

Clustering for Forensic Mitotype Quality Analysis

Introduction

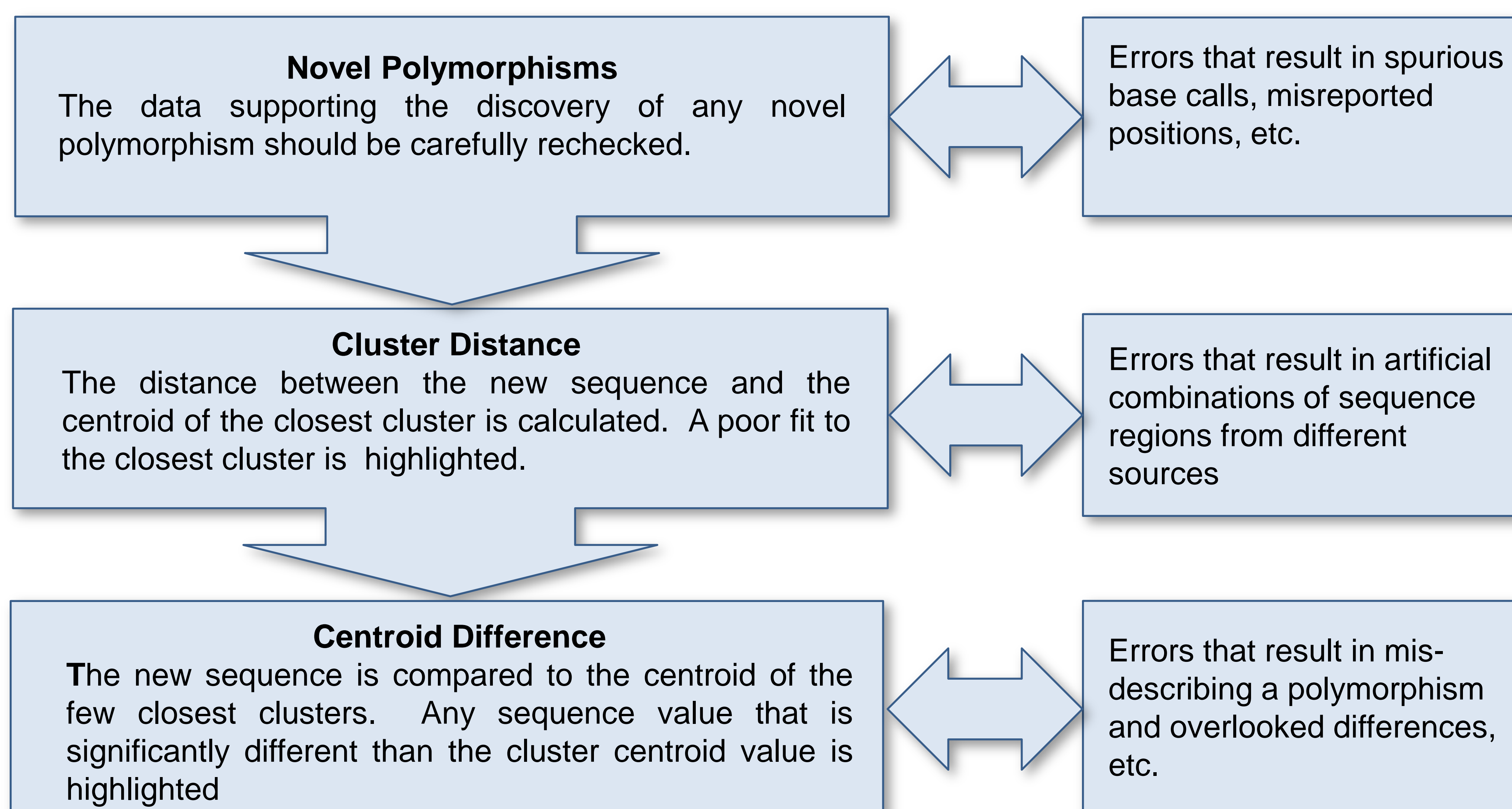
Phylogenetic clustering analysis has been used quite successfully to detect errors in mtDNA data by identifying unusual occurrences of sequence variation (e.g. Bandelt *et. al.* Int. J. Legal Med. 2001). Errors can affect individual polymorphisms in the data - which could be missed, misdescribed, mislocated, or spurious - and "artificial recombination" errors which result from mistakenly reporting sequence regions from different sources in the same sample consensus.

For quality analysis purposes, the hierarchical structure created by phylogenetic clustering algorithms is unnecessary and may confound the discovery of alternate patterns in narrower regions of the mtDNA genome. In this work, an approach to routine data quality review is described that employs standard statistical cluster analysis to identify unusual polymorphisms and sequence regions that may warrant further attention by the analyst. This approach is easily automatable and can be tailored to targeted population groups in individual investigations.

Methods:

- Binary descriptor vector: 0 or 1 for every observed unique polymorphism at 421 positions – 1010 bits.
- Standard k-means clustering with Euclidian distance. Other methods of clustering can also be employed.
- 2622 types from EMPOP and 3552 types from SWGDAM.

Data Review Process



Clusters

In this simple example, the 5212 unique mitotypes in the combined data sets were clustered into 200 randomly-seeded clusters using k-means clustering with a Euclidian distance. Adequate separation of the clusters was observed. In Figure 1, a box plot of the size of the clusters is shown on the bottom. On top of Figure 1 is a box plot of distance from each type to the next-closest cluster. In this summary view, the separation of the distributions shows that cluster membership can be used to identify one or two unusual polymorphisms by comparison on to the cluster's centroid.

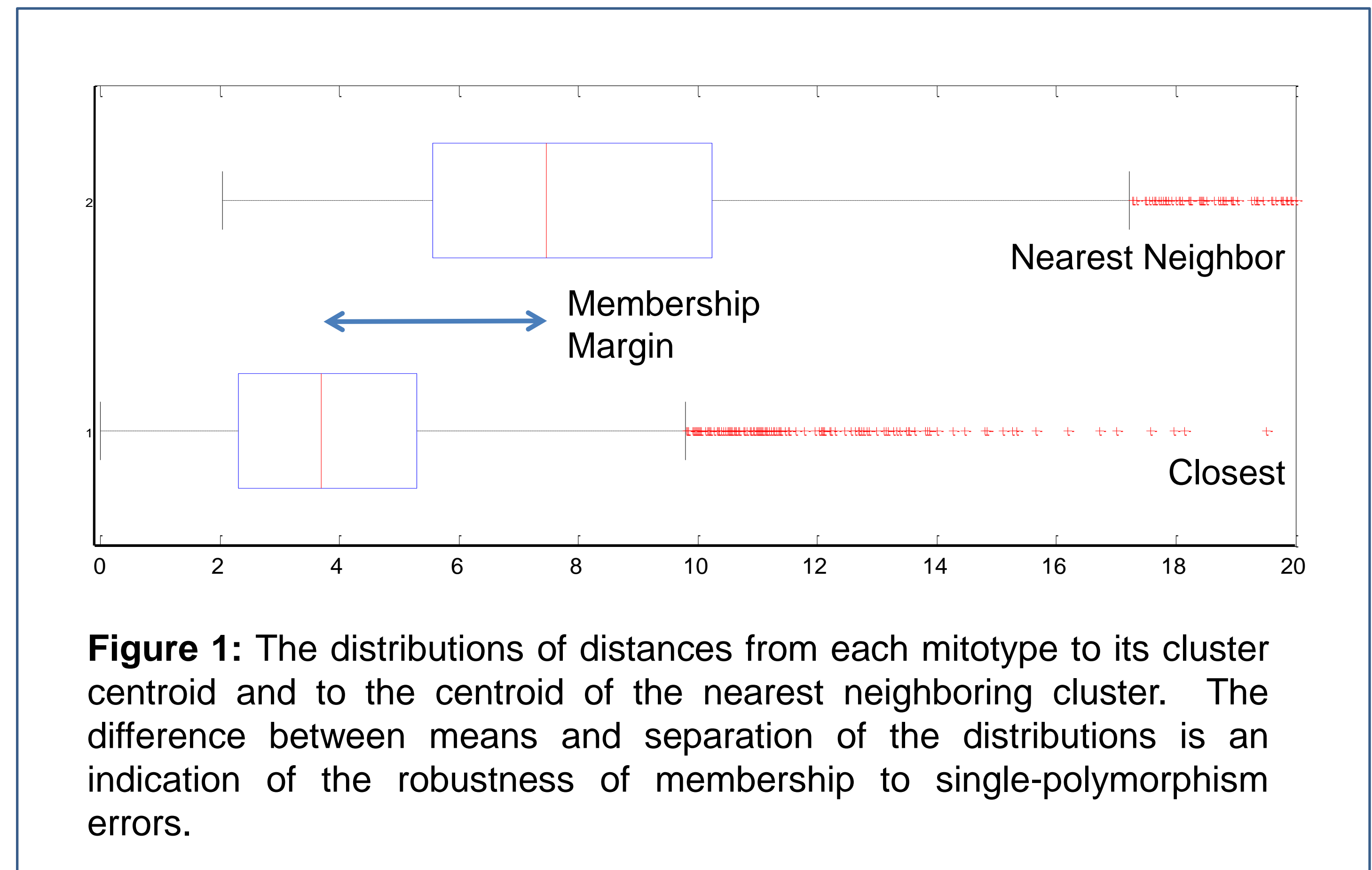


Figure 1: The distributions of distances from each mitotype to its cluster centroid and to the centroid of the nearest neighboring cluster. The difference between means and separation of the distributions is an indication of the robustness of membership to single-polymorphism errors.

Conclusions:

A straightforward procedure for routine quality review is described that employs cluster membership and difference to identify unusual mitotypes and individual polymorphisms that warrant review by the analyst.

The review process can be tailored to the specific investigation by clustering databases that are representative of the investigations. In this example, two diverse databases were combined for clustering, which is appropriate when no target population for the sequence being reviewed can be predicted.

Case Study: Novel Polymorphism and Poor Fit

In an earlier version of the SWGDAM database, the sequence of USA.HIS.000204 was found to contain two transcriptional errors. Table 1 shows the results presented to the analyst when this original sequence is reviewed. The two errors are highlighted as novel polymorphisms and the other significant differences between the centroids of the closest two clusters and the original USA.HIS.000204 sequences are shown.

Table 1	204	Clust 157	Clust 109
Site	Value		
315.1-	1	1.00	0.00
315.1C	0	0.00	1.00
153.0A	0	0.00	0.07
153.0C	1	Not Found	Not Found
153.0G	0	1.00	0.93
235.0C	1	Not Found	Not Found
235.0G	0	0.50	1.00
16182.0A	0	0.75	1.00
16182.0C	1	0.25	0.00
16183.0A	0	0.75	0.79
16183.0C	1	0.25	0.21
16335.0A	1	0.25	0.93
16335.0G	0	0.75	0.07
317.0-	1	0.25	0.00
317.0C	0	0.75	1.00
316.0-	1	0.25	0.00
316.0G	0	0.75	1.00
310.0-	1	0.50	0.00
310.0T	0	0.00	1.00
16189.0T	0	0.75	0.00

Case Study: Artificial Recombination

In an earlier version of the SWGDAM database, the sequence of sample USA.AFR.000942 was found to be an artificial combination of sequences from two sources and corrected. The Cluster Distance check clearly highlights the original sequence as an outlier as shown in Figure 2 (right).

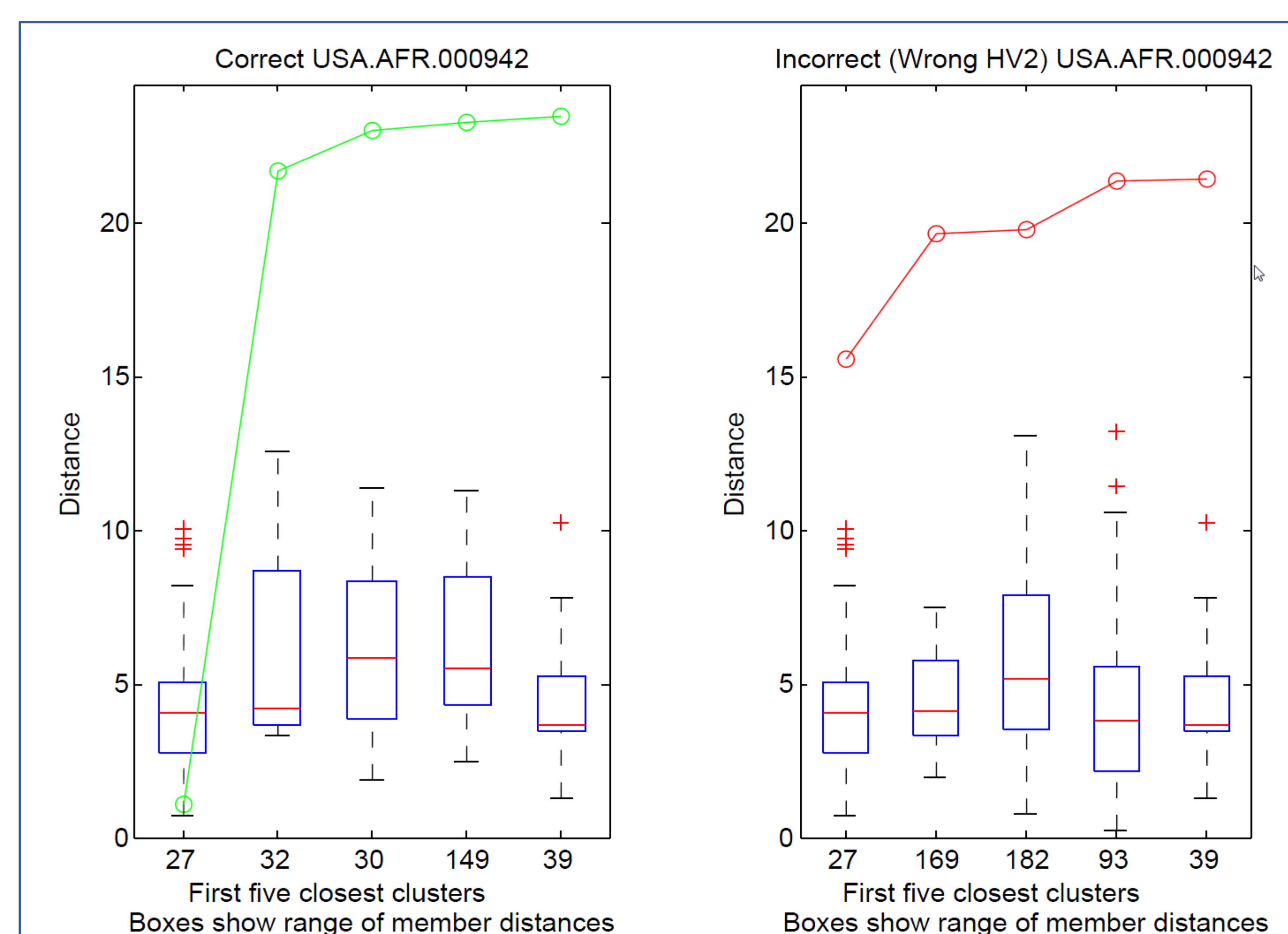


Figure 2: The distance of the original and corrected USA.AFR.00942 sample sequence from the five nearest clusters. The original erroneous sequence is a clear outlier from all five of the nearest clusters (right) and so is easily identified as suspicious. Once corrected, the sequence is a good fit within the closest cluster (left).