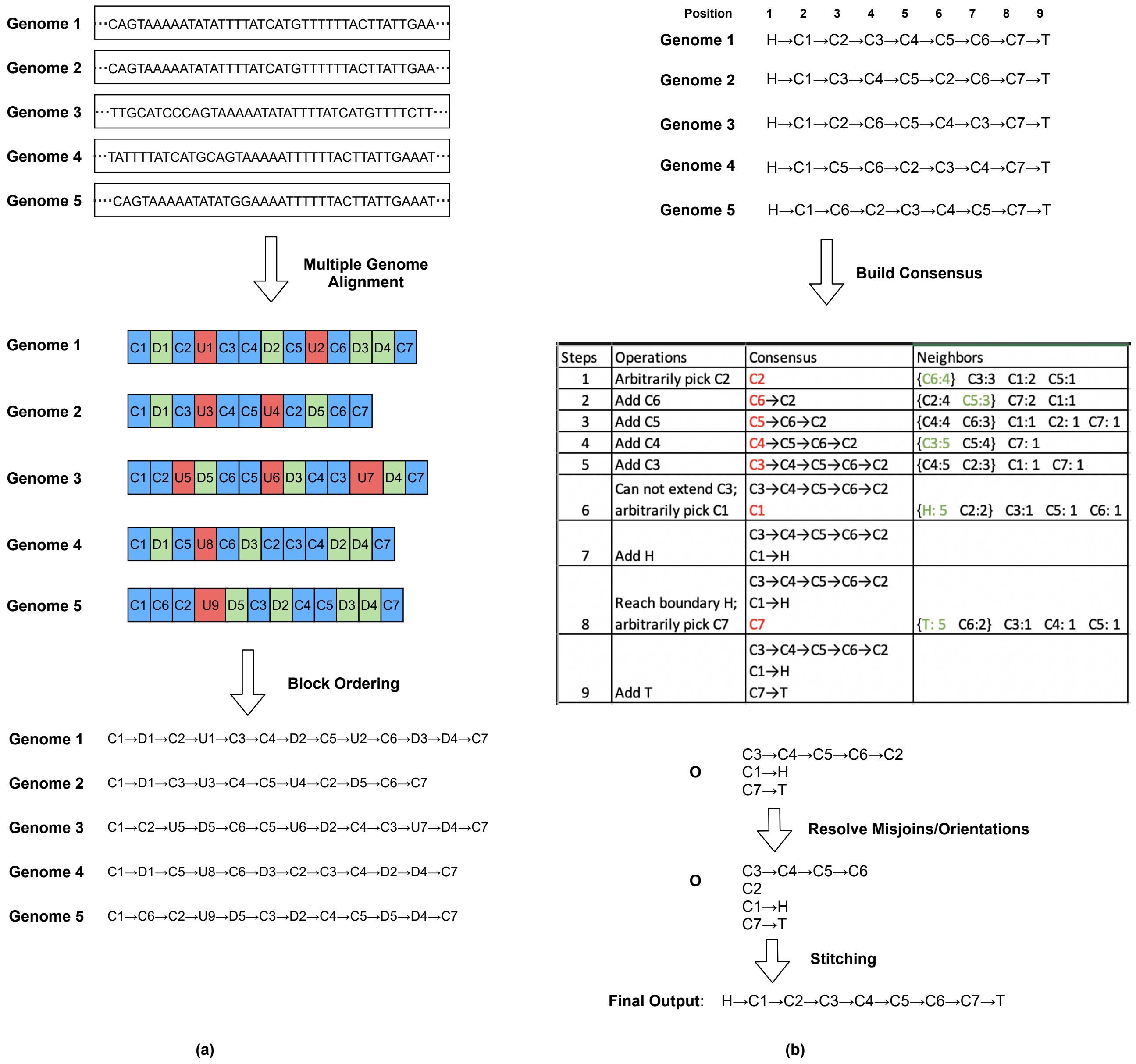
Reference-agnostic Representation and Visualization of Pan-genomes (Supplementary Materials)

Qihua Liang[1] **and Stefano Lonardi**[1,*]

[1]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA
[*]e-mail: stelo@ucr.edu

An Example of PGV's Consensus Algorithm

In the example in Figure S1(b) we assume that PGV arbitrarily starts from $C2$, and initializes the consensus ordering $O = \{C2\}$. Then, PGV collects the frequencies for the neighbors of $C2$: $C6$ occurs four times, $C3$ occurs three times, $C1$ occurs two times, and $C5$ occurs one time. $C6$ with majority vote is added to O, resulting in $O = \{C6 \to C2\}$. In Step 2, PGV collects the frequencies for the neighbors of $C6$. In the top two, only $C5$ is not in O, and thus we extend $O = \{C5 \to C6 \to C2\}$. Similarly, in Step 4 and 5, $C4$ and $C3$ are appended to the consensus. In Step 6, $C3$ cannot be extended because both its top two neighbors are already in O. Thus PGV starts a new path by arbitrarily picking $C1$, thus now $O = \{C1, C3 \to C4 \to C5 \to C6 \to C2\}$. The top two neighbors of C_1 are H and C_2. Since C_2 is in O, only H is appended to C_1. PGV cannot extend H because H is a boundary. In Step 8, a new path is created which is extended in Step 9 to the other chromosome boundary. The preliminary set of consensus ordering is thus $O = \{p_1 = C3 \to C4 \to C5 \to C6 \to C2, p_2 = C1 \to H, p_3 = C7 \to T\}$. At this point, PGV aligns p_1, p_2 and p_3 to the individual genome orderings to decide their orientations. For instance, the alignment score of $H \to C1$ is higher than the alignment score of $C1 \to H$, so p_2 is reversed. Paths p_1 and p_3 are left as is. Then, PGV checks whether p_1, p_2 and p_3 have a good agreement with the input genomes. For instance, the alignment score of p_1 against the five input genomes is 4, 3, 5, 3, and 3, respectively. The total score for p_1 is 18, which is lower than 80% of the highest possible score, which is 0.8*5*5=20. Based on this, PGV considers p_1 not to be a good ordering and it breaks it, as follows. The highest scoring sub-path of p_1 is $C3 \to C4 \to C5 \to C6$ on the majority of the input genomes, so PGV splits p_1 into $p_4 = C3 \to C4 \to C5 \to C6$ and $p_5 = C2$, thus now $O = \{p_2, p_3, p_4, p_5\}$. PGV again checks the alignments of p_2, p_3, p_4, p_5 in O. If any of them is not sufficiently high (i.e., at least 80% of the maximum score), it will be broken again. Once this iterative process is concluded, each path in O is aligned against the genomes and the starting position of its best alignment is recorded. The position with most votes (majority) determines the coordinate of each path. For instance, the best alignment of p_4 on the input genomes are at position 4,3,4,6 and 5, respectively. Thus, p_4 is given coordinate 4. Similarly, PGV assigns p_2 position 1, p_3 position 8 and p_5 position 3. Based on these coordinate, PGV orders the paths as $p_2 \rightarrow p_5 \rightarrow p_4 \rightarrow p_3$ which provides the final consensus ordering $H \rightarrow C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow C_6 \rightarrow C7 \rightarrow T$.


Genome 1 ···CAGTAAAAATATATTTTATCATGTTTTTTACTTATTGAA···
Genome 2 ···CAGTAAAAATATATTTTATCATGTTTTTTACTTATTGAA···
Genome 3 ···TTGCATCCCAGTAAAAATATATTTTATCATGTTTTCTT···
Genome 4 ···TATTTTATCATGCAGTAAAAATTTTTTACTTATTGAAAT···
Genome 5 ···CAGTAAAAATATATGGAAAATTTTTTACTTATTGAAAT···
Multiple Genome Alignment
Genome 1 C1 D1 C2 U1 C3 C4 D2 C5 U2 C6 D3 D4 C7
Genome 2 C1 D1 C3 U3 C4 C5 U4 C2 D5 C6 C7
Genome 3 C1 C2 U5 D5 C6 C5 U6 D3 C4 C3 U7 D4 C7
Genome 4 C1 D1 C5 U8 C6 D3 C2 C3 C4 D2 D4 C7
Genome 5 C1 C6 C2 U9 D5 C3 D2 C4 C5 D3 D4 C7
Block Ordering
Genome 1 C1→D1→C2→U1→C3→C4→D2→C5→U2→C6→D3→D4→C7
Genome 2 C1→D1→C3→U3→C4→C5→U4→C2→D5→C6→C7
Genome 3 C1→C2→U5→D5→C6→C5→U6→D2→C4→C3→U7→D4→C7
Genome 4 C1→D1→C5→U8→C6→D3→C2→C3→C4→D2→D4→C7
Genome 5 C1→C6→C2→U9→D5→C3→D2→C4→C5→D5→D4→C7
(a)
Position 1 2 3 4 5 6 7 8 9
Genome 1 H→C1→C2→C3→C4→C5→C6→C7→T
Genome 2 H→C1→C3→C4→C5→C2→C6→C7→T
Genome 3 H→C1→C2→C6→C5→C4→C3→C7→T
Genome 4 H→C1→C5→C6→C2→C3→C4→C7→T
Genome 5 H→C1→C6→C2→C3→C4→C5→C7→T
Build Consensus
O
C3→C4→C5→C6→C2
C1→H
C7→T
Resolve Misjoins/Orientations
O
C3→C4→C5→C6
C2
C1→H
C7→T
Stitching
Final Output: H→C1→C2→C3→C4→C5→C6→C7→T
(b)

Steps	Operations	Consensus	Neighbors
1	Arbitrarily pick C2	C2	{C6:4} C3:3 C1:2 C5:1
2	Add C6	C6→C2	{C2:4 C5:3} C7:2 C1:1
3	Add C5	C5→C6→C2	{C4:4 C6:3} C1:1 C2: 1 C7: 1
4	Add C4	C4→C5→C6→C2	{C3:5 C5:4} C7: 1
5	Add C3	C3→C4→C5→C6→C2	{C4:5 C2:3} C1: 1 C7: 1
6	Can not extend C3; arbitrarily pick C1	C3→C4→C5→C6→C2 C1	{H: 5 C2:2} C3:1 C5: 1 C6: 1
7	Add H	C3→C4→C5→C6→C2 C1→H	
8	Reach boundary H; arbitrarily pick C7	C3→C4→C5→C6→C2 C1→H C7	{T: 5 C6:2} C3:1 C4: 1 C5: 1
9	Add T	C3→C4→C5→C6→C2 C1→H C7→T	

Figure S1. A detailed example of PGV's processing steps. (a) the input to PGV is a set of $n = 5$ genomes; PGV first carries out a multiple sequence alignment, then classifies each alignment block into core blocks (C), dispensable block (D) and unique block (U); each genome is then converted in an ordered sequence of C-, D-, and U-blocks, each with its corresponding identifier; (b) in the second phase, PGV computes the consensus ordering of the common blocks; red C-nodes are the active nodes; green C-nodes are the neighbors selected to be added to the linear ordering

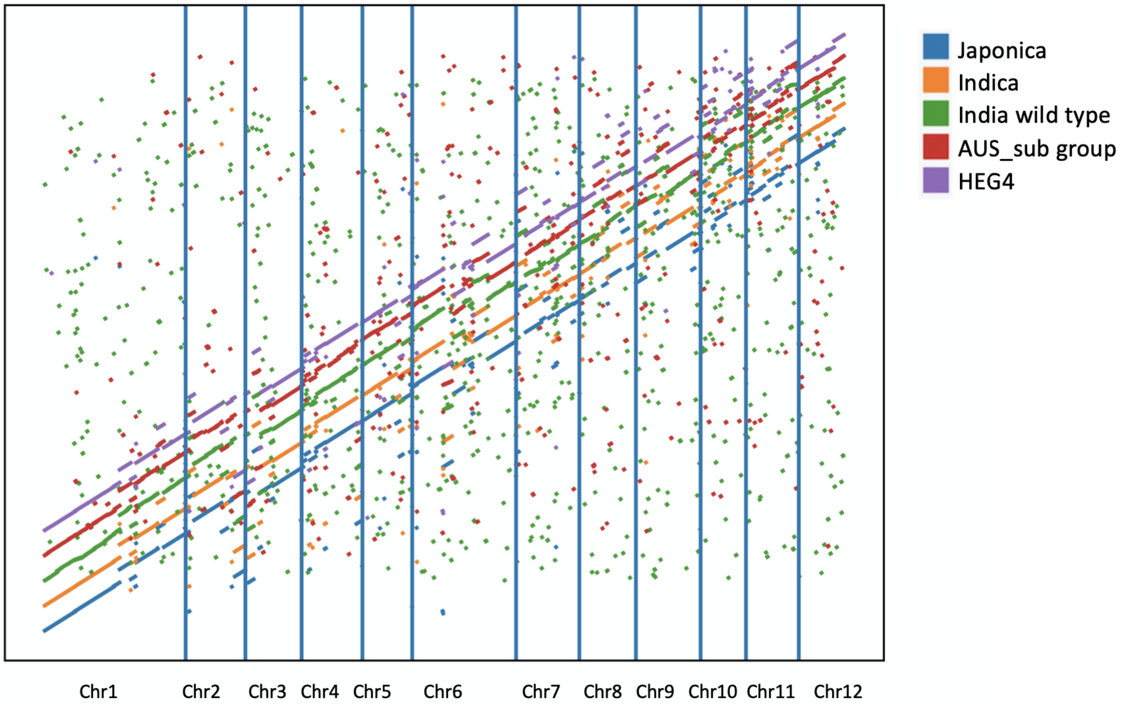


Figure S2. Rice pan-genome analysis using PGV. The x-axis represents the coordinates of the consensus ordering of core blocks computed by PGV. Genomes coordinates for the core blocks are used on the y-axis (staggered to avoid overlapping lines).