Additional information on haplotype analysis and visualization

This analysis method for the identification of homologous haplotypes is both simple and computationally efficient.

Sequences were windowed and stepped according to user specifications. Next, a hierarchical agglomerative clustering model is built using Ward distance. Sequences in a window are then clustered together into homologous haplotypes if their pairwise distance falls under a distance threshold.

The program determines the distance threshold, d, by iterating through a user-specified range and identifying the value of d that maximizes the mean silhouette coefficient for all sequences, thus allowing the clustering algorithm to account for unequal information content across genomic windows. The code is built on Scikit-learn's clustering module and is available for download as a Python3 command-line script.

The optimal range for d depends on the size of the genomic window under consideration. Because the genetic information content and mean pairwise distance among samples increases along with window size, the optimal value of d that returns the maximal number of meaningful clusters will also. A value of d that is too small for a genomic window may cause homologous haplotypes to be split into separate clusters because of minor genotyping errors or residual heterozygosity in seed lots. A d value that is too large can result in the clustering of divergent haplotypes. Visualization of d values returned by the analysis gives insight into the chosen d range, with an ideal d distribution, in our experience, being uniform to somewhat left-skewed.

We evaluated our haplotype analysis methodology by detecting introgressions in the well-characterized breeding line CU17NBL that is known to have six S. pennellii LA0716 introgressions across five chromosomes (MA Mutschler, unpublished).

Using the cluster data from our windowed analysis, we defined LA0716 introgressions as those where the haplotypes from breeding line CU17NBL clustered with LA0716 but not with two cultivated breeding lines without LA0716 introgressions: M82 and NC 1 CELBR. With a 100 Kb window size, 25 Kb step size, and d range of {2-80}, the clustering algorithm identified 12 putative LA0716 introgressions, which is greater than the true number of introgressions.

All distance thresholds returned for the erroneous windows fell in the top 10% of the d distribution, indicating little distinguishing genetic information in these windows. Seven putative introgressions were detected, including all six known introgressions, when the input parameters were changed to a 250 Kb window size, 100 Kb step size, and d range of {2-100}. With a window size of 500 Kb or greater, a step size of 100 Kb, and d range {20-200}, we detected all but the smallest known introgression on chromosome 7. Thus, larger window sizes showed decreased sensitivity to the identification of smaller introgressions, but greater accuracy in the detection of known introgressions.