Evaluation of Clustering Algorithms for Inferring Inversion Genotypes

Matthew Aleck[1], Ronald J. Nowling[1]

[1] Electrical Engineering and Computer Science, Milwaukee School of Engineering, Milwaukee, WI USA

Genotypes of large polymorphic inversions can be inferred from single nucleotide polymorphism (SNP) data. Recombination is repressed in inversion regions, allowing alleles to remain private to a single inversion orientation. For large inversions, there are sufficient numbers of private alleles such that samples segregate according to inversion genotypes.

In principal component analysis (PCA), samples form groups according to their inversion genotypes. Clustering algorithms can be used to determine the distinct groups and their members. Depending on the particular data set, however, inversions cause different patterns. For example, *D. melanogaster* samples form three distinct, circular clusters with only a handful of outliers when analyzing SNPs on the 2L chromosome arm, while the origin separates *An. gambiae* samples by their 2Rb inversion genotypes, but the clusters are not clearly defined since the intra- and inter-cluster variances are nearly equal.

Clustering algorithms use a range of heuristics to infer cluster membership. For example, k-means partitions the global space into regions of approximately equal size, while DBSCAN uses the local topology of points to infer clusters. The differences in choices make each clustering algorithm suitable to a different type of data. For example, k-means is most appropriate for evenly sized, circular clusters, while DBSCAN works on complex shapes with uneven sizes as long as inter-cluster distances are less than intra-cluster distances.

Our aim was to characterize the different cluster patterns caused by large polymorphic inversions and determine which clustering algorithm is best suited for each type of pattern. We performed PCA of SNPs from chromosome arms of *An. gambiae*, *An. coluzzii*, and *D. melanogaster*. We developed a set of descriptive criteria which we applied to characterize the observed patterns for each inversion. We then applied several clustering algorithms to the data sets and evaluated their predictions against the known genotypes for the samples. We use our characterizations and clustering results to recommend the most appropriate clustering algorithms for each type of pattern.