MAUVE: MULTIPLE ALIGNMENT OF CONSERVED GENOMIC SEQUENCES WITH REARRANGEMENTS

Aaron C.E. Darling, Bob Mau, Frederick R. Blattner and Nicole T. Perna

University of Wisconsin – Madison (2004).

Presented by Natalie Yudin

Evolutionary Events

SMALL-SCALE

- Single Mutations
- Deletions
- Insertions

LARGE-SCALE

- Recombination
  - Observed in organisms of all types
- Gene Loss
- Gene duplication and Horizontal transfer
  - Commonly observed in higher eukaryotes and bacteria

Computational Methods

Short sequences

- Pairwise
  - Needleman-Wunsch global
  - Smith-Waterman local
- Multiple
  - Dynamic programming

◉ $O(n^2)$

◉ Typically n > 10 kb

Long sequences

- Assumptions
  - Highly similar subsequences will be part of the correct global alignment
  - Reduce the number of possible global alignments considered in dynamic step.
  - Spurious local alignments can be found
- Pairwise
  - MUMmer, GLASS, AVID, WABA
- Multiple
  - MAVID, MLAGAN, MGA

Computational Methods

All of them assume that the sequence is free from significant rearrangements

Let's consider it

Pairwise

- Shuffle-LAGAN
  - Selects anchors collinear in the first sequence with rearrangements permitted in the second sequence

Multiple

- MultiPipMaker
  - Uses BLASTZ to align multiple genome to a single genome with a presence of rearrangement
  - Does local alignment on pairs then constructs rough global alignment

Do not identify the breakpoints of multiple genome rearrangements.

Mauve

- Identifies large-scale evolutionary events
  - conserved genomic regions
  - Rearrangements
  - Inversions in conserved regions
  - Exact sequence breakpoints of rearrangements
- Provides traditional multiple alignment of conserved regions, identifying small insertions and deletions

Anchored Alignment Approach

- Identify liner collinear blocks (LCB) that contain sequence elements conserved among all genomes being aligned
  - Particular motif appears *r* times in each of *G* sequences will result in r^G possible anchors
  - Mauve's solution - Multiple Maximal Matches (Multi-MUMs)
- Those anchor blocks are required to meet user-set minimum weight – the measure of confidence that is it true rearrangement and not a spurious match.
- Apply ClusterW progressive global alignment at each LCB.
- Visualize the results

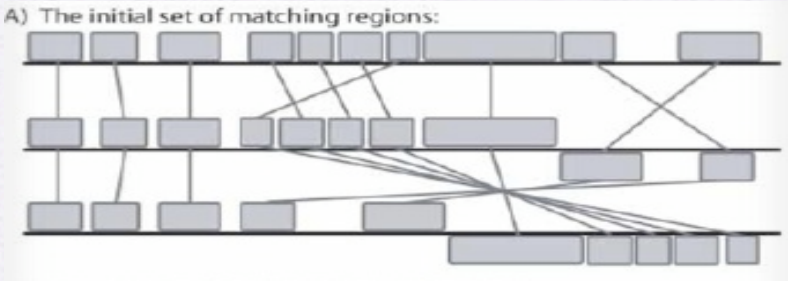
Algorithms

- Find local alignments (multi-MUMs)
- Use the multi-MUMs to calculate a phylogenetic guide tree
- Select a subset of the multi-MUMs to use as anchors – these anchors are partitioned into collinear groups called LCBs
- Perform recursive anchoring to identify additional alignment anchors within and outside each LCB
- Perform a progressive alignment of each LCB using the guide tree

Finding Multi-MUMs

What is it?

- Mutli-MUMs are exactly matching subsequences shared by two or more genomes and are bounded on either side by mismatched nucleotides
- If multi-MUM includes a region in reverse orientation in sequence *j* we define $M_i \bullet S_j$ to be negative
- If multi-MUM does not exist in the sequence, we define $M_i \bullet S_j = 0$
  - Multiplicity(M_i) = the number of genomes for which $M_i \bullet S_j \neq 0$



Formal Definition

- Multi-MUM is a tuple
  - $<L, S_1, \ldots, S_G>$
  - L is a length
  - S_j is a left end position of the multi-MUM in j-th genome
- Multi-MUMs
  - $\mathbf{M} = \{M_1, \ldots, M_N\}$
  - $M_i \bullet L$ is length of i-th Multi-MUM
  - $M_i \bullet S_j$ left end positionof i-th Multi-MUM
- $O(G^2n+Gn \log Gn)$

Algorithms

- Find local alignments (multi-MUMs)
- Use the multi-MUMs to calculate a phylogenetic guide tree
- Select a subset of the multi-MUMs to use as anchors – these anchors are partitioned into collinear groups called LCBs
- Perform recursive anchoring to identify additional alignment anchors within and outside each LCB
- Perform a progressive alignment of each LCB using the guide tree

Calculate a Guide Tree

- Exploit the subset information provided by subset multi-MUMs as a distance metric to construct a phylogenetic guide tree using Neighbor Joining .
- Resolves overlaps between Multi-MUMs
- Sequence similarity = $\frac{\text{base pairs shared}}{\text{avg. genome length}}$
- This similarity is converted into a distance value for Neighbor Joining.

Algorithms

- Find local alignments (multi-MUMs)
- Use the multi-MUMs to calculate a phylogenetic guide tree
- Select a subset of the multi-MUMs to use as anchors – these anchors are partitioned into collinear groups called LCBs
- Perform recursive anchoring to identify additional alignment anchors within and outside each LCB
- Perform a progressive alignment of each LCB using the guide tree

Selecting a Set of Anchors

- Multi-MUMs with Multiplicity < G are removed from **M**
- Partition into LCBs
  - *lcb* is a subset of M: $lcb = \{M_1, \ldots, M_{|lcb|}\}$
  - Satisfies total ordering: $M_i \bullet S_j \leq M_{i+1} \bullet S_j$
  - Holds for all i $1 \leq i \leq |lcb|$ and all j $1 \leq j \leq G$
- Uses greedy breakpoints elimination algorithm to remove low-weight collinear blocks of **M**, until **M** meets the minimum weight requirement.

Selecting a Set of Anchors

Substep 1. Determine a partitioning of **M** into collinear blocks **CB**.

Substep 2. Calculate the weight, $w(cb_i)$ of each collinear block $cb_i \in \mathbf{CB}$.

Substep 3. Let $z = \min_{cb \in \mathbf{CB}} w(cb)$.

Substep 4. Stop if $z \geq MinimumWeight$.

Substep 5. Identify the collinear subsets $\mathbf{MinCB} \subseteq \mathbf{CB}$ that satisfy $w(cb_i) = z$.

Substep 6. For each $cb \in \mathbf{MinCB}$, remove each multi-MUM $M \in cb$ from **M**.

Substep 7. Go to substep 1.

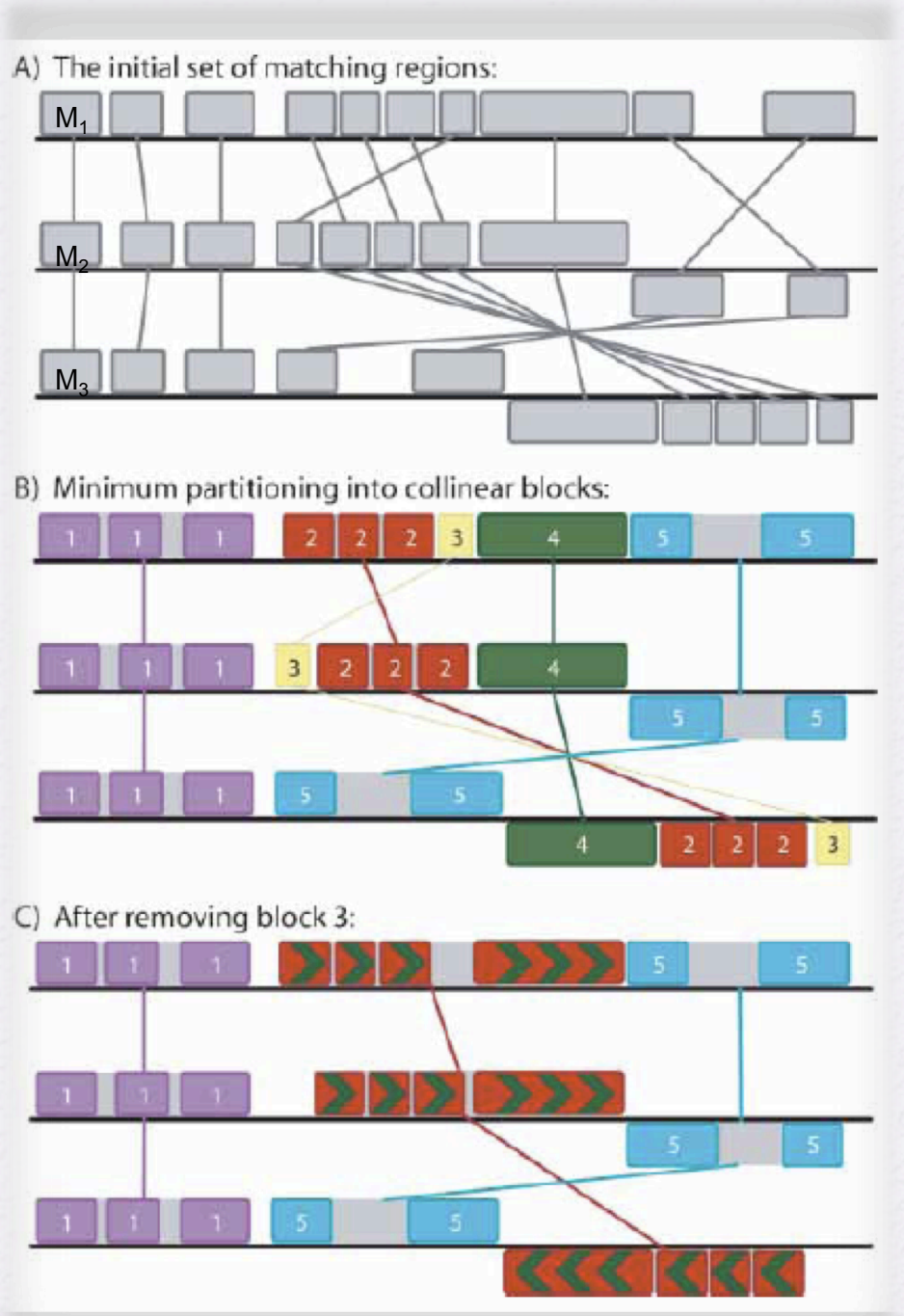
Picture!

A pictorial representation of greedy breakpoint elimination in three genomes.

(A) The initial set of matching regions (multi-MUMs) represented as connected blocks. Blocks below a genome's center line are inverted relative to the reference sequence.

(B) The matches are partitioned into a minimum set of collinear blocks. One connecting line is drawn per collinear block. Block 3 (yellow) has a low weight relative to other collinear blocks.

(C) As low-weight collinear blocks are removed, adjacent collinear blocks coalesce into a single block, potentially eliminating one or more breakpoints. Gray regions within collinear blocks are targeted by recursive anchoring.



Algorithms

- Find local alignments (multi-MUMs)
- Use the multi-MUMs to calculate a phylogenetic guide tree
- Select a subset of the multi-MUMs to use as anchors – these anchors are partitioned into collinear groups called LCBs
- Perform recursive anchoring to identify additional alignment anchors within and outside each LCB
- Perform a progressive alignment of each LCB using the guide tree

Recursive Anchoring and Gapped Alignment

- Using Multi-MUMs with smaller lengths
- Two types of recursive anchoring
  - Unanchored regions within LCB.
    - Recursive until the length of Multi-MUM is smaller than the sequence
  - Regions outside of LCB
    - Outside sequences in the entire genome is concatenated into a single sequence per genome
    - The left-hand coordinate has to be transposed back into the original coordinate system.
    - Any matches spanning two concatenated subsequences has to be split

Recursive Anchoring and Gapped Alignment

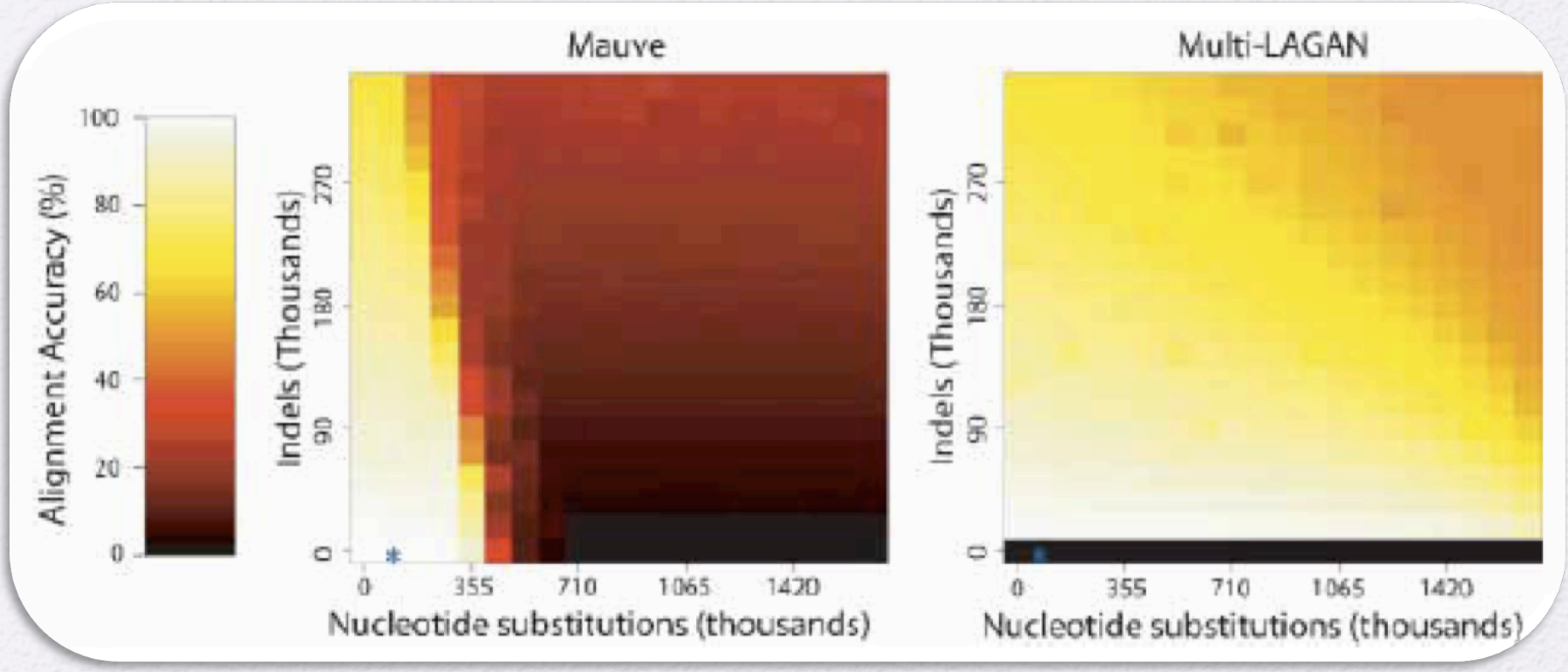
"Armed with a complete set of alignment anchors, Mauve performs a CLUSTAL W progressive alignment using the genome guide tree calculated previously."

Evaluating Alignment Quality

- Due to the lack of manually curated “correct” alignment , they estimated the alignment accuracy by modeling evolution and aligning simulated data sets.
  - Used HKY model implemented in Monte Carlo simulated package called Seqgen.
  - Model includes small insertions and deletion, large and small horizontal transfers and inversions
  - All events are simulated using Poisson distribution

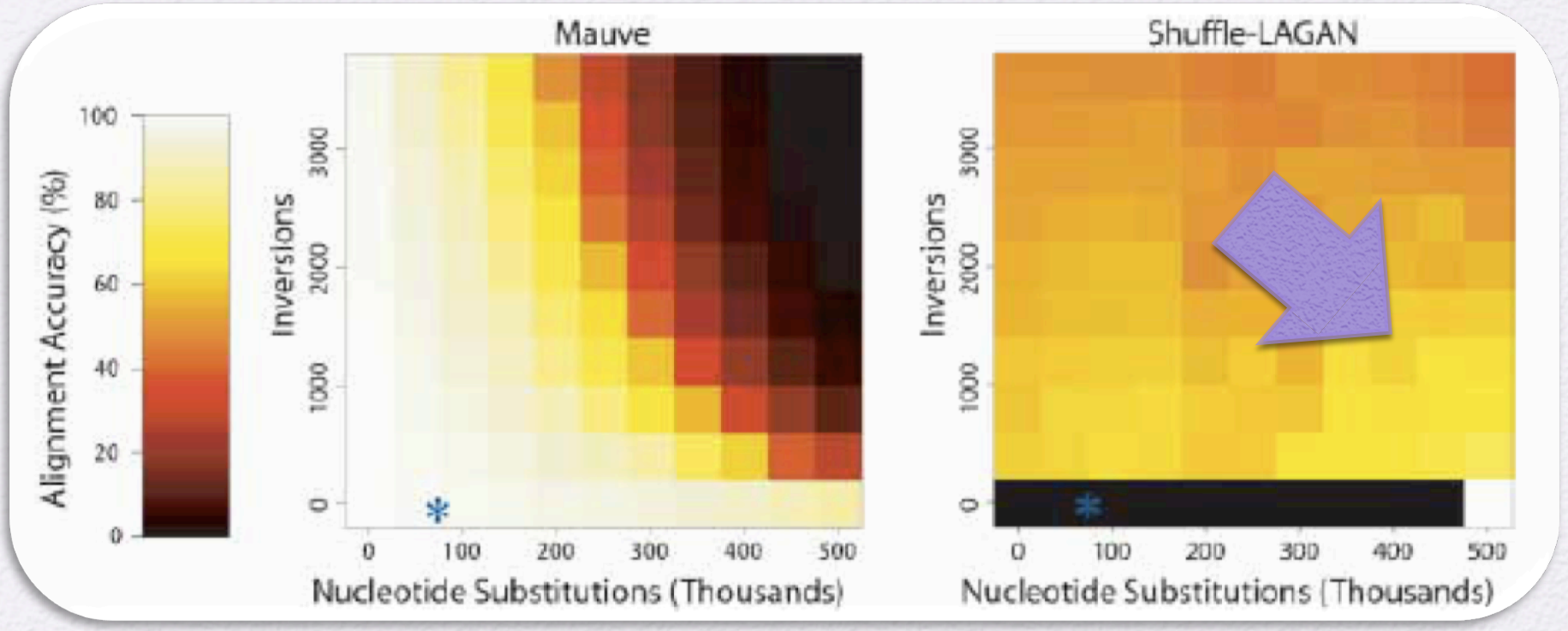
Mauve vs. Multi-LAGAN

- Genomes with high nucleotide substitutions and indel rate
- Experiment tests the sensitivity of the anchoring method



Mauve vs. Shuffle-LAGAN

- Genomes with presence of rearrangement



Alignment of Nine Enterobacterial Genomes

- Previous studies have shown that these genomes have undergone significant horizontal transfer and numerous genome rearrangements since their divergence.

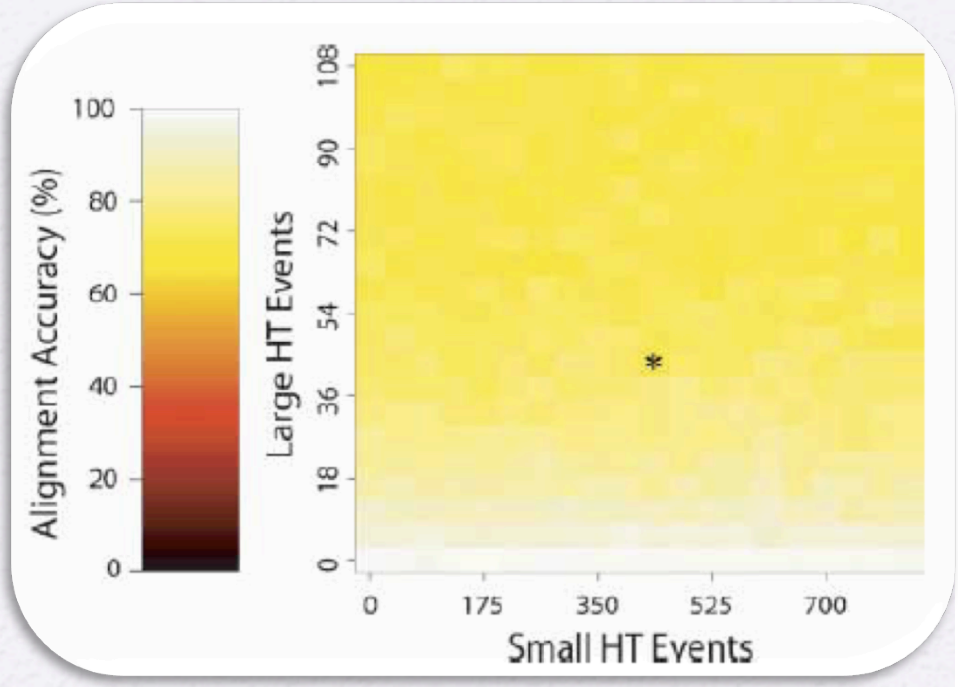
Species	Genome size	Reference
E. coli K12 MG1655	4,639,221	Blattner et al. 1997
E. coli O157:H7 EDL933	5,524,971	Perna et al. 2001
E. coli O157:H7 VT-2 Sakai	5,498,450	Hayashi et al. 2001
E. coli CFT073	5,231,428	Welch et al. 2002
S. flexneri 2A 2457T	4,599,354	Wei et al. 2003
S. flexneri 2A	4,607,203	Jin et al. 2002
S. enterica Typhimurium LT2	4,857,432	McClelland et al. 2001
S. enterica Typhi CT18	4,809,037	Parkhill et al. 2001
S. enterica Typhi Ty2	4,791,961	Deng et al. 2003

But first…

The performance of Mauve when aligning sequences evolved with rates similar to those observed among the group of nine enterobacteria.

In this experiment, the substitution, indel, and inversion frequencies were held constant at rates similar to those observed in the enterobacteria.

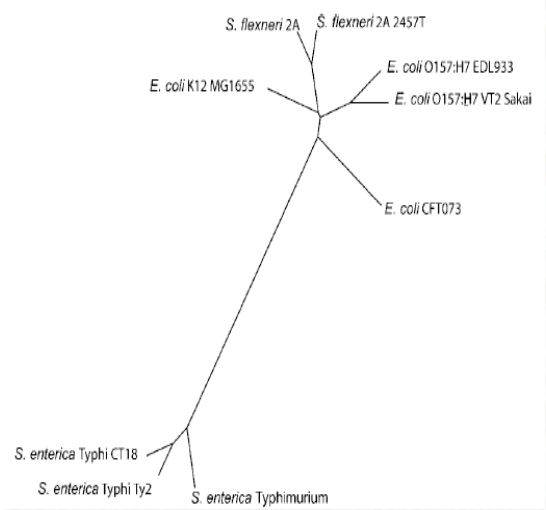
The asterisk (*) denotes the combination of large and small horizontal transfer rates observed in the enterobacteria.



- When scored only on regions considered backbone sequence, the accuracy is consistently above 98%.

Results for Nine Enterobacterial Geomes

- Took 3 hours to align



Data

- Represent extracted backbone sequences from the alignment.
- Significant lineage specific regions (grey areas) remained unaligned

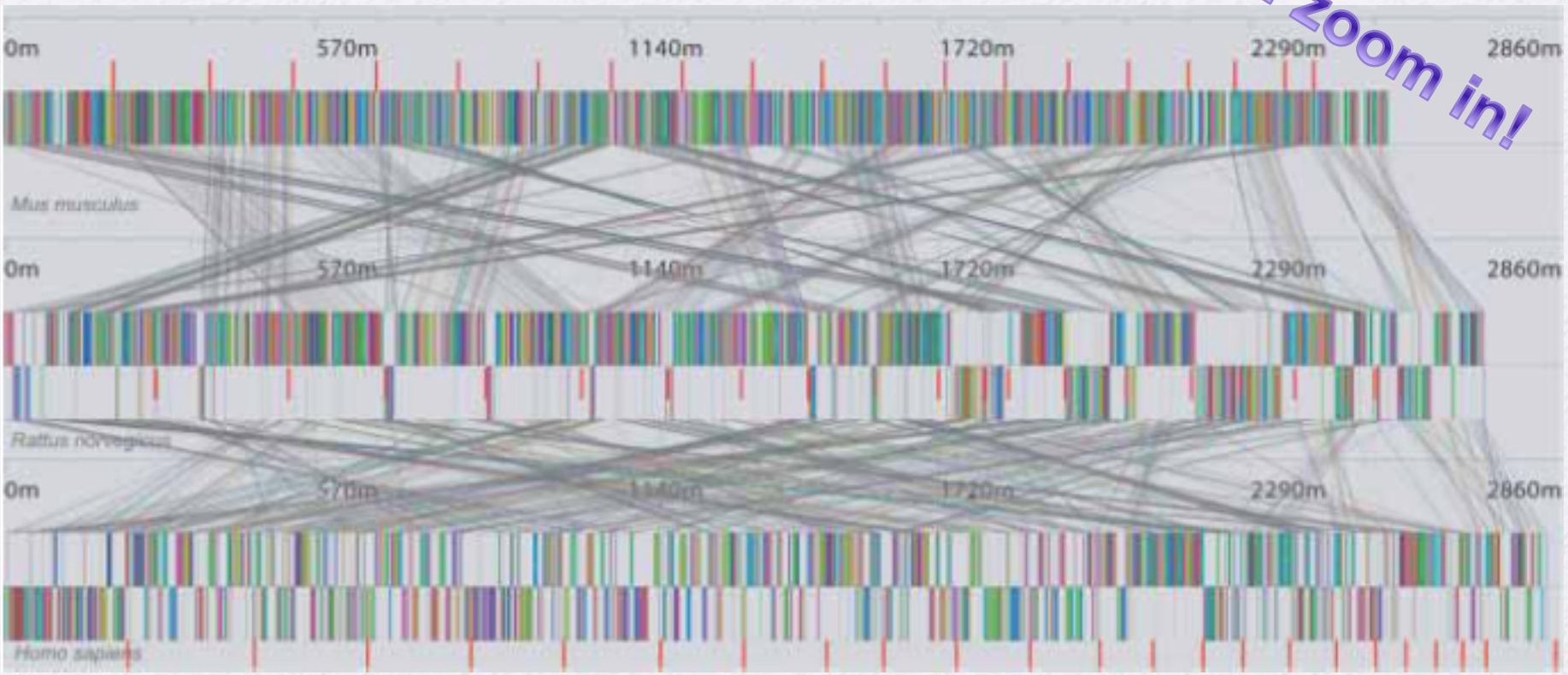
		1	2	3	4	5	6	7	8	9
1	*E. coli* K12 MG1655	1.000	—	—	—	—	—	—	—	—
2	*E. coli* EDL933	0.977	1.000	—	—	—	—	—	—	—
3	*E. coli* VT-2 Sakai	0.978	1.000	1.000	—	—	—	—	—	—
4	*E. coli* CFT073	0.965	0.966	0.967	1.000	—	—	—	—	—
5	*S. flexneri* 2a	0.976	0.975	0.975	0.963	1.000	—	—	—	—
6	*S. flexneri* 2a 2457T	0.976	0.975	0.975	0.962	0.999	1.000	—	—	—
7	*S. Typhimurium*	0.794	0.793	0.793	0.793	0.791	0.791	1.000	—	—
8	*S. typhi* CT18	0.792	0.791	0.791	0.792	0.790	0.789	0.981	1.000	—
9	*S. typhi* Ty2	0.793	0.793	0.793	0.793	0.791	0.791	0.984	0.996	1.000

Although an average of only 58% of the genomes is conserved across species, the level of sequence identity is remarkably high, suggesting that horizontal transfer and differential gene loss may account for the majority of phenotypic diversity among bacteria in this group.

For fun!

- Human genome build 34, mouse genome build 32 and rat genome RGSC build 3.1
- Took 12 hours
- Each of the 1251 blocks has a minimum weight of 90.
- Red vertical bars demarcate interchromosomal boundaries.

You can zoom in!



Limitations

- Does not do very well with divergent sequences (comparing to Mulit-LAGAN)
- Large regions (lineage specific) are not aligned
- Minimum LCB weight has to be manually determined for accurate estimation
- It was difficulty aligning genome with vast amount of segmental duplication

Progressive Mauve

Strengths

- Can align lager taxa
- Can align more divergent genome (less than 50% nucleotide identity is usually alignable)
- Manual adjustment of alignment scoring parameters is not necessary
- Can align conserved sequences in the subset of the input genomes

Limitations

- Significantly slow than the original Mauve
- Consumes more memory than original Mauve
- Manual adjustment of the breakpoint penalty might still be required

Questions?