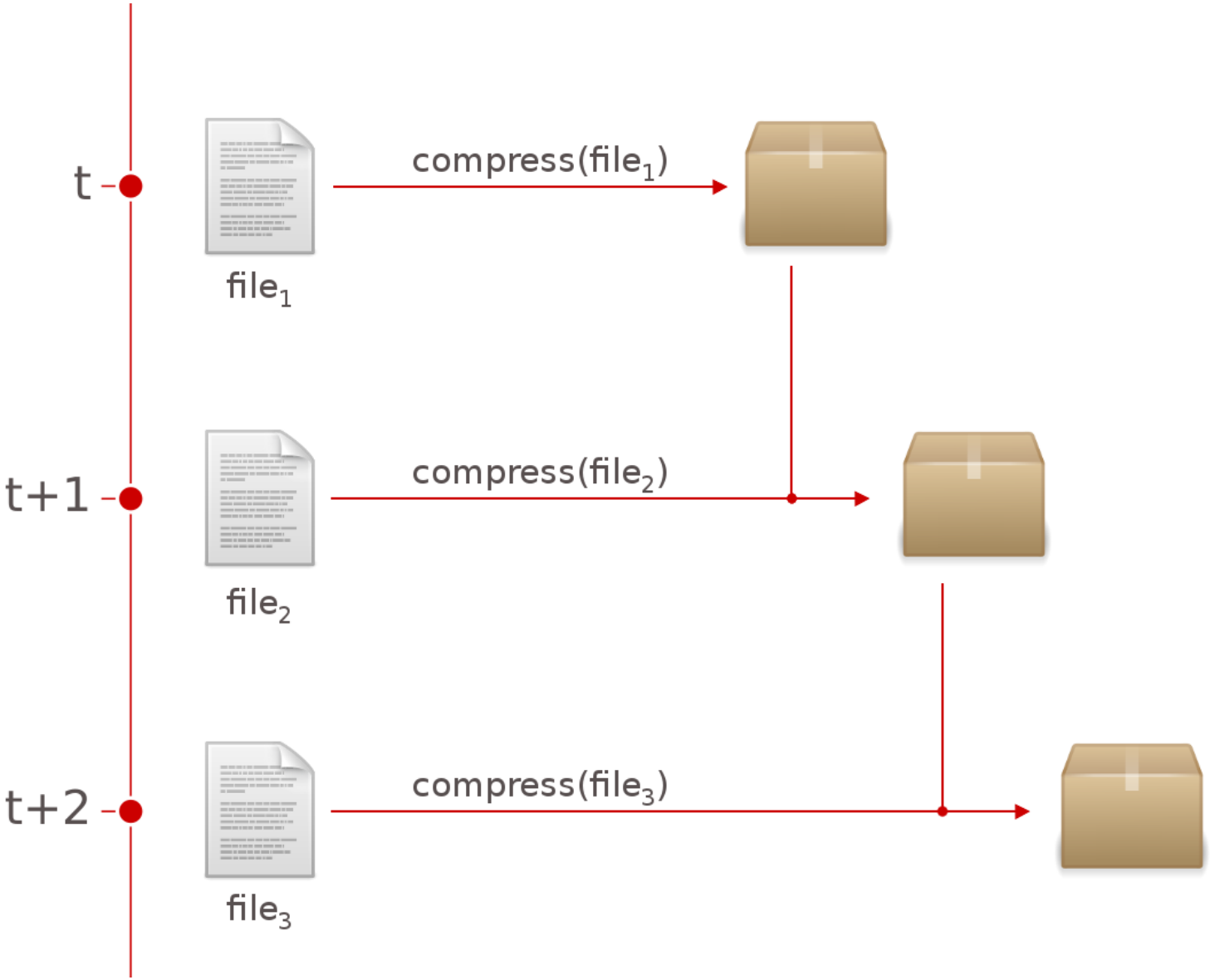
1000 GENOMES PROJECT DATA RELEASE DATES

- 2008 – Pilot 1 - Pilot 2
- Pilot 3
- 2009 – CEPH - Japanese - Toscan - YRI Yoruba
- Han Chinese CHB
- 2010 – Luhya
- CHS - Finnish - GBR - MXL - PUR
- CLM
- IBS
- 2011 – ACB - CDX - KHV - PEL
- GIH
- 2012 – BEB - ESN - GWD - ITU - MSL - PJL - STU
- CEU
- 2013 – ITU - PUR - STU (blood)
- ASW

SEQUENCING PROJECT TIMELINE



SEQUENCING PROJECT TIMELINE



Encoding

ENCODING A SEQUENCE AS A DBG: THE PROBLEM

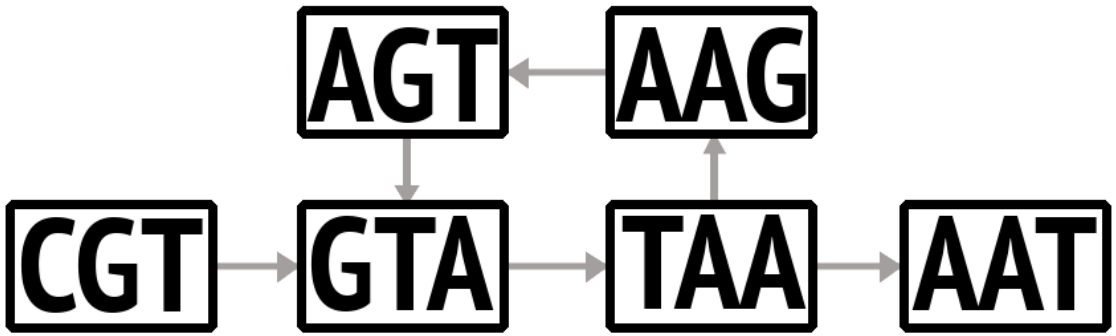


Figure: dBG of $S = CGTAAGTAAT$ using 3-mers.

GUIDED DE BRUIJN GRAPH (GDBG)

A *partition*(G, S) is a subgraph G' of G such that G' is a simple path that reconstructs a substring of S.

A guided de Bruijn graph (gdBG) is a de Bruijn graph with partitions.

ENCODING A SEQUENCE AS A GDBG

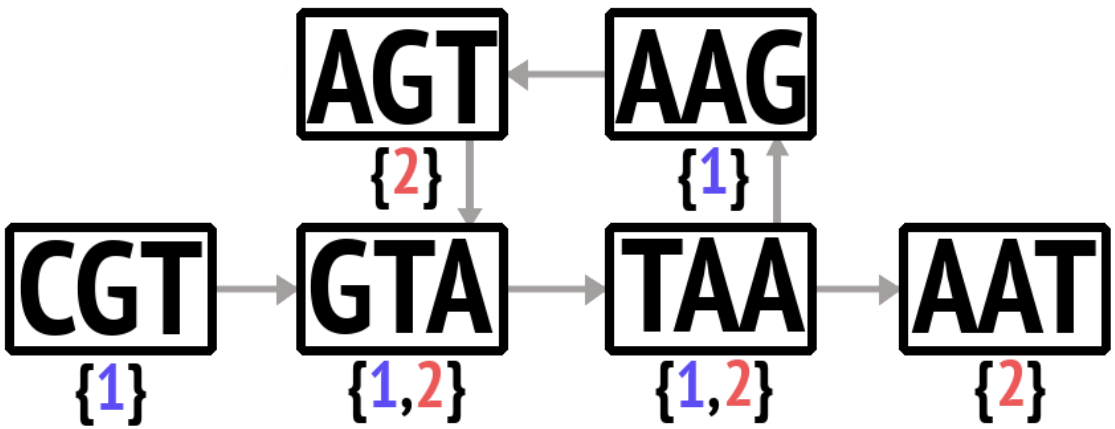


Figure: gdBG of $S = CGTAAGTAAT$ using 3-mers.

ENCODING A SEQUENCE AS A GDBG

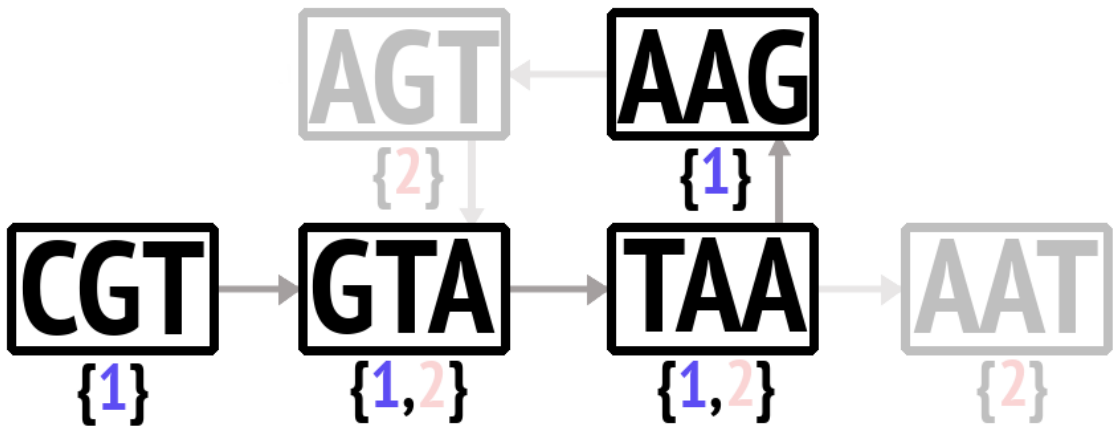


Figure: gdBG of S = *CGTAAGTAAT* using 3-mers.

ENCODING A SEQUENCE AS A GDBG

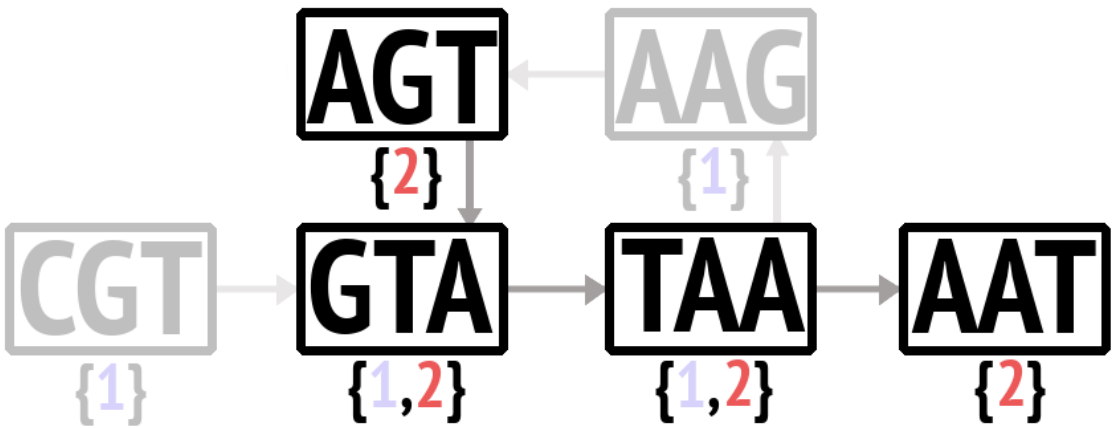
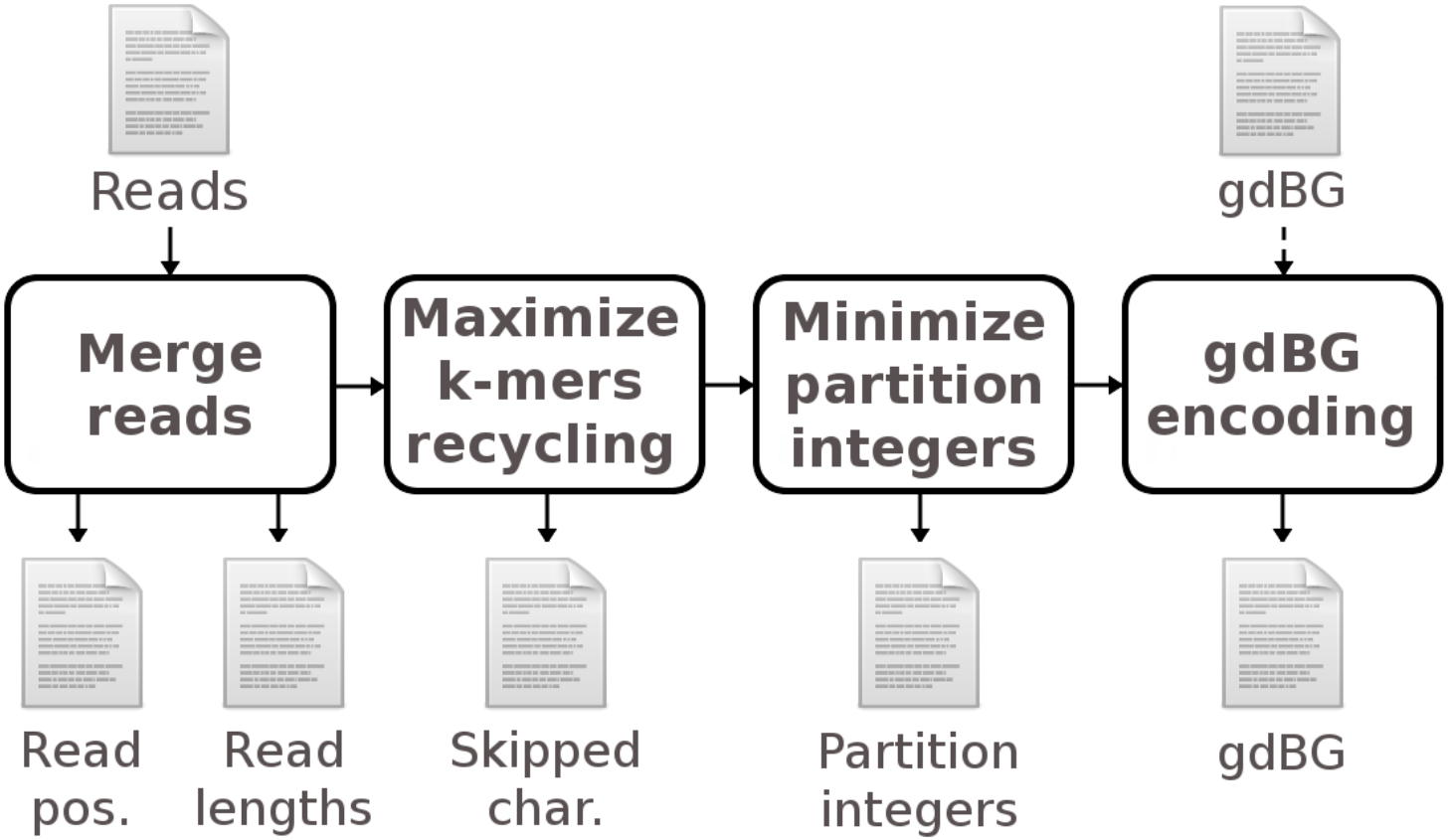


Figure: gdBG of $S = CGTAAGTAAT$ using 3-mers.

Compression pipeline

COMPRESSION PIPELINE



Results

RESULTS

Dataset: Bioproject PRJEB5438

473 isolates of *Pseudomonas aeruginosa* → 339 Gbp
Illumina HiSeq 2000 paired-end reads (2x100 bp)

DNA stream compression tools:

→ Mince (Patro *et al.*, 2015)

→ LEON (Benoit *et al.*, 2015)

→ ORCOM (Grabowski *et al.*, 2015)

RESULTS

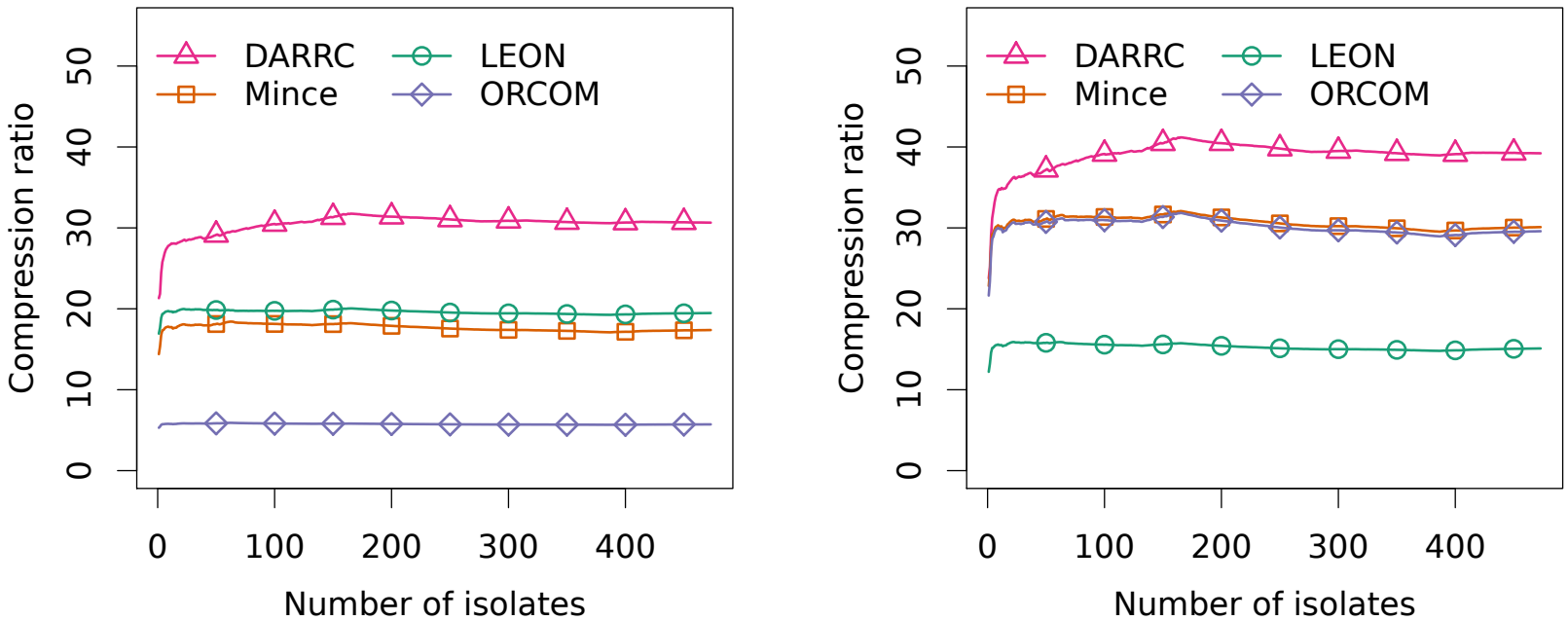


Figure: Compression ratios in paired-end mode (left) and single-end mode (right).

RESULTS

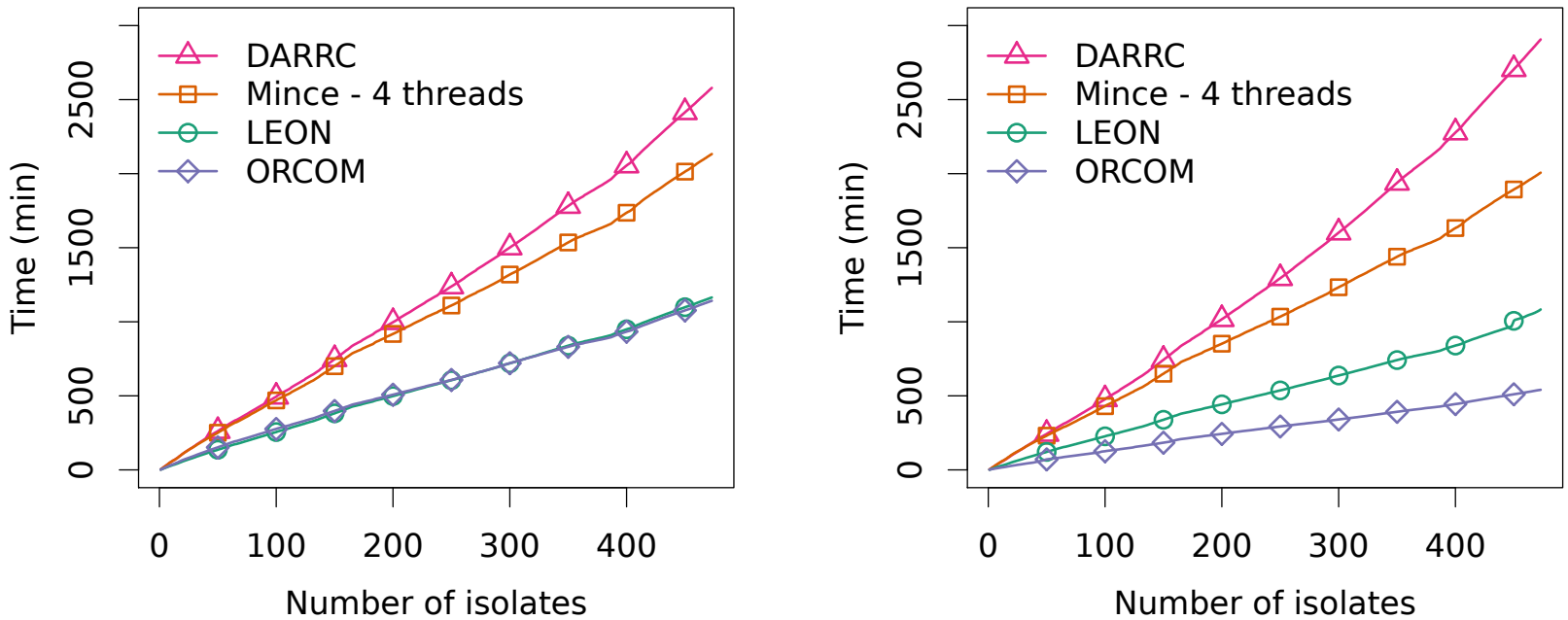


Figure: Compression times in paired-end mode (left) and single-end mode (right).

RESULTS

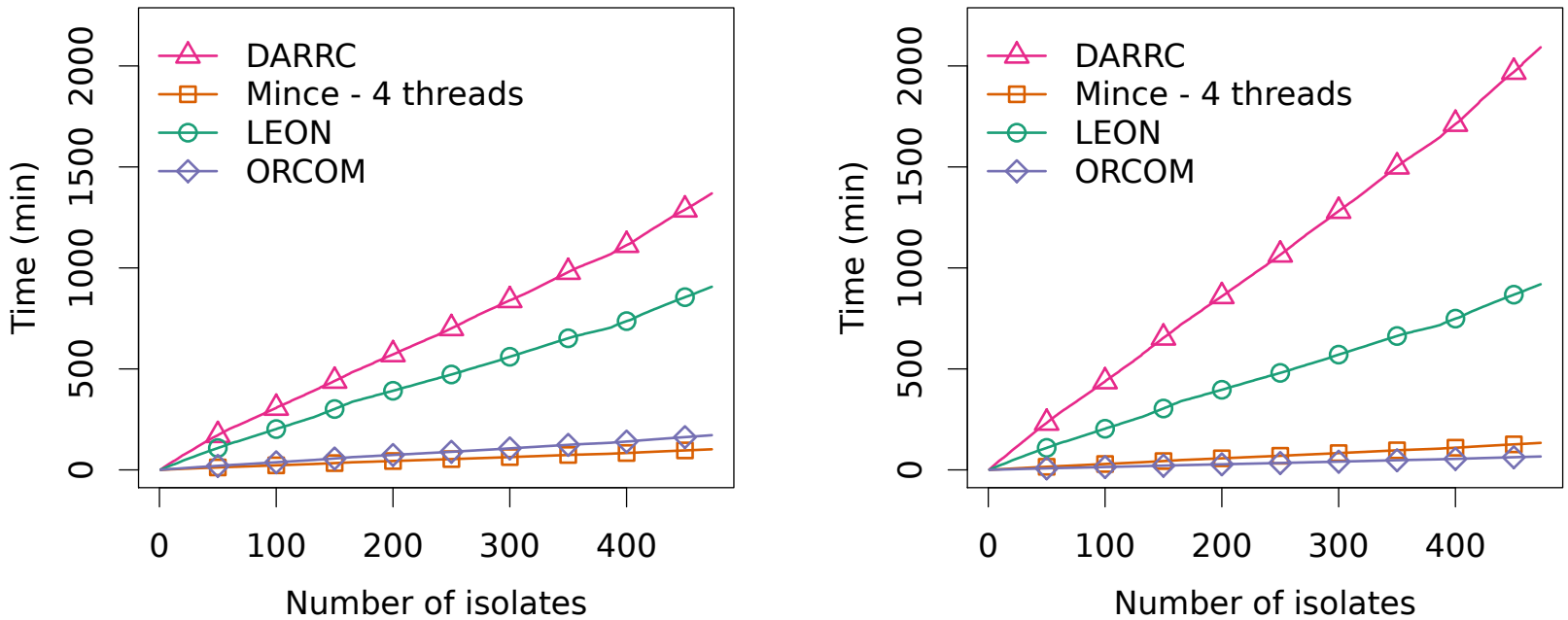


Figure: Decompression times in paired-end mode (left) and single-end mode (right).

RESULTS

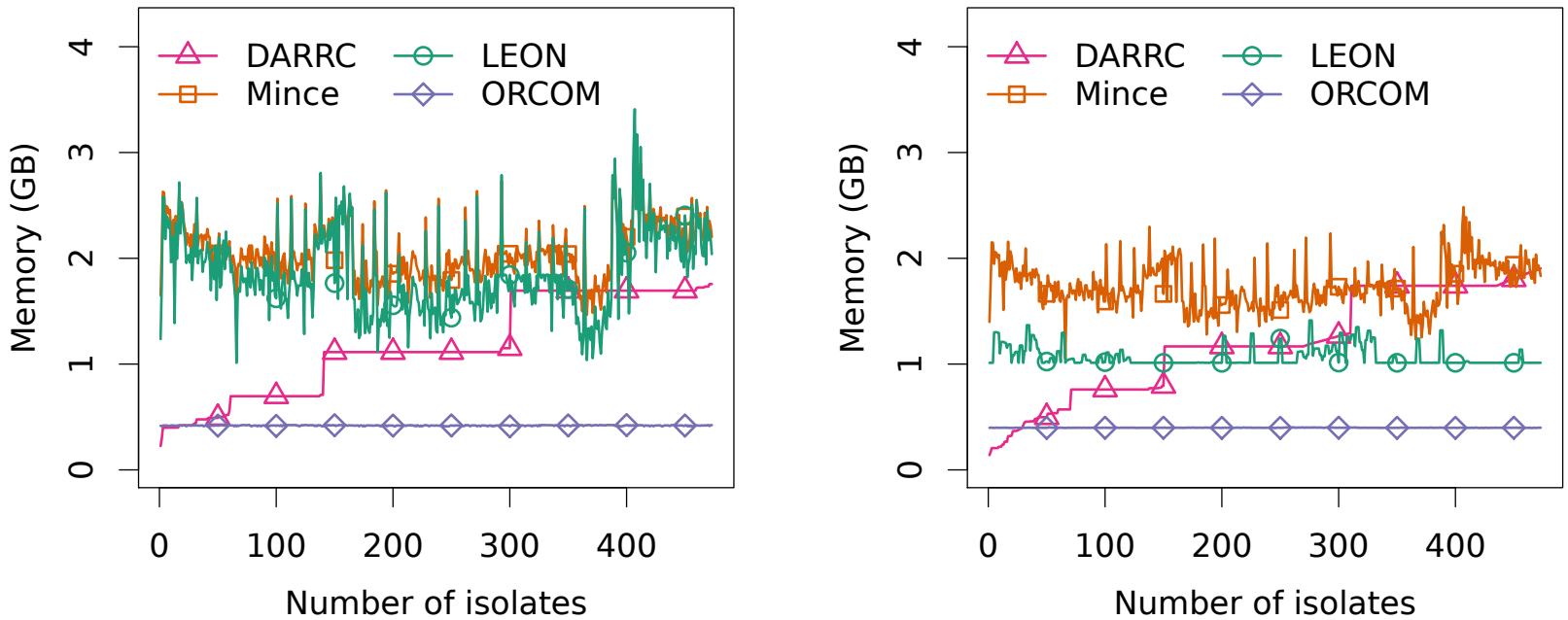


Figure: Compression main memory peaks in paired-end mode (left) and single-end mode (right).

RESULTS

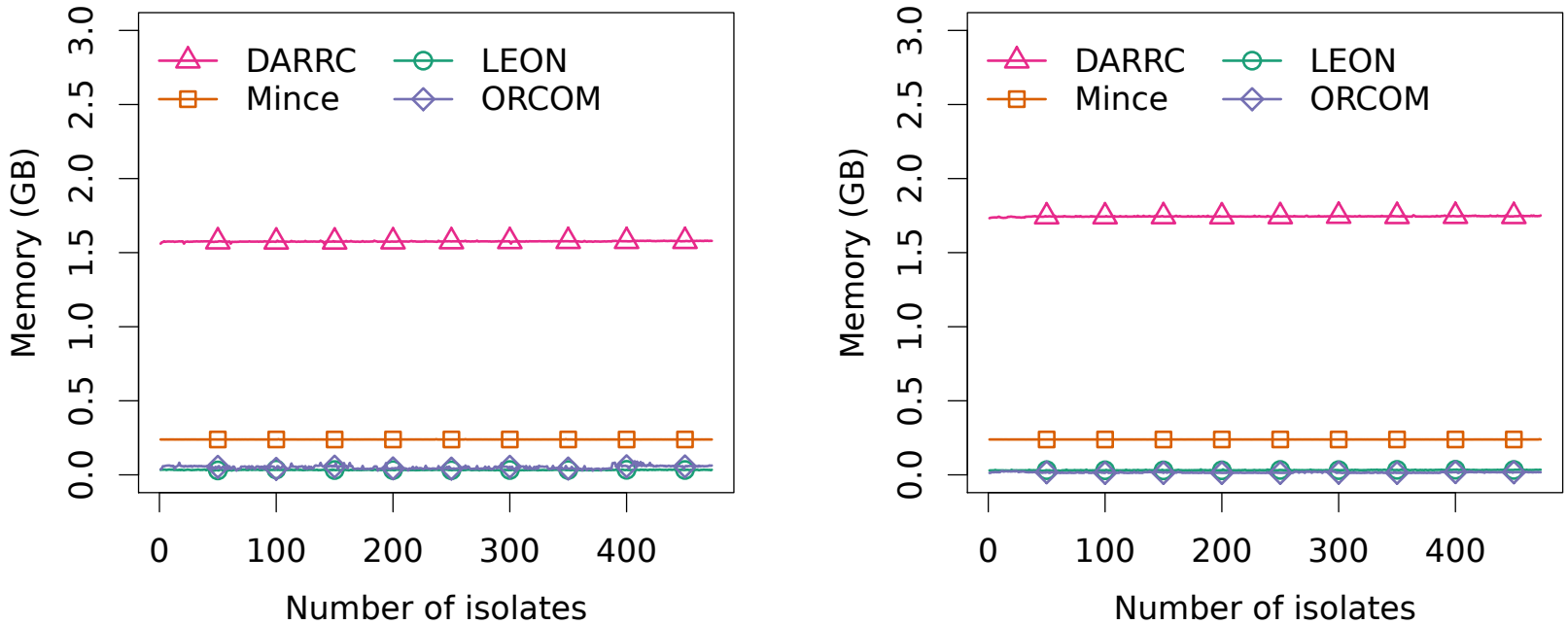


Figure: Decompression main memory peaks in paired-end mode (left) and single-end mode (right).

Conclusion

CONCLUSION

BFT:

→ pan-genome as a colored de-Bruijn graph stored in a trie

→ indexes and compresses colored *k*-mers

→ new representation of nodes and containers proposed

DARRC:

→ compress pan-genome sequencing reads *de novo*

→ compressed reads of any length, pair-ended or not, support full IUPAC nuc. alphabet

→ encode reads as a guided de Bruijn graph using the BFT

→ archive can be updated without full decompression

Thank you !