SUPPLEMENTAL FIGURE 2

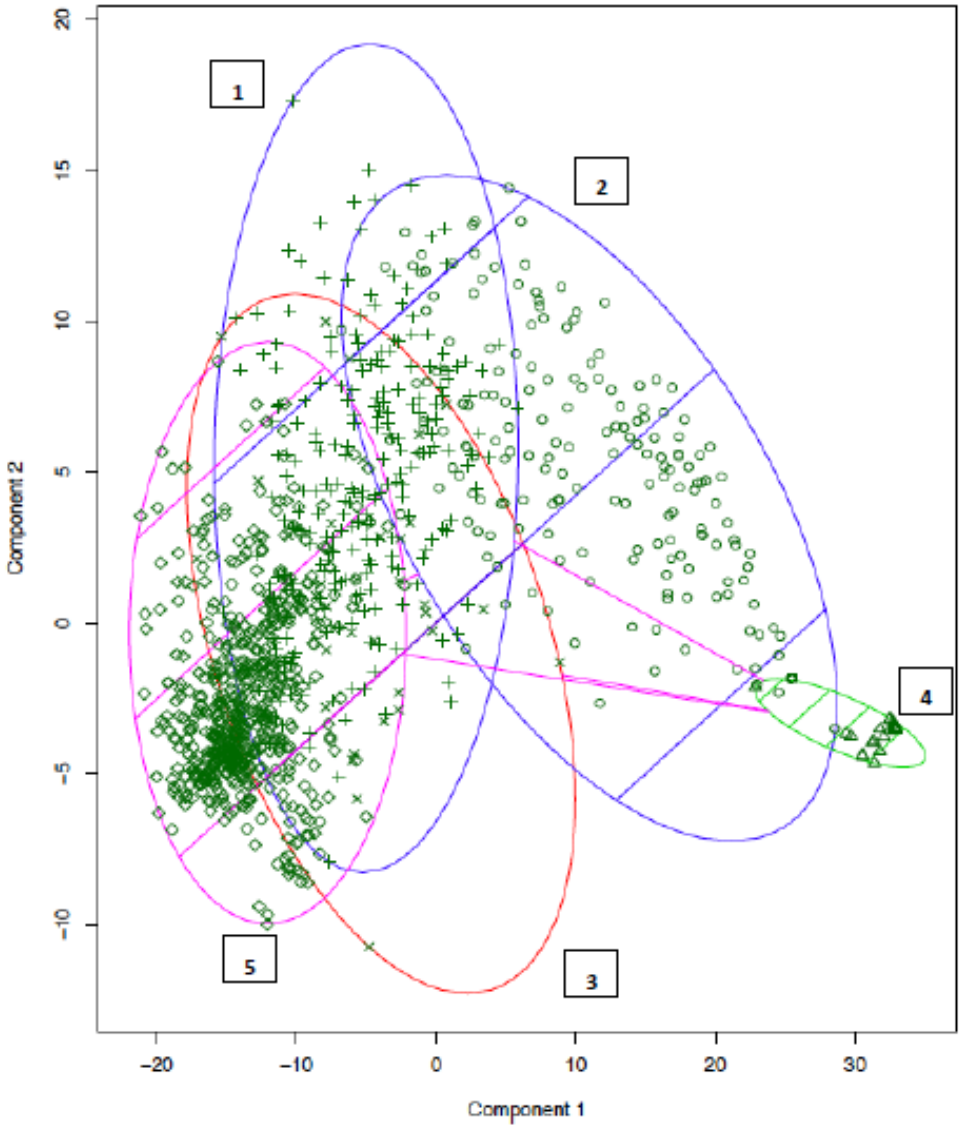

Figure S2. k-means 5 clusters plot. The numbers in the boxes refer to the 5 clusters.

All 1,222 included sequence IDs grouped based on all 519 included variables after cleaning the data. Due to the high number of variables included, heterogeneity is challenging to plot on 3 axes (x-, y-, and z-axes); therefore, the distance cannot be accurately assessed via this diagram. The lines between clusters indicate 3-dimensional distance between clusters. Based on measures of sum of squares, in-between cluster variability for the 5 clusters model was found to be 63.08%. This value represented the highest variability calculated among all the suggested k-values in the elbow diagram in Figure S1. Components 1 and 2 account for the greatest variance attributed to variables.

Euc.dist <- function (x1, x2) sqrt (sum(x1-x2) ^2; where sqrt is square root measure. This is repeated multiple times for all data points. Based on all those distances, for the 1,222 participants, the model finds the optimal centroid (beginning with an assumption) to fit the k-5 model. This is a direct explanation of it being an unsupervised learning method.