Table. Calculation of pairwise sequence similarity/distance statistics at different stages of the alignment parsing procedure

Statistic	Formula	Description
Stage A: Statistics calculated from non-overlapping alignment intervals from each genome of a pairwise comparison		
Coverage breadth genome 1	$\frac{L_{G1}}{\lambda(G1)}$	Proportion of genome 1 covered by non-overlapping alignment intervals
Coverage breadth genome 2	$\frac{L_{G2}}{\lambda(G2)}$	Proportion of genome 2 covered by non-overlapping alignment intervals
Percent identity genome 1	$\frac{N_{G1}}{L_{G1}}$	Number of identical bases across non-overlapping alignments relative to total non-overlapping alignment length, for genome 1 alignment intervals
Percent identity genome 2	$\frac{N_{G2}}{L_{G2}}$	Number of identical bases across non-overlapping alignments relative to total non-overlapping alignment length, for genome 2 alignment intervals
Stage B: Statistics calculated from the subset of alignments from stage A with an alignment interval in both genomes of a pairwise comparison		
Breakpoints	NA	Number of cases where an adjacent pair of alignments in one genome is not adjacent in the same relative order in the other genome
Alignments	NA	Total number of alignments per genome
Alignment pairs	NA	For each aligned sequence, alignment pairs = alignments – 1 (assuming linear sequence topology). e.g. given 4 alignments A,B,C,D; the alignment pairs are A,B; B,C; C,D
Breakpoint distance d0	$\frac{breakpoints}{alignment\ pairs}$	If the denominator is 0 (1 alignment), breakpoint distance is assigned 0
Breakpoint distance d1	$\frac{breakpoints}{L/1000}$	Breakpoints expressed per kilobase of aligned sequence
L50	NA	After ordering alignments from large to small, L50 is the number of alignments that must be cumulatively summed to reach 50% of total alignment length
N50	NA	As above, but the size of the alignment at the 50% quantile is given
Stage C: Statistics calculated from 'trimmed' alignment intervals (for each alignment from stage B, the shorter alignment interval from query/subject genome is selected)		
Coverage breadth (trimmed)	$\frac{L}{\lambda(G)/2}$	Total trimmed alignment length relative to mean genome length
Coverage breadth of smaller genome (trimmed)	$\frac{L}{\lambda_{min}(G)}$	Total trimmed alignment length relative to length of the smaller genome
Distance score d0	$1 - \frac{L}{\lambda(G)/2}$	Distance metric of trimmed coverage breadth

Distance score d1	$1 - \frac{L}{\lambda_{min}(G)}$	Distance metric of trimmed coverage breadth of the smaller genome
Distance score d2	$-log\ \frac{L}{\lambda(G)/2}$	Rescaled variant of distance score d0
Distance score d3	$-log\ \frac{L}{\lambda_{min}(G)}$	Rescaled variant of distance score d1
Average nucleotide identity	$\frac{N}{L}$	Number of identical bases across trimmed alignments relative to total trimmed alignment length
Distance score d4	$1 - \frac{N}{L}$	Distance metric of average nucleotide identity
Distance score d5	$-log\ \frac{N}{L}$	Rescaled variant of distance score d4
Distance score d6	$1 - \frac{N}{\lambda(G)/2}$	Number of identical bases across trimmed alignments relative to mean genome length
Distance score d7	$1 - \frac{N}{\lambda_{min}(G)}$	Number of identical bases across trimmed alignments relative to length of the smaller genome
Distance score d8	$-log\ \frac{N}{\lambda(G)}$	Rescaled variant of distance score d6
Distance score d9	$-log\ \frac{N}{\lambda_{min}(G)}$	Rescaled variant of distance score d7

- Genome 1 (G1) and genome 2 (G2) are the query/subject genomes re-ordered alphabetically.
- λ(G1) and λ(G2) are the lengths of genome 1 and genome 2, respectively.
- L_{G1} and L_{G2} are the total length of non-overlapping alignment intervals covering genome 1 and genome 2, respectively.
- N_{G1} and N_{G2} are the total number of identical bases across non-overlapping alignment intervals in genome 1 and genome 2, respectively.
- λ(G)/2 is the mean genome length, and $\lambda_{min}(G)$ is the length of the smaller genome of the pair.
- L is the total length of trimmed alignment intervals.
- N is the total number of identical bases across trimmed alignment intervals.

Note that if bidirectional BLAST is run, alignment length statistics (L_{G1}, L_{G2}, L), identical base statistics (N_{G1}, N_{G2}, N), and structural similarity/alignment contiguity statistics (breakpoints, alignments, alignment pairs, L50, N50) are calculated as a mean of the statistics obtained from both directions.

The different distance scores reflect different distance concepts. For example, distance scores d6 and d7 represent resemblance and containment respectively.