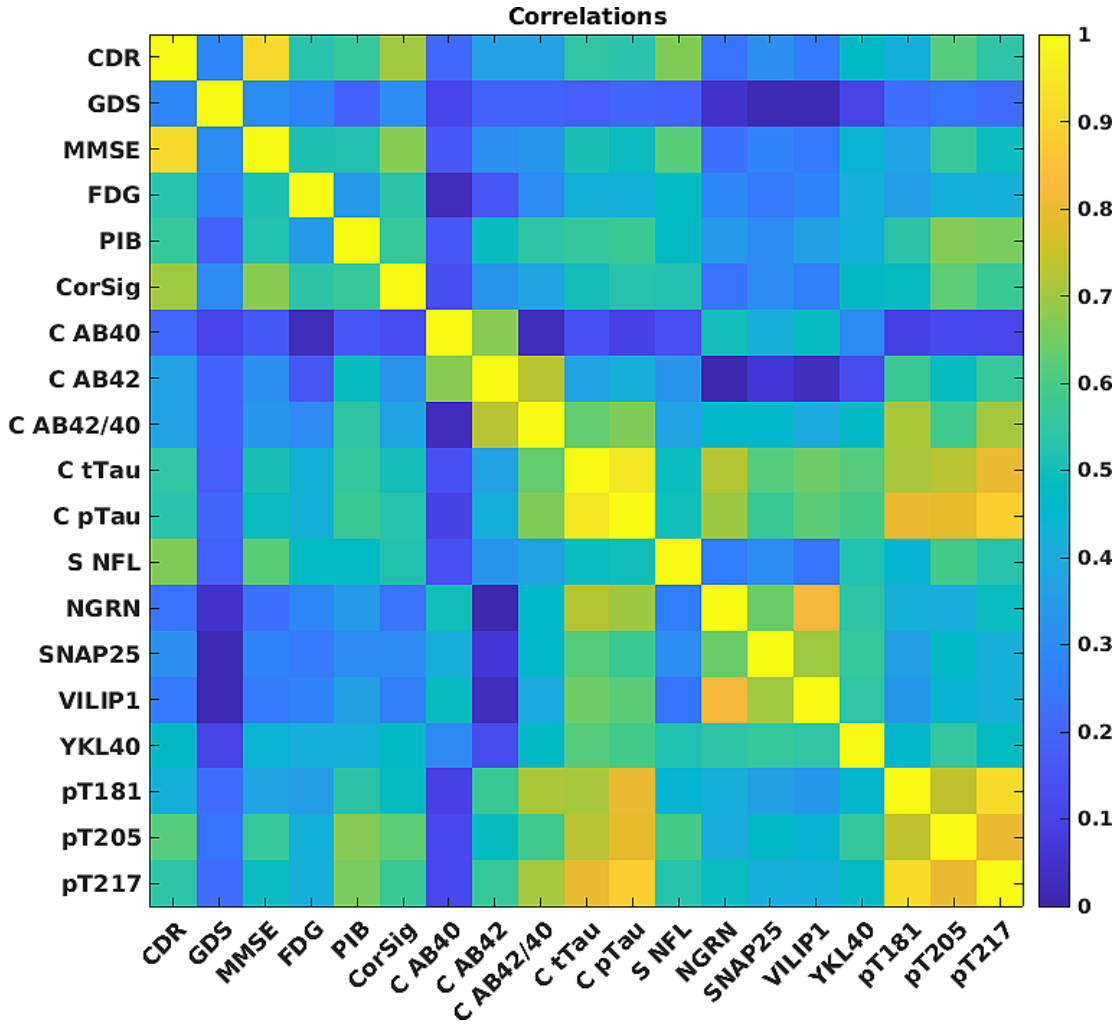
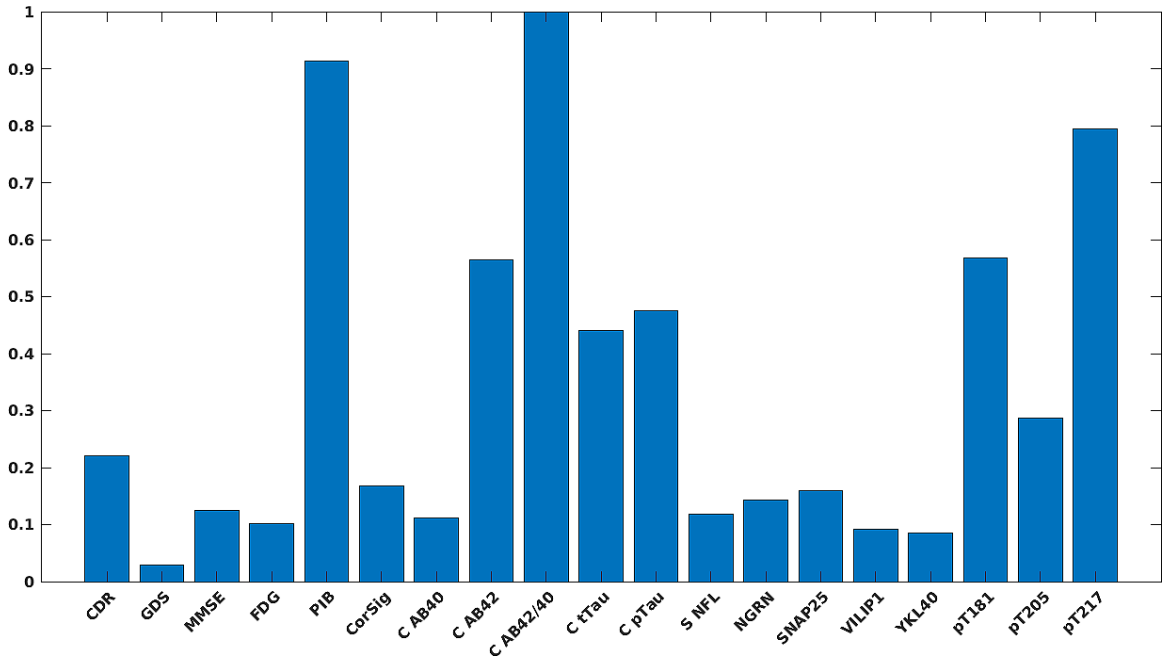
Cluster Validation

Cluster stability and variance were evaluated using the elbow method[1] with both distortion[2] and principal component analysis (PCA). The elbow method relies on identifying cluster solutions based on a minimum amount of explained variance and a point of diminishing return (i.e. at what point does adding an additional cluster no longer provide a meaningful increase in the explained variance). The method provides cluster solutions that yield a balance between under fitting and over fitting a model. This method was applied using both PCA and distortion (sum of squared distances from each point to its assigned center) as variance metrics. Further, to assess the stability of the results, the elbow method using PCA and distortion was calculated for 1) 2000 random permutations of 80% of the clustering feature space (corresponding to removing 20% of participants from the data set), and 2) 18 leave-one-out partitions (corresponding to removing a single biomarker from each partition). The mean and standard deviation were then calculated for each of the four analysis (see Supplementary Figure 6).

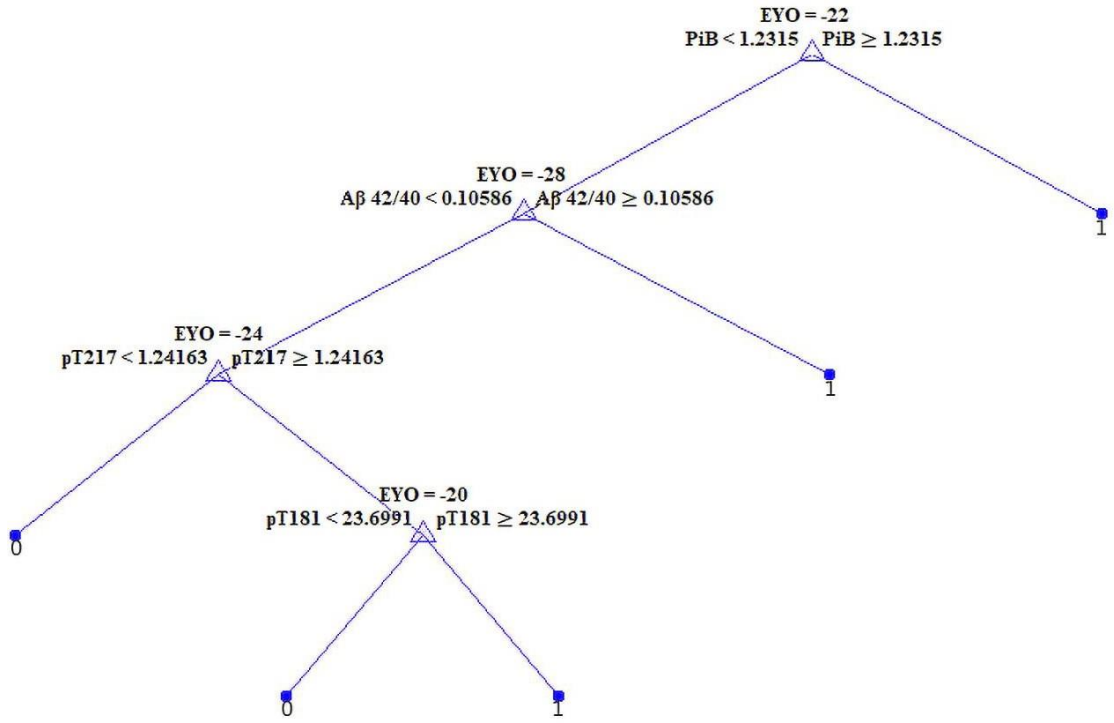
Supplemental Figures



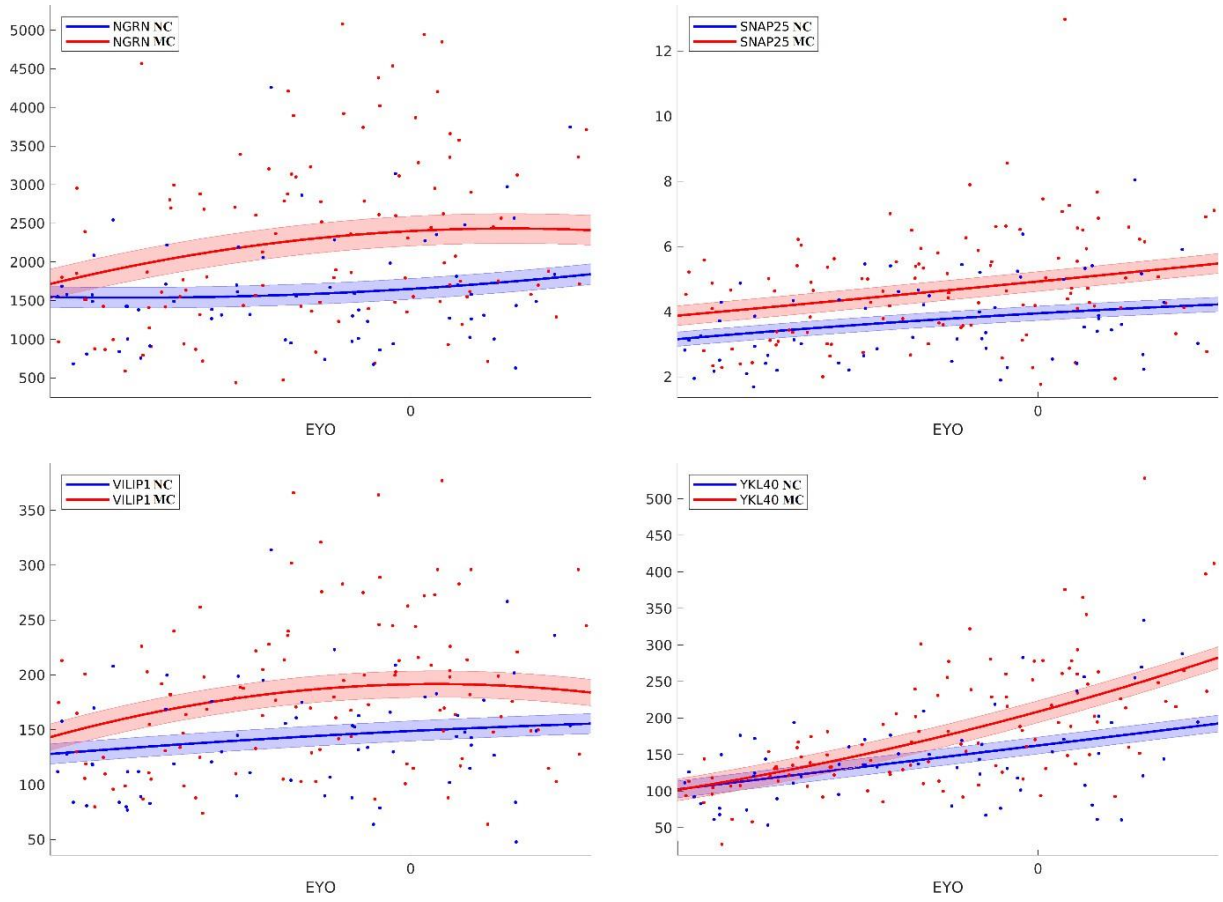
Supplemental Figure 1. Similarity matrix used for clustering of biomarkers. Values were calculated by taking the absolute values of the Pearson correlations. Three main clusters were seen with markers of neurodegeneration (CDR, FDG, CorSig) forming one group, markers of amyloid and tau (CSF AB42, CSF t-tau, CSF p-tau) forming another group, and third group of emerging markers of neuronal integrity and inflammation (serum NFL, CSF SNAP25, VILIP 1, YKL40). These clusters were subsequently shown in Figure 1.



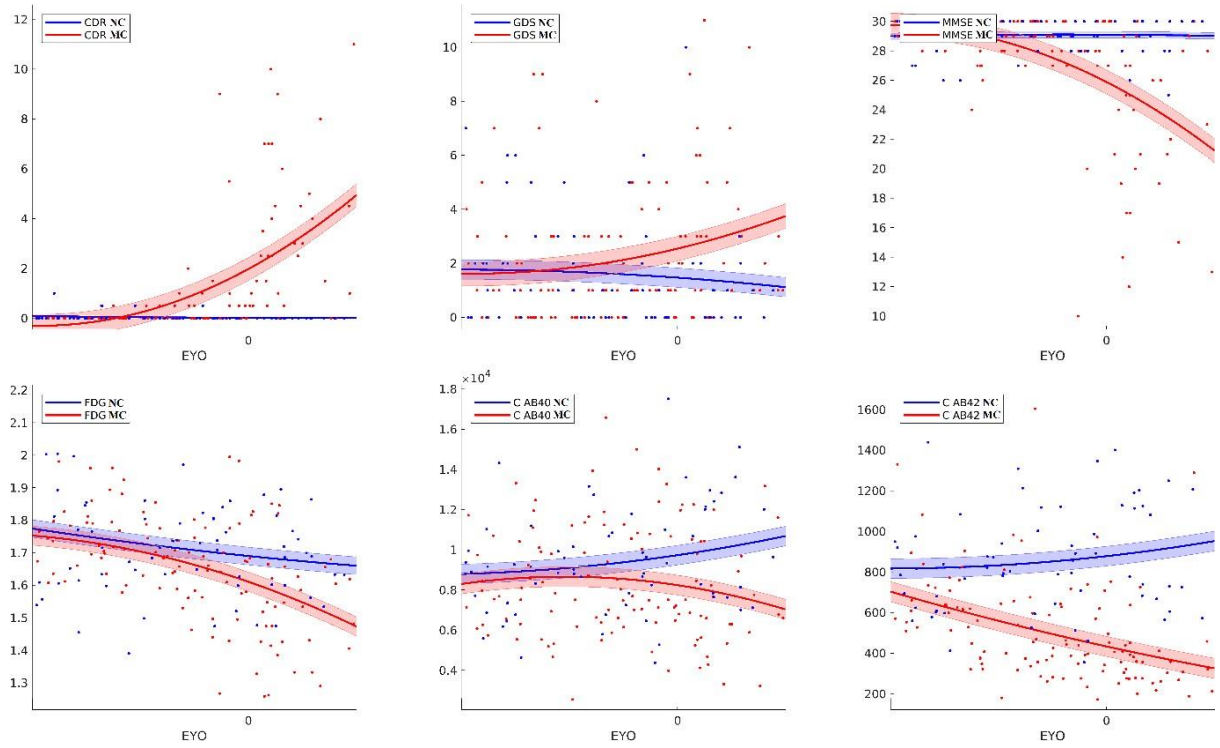
Supplemental Figure 2. The strongest predictors of mutation status across EYO. These results correspond with those shown in Figure 2. The strongest predictors for MC were markers of amyloid (CSF Aβ42/40, PiB) and tau (pTau 217).



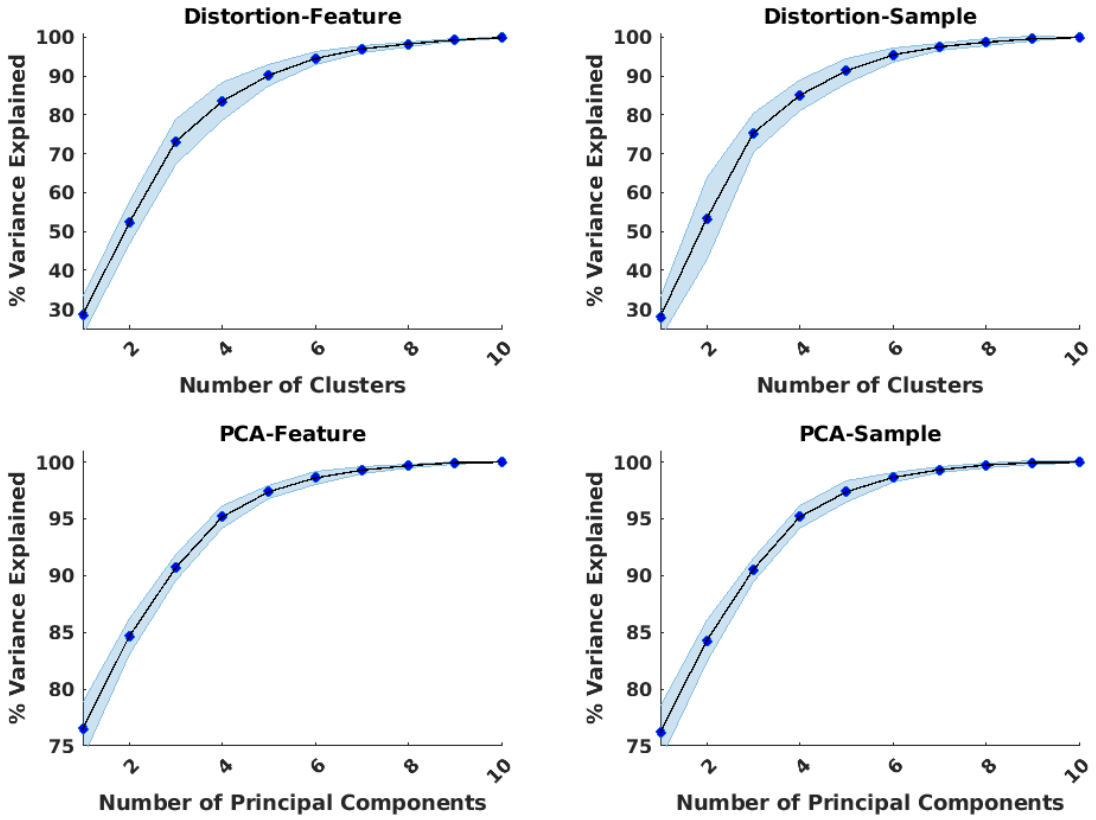
Supplementary Figure 3. Optimal decision tree for classifying mutation status. PiB, CSF Aβ42/40, CSF pT217, and CSF pT181 gave the optimal results for the model. At the top of the tree, the optimal SUVR cutoff for PiB was 1.2315. If PiB was less than this cutoff, CSF Aβ42/40 was then included in the evaluation. A participant was classified as a MC if the CSF Aβ42/40 was greater than .105861. If CSF Aβ42/40 was less than this value, CSF pT217 was included using a cutoff of 1.24163. If a participant had a CSF pT217 value greater than 1.24163 and a CSF pT181 value greater than 23.6991 then (s)he was classified as MC. EYO values were identified by associating the cutoffs with the polynomial fit curves shown in Figures 4 and 5.



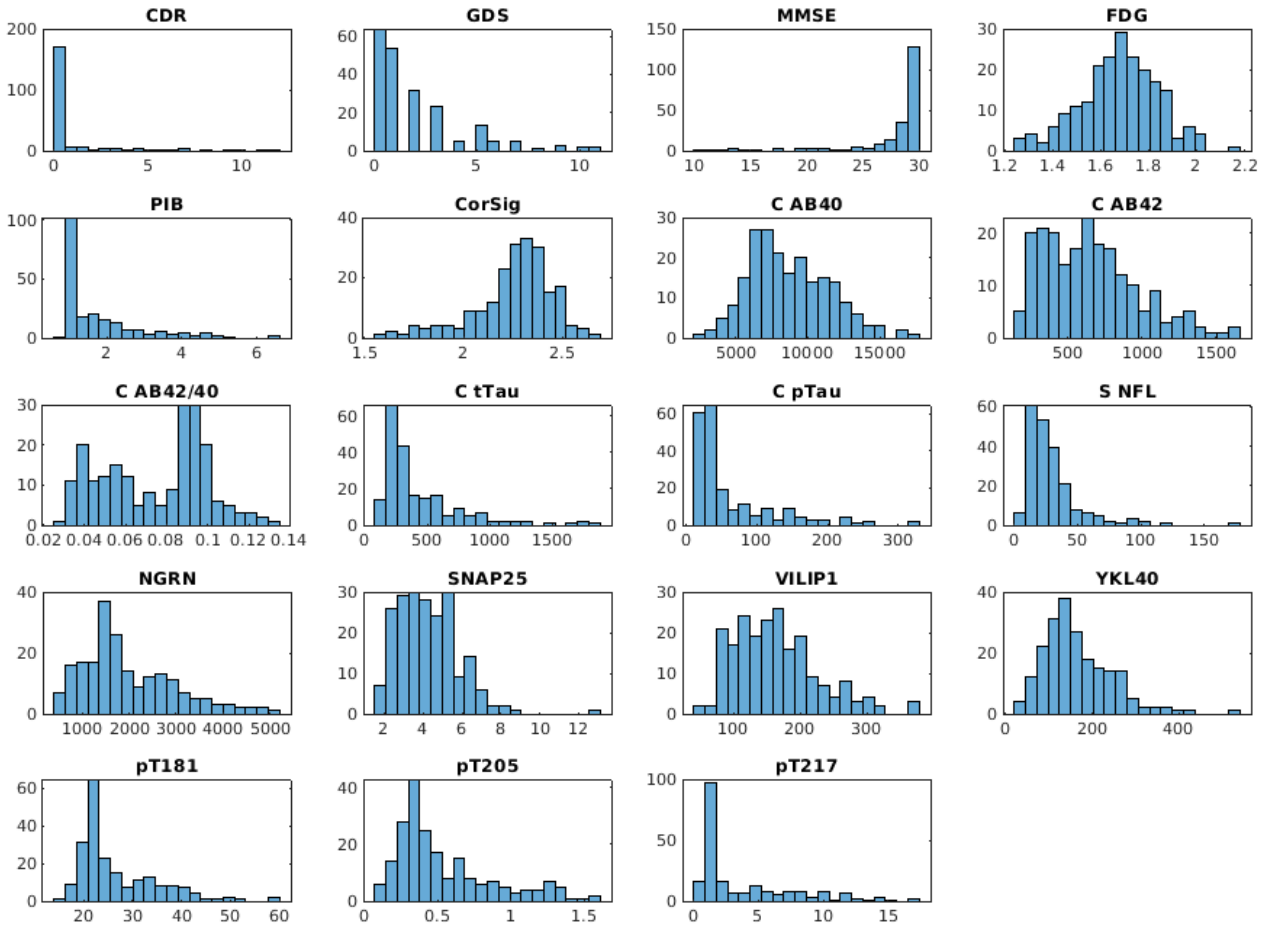
Supplementary Figure 4. Trajectories for emerging biomarkers (NGRN, SMAP25, VILIP1, and YKL40) for NC (blue) and MC (red) from -20 to +10 EYO. Trajectories were fit using a 2 degree polynomial with shaded regions representing standard error. Both CSF NGRN and VILIP-1 showed inverted U shapes curves for MCs but not for NCs. YKL40 increased with age for both MCs and NCs suggesting a possible aging effect.



Supplementary Figure 5. Trajectories for emerging cognitive and clinical measures for NC (blue) and MC (red) from -20 to +10 EYO. Trajectories were fit using a 2-degree polynomial with shaded regions representing standard error. CDR-SB was the strongest clinical predictor of both mutation status and EYO (See Figures 2 and 3) with changes seen at -10 EYO. Changes in GDS were seen close to EYO (-5 EYO).



Supplementary Figure 6. Elbow method for cluster evaluation and identification of the optimal number of clusters. Top left shows the mean (solid line) and 2 standard deviations (shaded region) of cluster variance based on distortion over 2000 random permutations of 80% of the clustering feature space (each feature is 1 participant). Top right shows the mean and 2 standard deviations of cluster variance based on distortion with leave-one-out partitions of the clustering sample space (each sample is 1 biomarker). Similar results for bottom left and right panels using principal component analysis. These results indicate a minimum of 3 to 5 clusters are needed to explain 90% of variance.



Supplementary Figure 7. Distribution of biomarkers

1. Nainggolan R, Perangin-angin R, Simarmata E, et al. Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the elbow method. In: *Journal of Physics: Conference Series*. Vol 1361. IOP Publishing; 2019:12015.
2. Sugar CA, James GM. Finding the number of clusters in a dataset: An information-theoretic approach. *J Am Stat Assoc*. 2003;98(463):750-763.