mumemto: efficient maximal matching across multiple genomes

Vikram Shivakumar[1] and Ben Langmead[1]

[1] Department of Computer Science, Johns Hopkins University

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Introduction & Motivation

- Intra-species collections of genome assemblies are growing, shifting genomics from single reference to pangenome-based analyses. These analyses rely inherently on a common coordinate system (such as a multiple sequence alignment, MSA).
- Recent breakthroughs[1] in compressed text indexing enable the computation of Burrows-Wheeler Transforms (BWT), suffix arrays, and other auxiliary arrays which were previously intractable for pangenome-sized sequence collections.
- Multiple Maximal Unique/Exact Matches (multi-MUM and multi-MEMs) can form a basis for an underlying MSA, while providing insights into genome structural diversity and conservation. Previously proposed enhanced suffix array-based methods[2] for computing multi-MUM/MEMs are now practical for pangenome-sized collections.
- We introduce **mumemto**: an efficient, multi-purpose tool to compute multi-MUMs and MEMs that scales to many genome assemblies and enables core genome alignment and visualization of pangenomes.

Computing multi-MUMs in the human pangenome

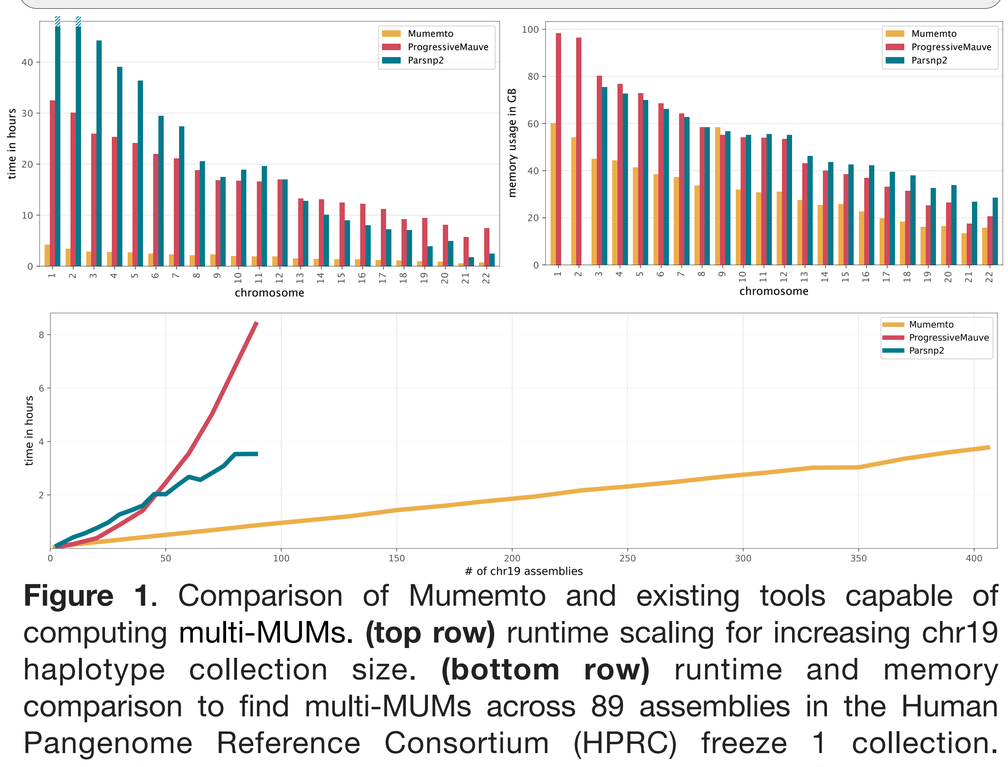


Figure 1. Comparison of Mumemto and existing tools capable of computing multi-MUMs. **(top row)** runtime scaling for increasing chr19 haplotype collection size. **(bottom row)** runtime and memory comparison to find multi-MUMs across 89 assemblies in the Human Pangenome Reference Consortium (HPRC) freeze 1 collection. *Parsnp2 could not compute multi-MUMs for chr1 and chr2.

Dataset	# partitions	# seqs	Serial		Parallel	
			Time (hrs)	Memory (GB)	Time (hrs)	Memory (GB)
chr19 HPRC (*N* = 474)	1	474	4.77	44.05	-	-
	5	96	5.00	13.96	1.11	66.78
	10	48	5.91	9.56	0.63	88.63
	20	24	7.21	7.16	0.39	125.29
	40	12	8.32	**5.35**	**0.26**	190.39
A. thaliana (*N* = 69)	1	69	2.58	70.34	-	-
	5	15	3.75	27.14	0.80	128.08
	10	7	4.92	18.79	0.57	172.00
	20	4	6.48	**13.58**	**0.40**	242.36

Table 1. Mumemto can be run over multiple partitions, enabling a time-memory tradeoff with parallelization. Runtime and memory for HPRC chr19 and *A. thaliana* assemblies in different partition schemes.

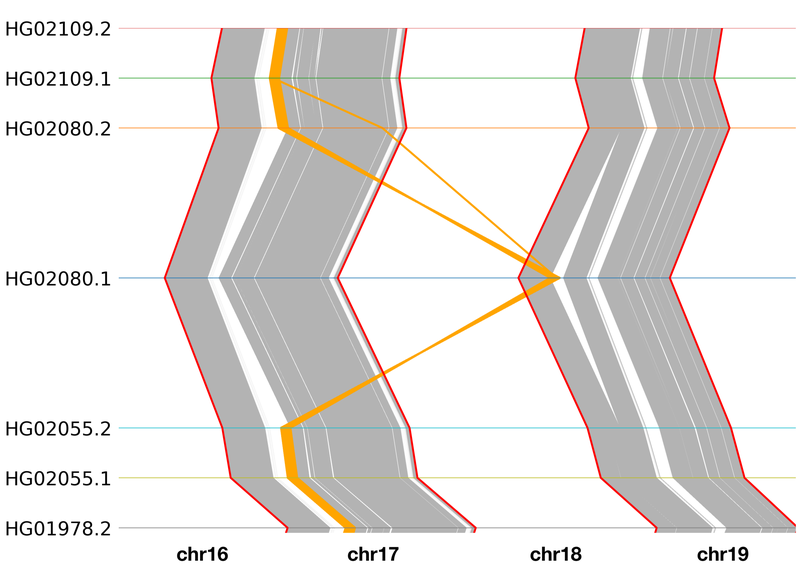


Figure 4. Evidence of misassembly from broken multi-MUM collinearity reveals a previously reported[6] misassembly (highlighted in gold) in HPRC assembly HG02080 (paternal haplotype).

mumemto

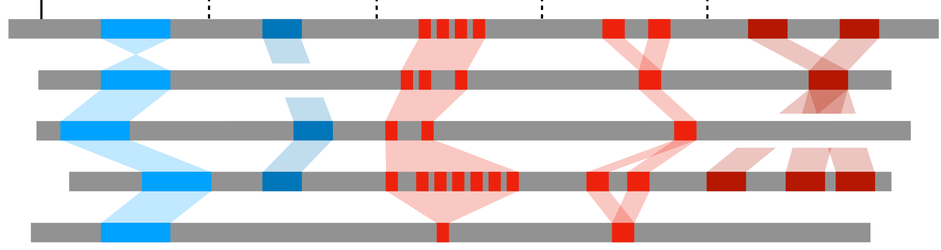
BWT	BWM	ID	LCP
…	…	…	…
A	ACAAAAGGACTGAA	1	3
T	ACAACCTAATCAGA	4	4
A	ACAACGTAATCAGA	2	5
T	ACAACGTACCAGAA	1	8
G	ACAACGTACGGGTA	3	9
T	ACAACGTACGGTTA	4	11
G	ACAACGTACGTGTA	2	10
C	ACCAATAGGATAGA	3	2
…	…	…	…
T	CCTTACATCATAGT	3	0
C	CCTTAAAATGACTA	2	5
C	CCTTAAGTGACTAG	2	5
C	CTTTAATGCAAGTC	4	1
…	…	…	…

- Given a collection of input sequences, Mumemto leverages **prefix-free parsing**, a compressed full-text indexing technique, to efficiently stream out the **BWT**, **suffix array (SA)**, and **longest common prefix (LCP) array**
- multi-MEMs are maximal exact matches that appear in each sequence and are non-extendable[2]. multi-MUMs are maximal exact matches that appear in each sequence in the collection **exactly once**[2]
- Mumemto can **visualize multi-MUM synteny**, revealing genomic structural diversity and sequence conservation within a pangenome collection
- Mumemto **integrates with alignnment tools** (Parsnp2[3]) to scale core genome alignment from bacterial to human genome-length pangenomes, and can accelerate pangenome graph construction

mumemto is multi-purpose and versatile, computing multi-MUMs and MEMs across any subsets of sequences within a collection

Partial multi-MUMs can reveal evolutionary relationships between sequences, rare MEMs can highlight duplication events, and partial multi-MEMs can represent copy-number polymorphisms within a population

	multi-MUM	partial multi-MUM	multi-MEM	multi-MEM	partial multi-MEM
# of sequences	all	some	all	all	some
# occ / sequence	one	one	no limit	at most f	no limit
mumemto flags	default	-k <INT>	-f 0	-f <INT>	-k <INT> -f 0



Pangenome synteny reveals assembly errors

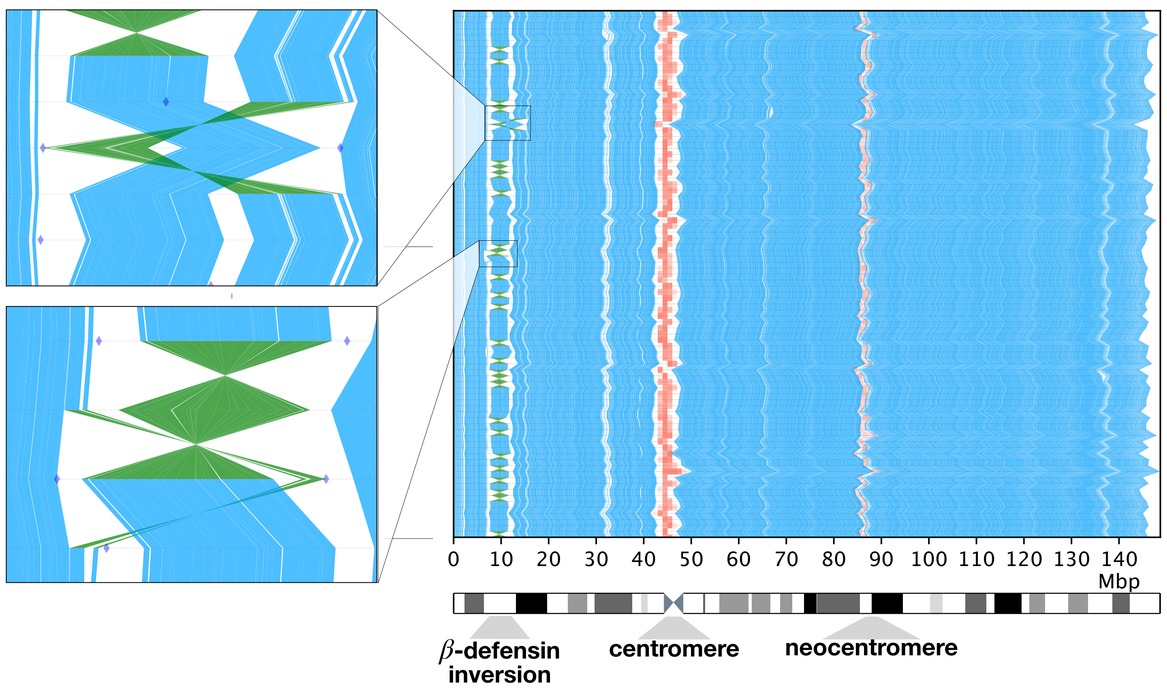


Figure 2. chr8 synteny across HPRC assemblies (computed in ~2 hrs). Centromeres are visible as gaps in multi-MUMs, but multi-MEMs reveal satellite repeats that are integral to centromeric function (shown in red heatmap). (Side panels) Potential contig orientation errors caused by single-reference-based scaffolding revealed by multi-MUM synteny.

Partial multi-MUMs reveal evolutionary insights

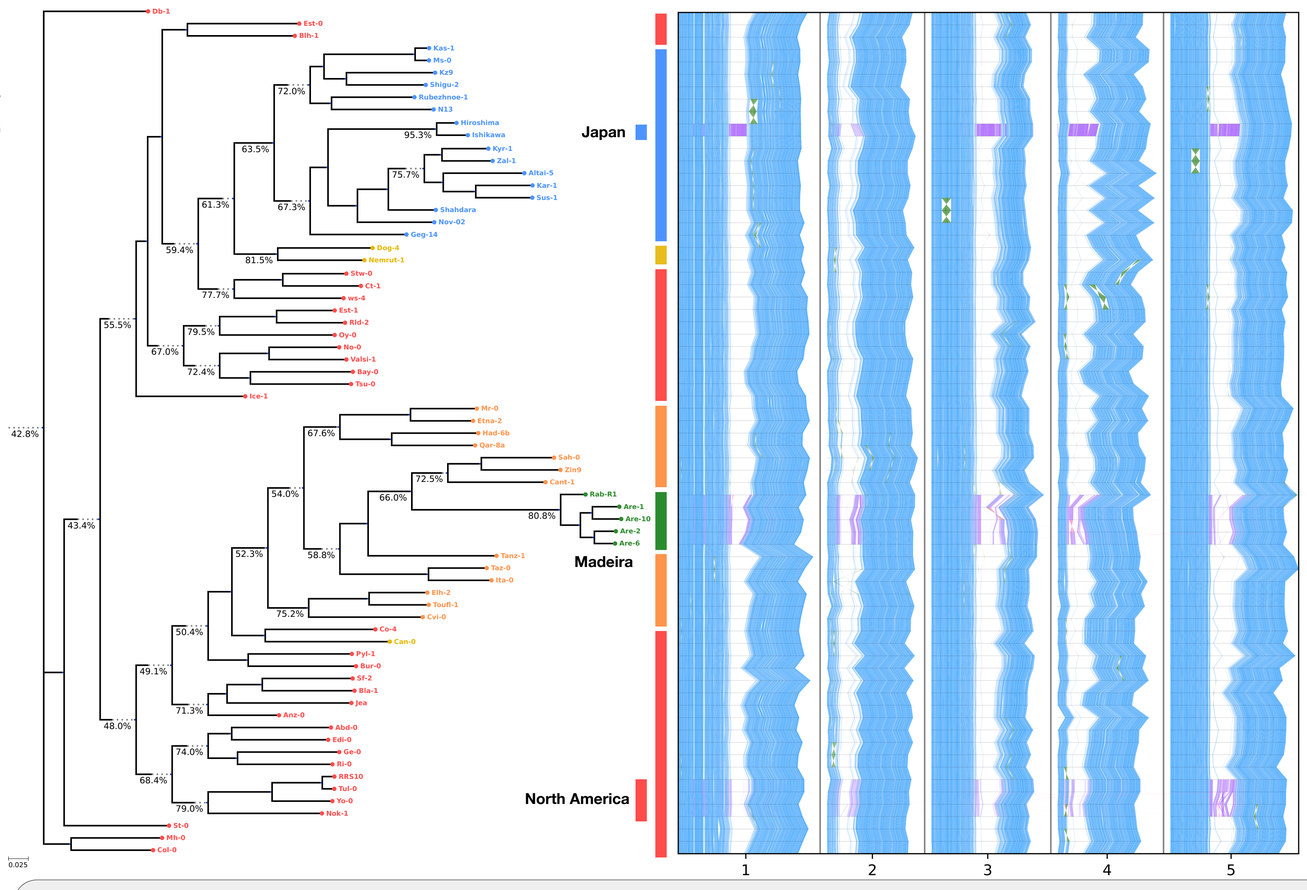


Figure 3. (left) Phylogeny of 69 geographically diverse *A. thaliana* accessions[5], with regions colored. Internal nodes labeled with MUM coverage over subcluster of genomes. **(right)** multi-MUM synteny across full dataset (blue, inversions in green), and subgroup specific multi-MUMs (purple, inversions in pink) for three high coverage subgroups.

multi-MUMs improve pangenome read mapping

- multi-MUMs serve as guideposts for collinearity in a pangenome, and they can be computed *while* building a full-text indexes with PFP.
- Integrating MUM information into compressed pangenome indexes (such as *r*-index) can provide pangenome collinearity information during read mapping.
- col-BWT (collinear BWT)[7] implements multi-MUM-aware read mapping with the *r*-index and MONI algorithm, improving mapping over SPUMONI.

References

1. Boucher, C., *et al*. (2019). Prefix-free parsing for building big BWTs. *Algorithms for Molecular Biology*, *14*, 1-15.
2. Abouelhoda, M.I., *et al*. (2002). The enhanced suffix array and its applications to genome analysis. *WABI 2002 Proceedings 2* (pp. 449-463).
3. Kille, et al (2024). Parsnp 2.0: Scalable Core-Genome Alignment for Massive Microbial Datasets. *bioRxiv*, 2024-01.
4. Darling, et al (2010). Progressive Mauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*. 5(6)
5. Lian, et al. (2024). A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. *Nature genetics*, *56*(5), 982-991.
6. Liao, et al (2023). A draft human pangenome reference. *Nature*, 617(7960), 312-324.
7. Brown, et al (2024). Improved pangenomic classification accuracy with chain statistics. *RECOMB 2025.*

Work supported by NIH grants R01HG011392, NSF BIO grant DBI-2029552, NSF DGE2139757

Github

Papers