

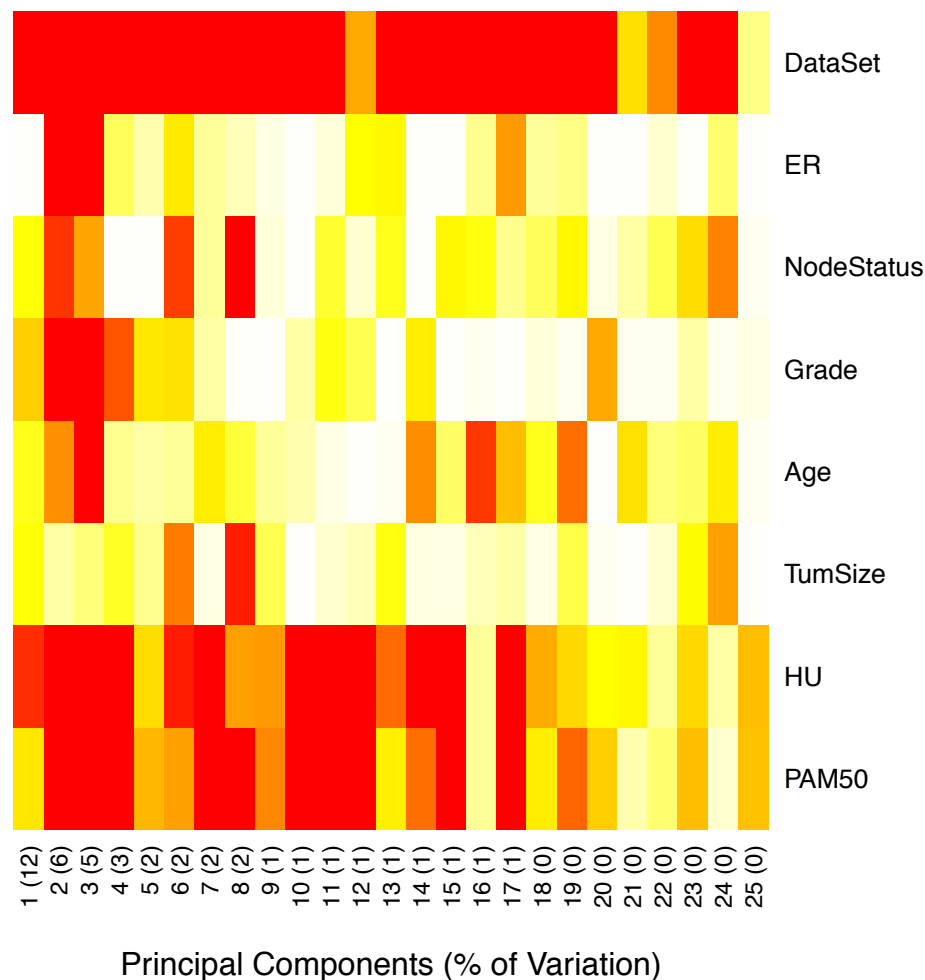
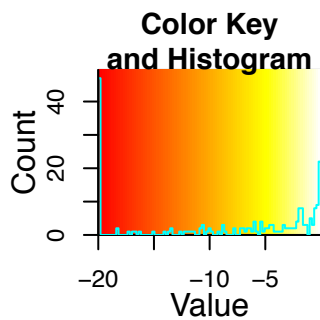


Figure S1. PCA analysis to investigate associations between variation in the 1881-sample gene expression data set merged on Entrez Gene ID and various sample annotations. The principal components are shown horizontally, with the fraction of the variation associated with each principal component shown within parenthesis. The evaluated sample annotations are shown vertically. Each principal component was associated to each sample annotation using a linear model fit. The linear model fits were evaluated using the F-statistic and log10 of the P-values are pseudocolored in the plot. Only the first 25 principal components are shown.

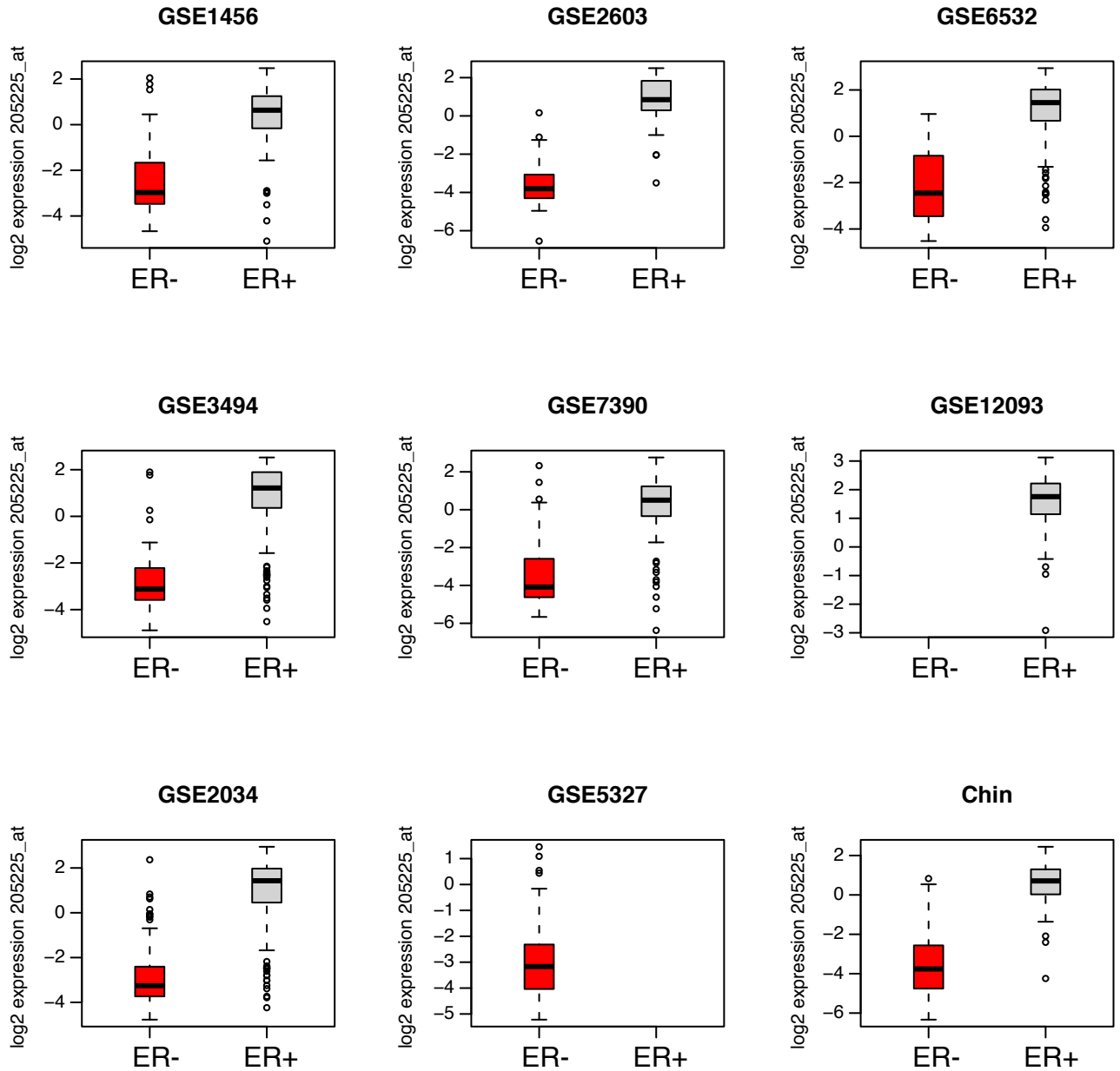


Figure S2. Log2 gene expression values for Affymetrix probe 205225_at mapping to *ESR1* across nine different data sets with ER-status included in the pooled 1881-data set. Affymetrix probe 205225_at have been reported to have the highest concordance with biochemical ER-status. (Karn et al. [21]).

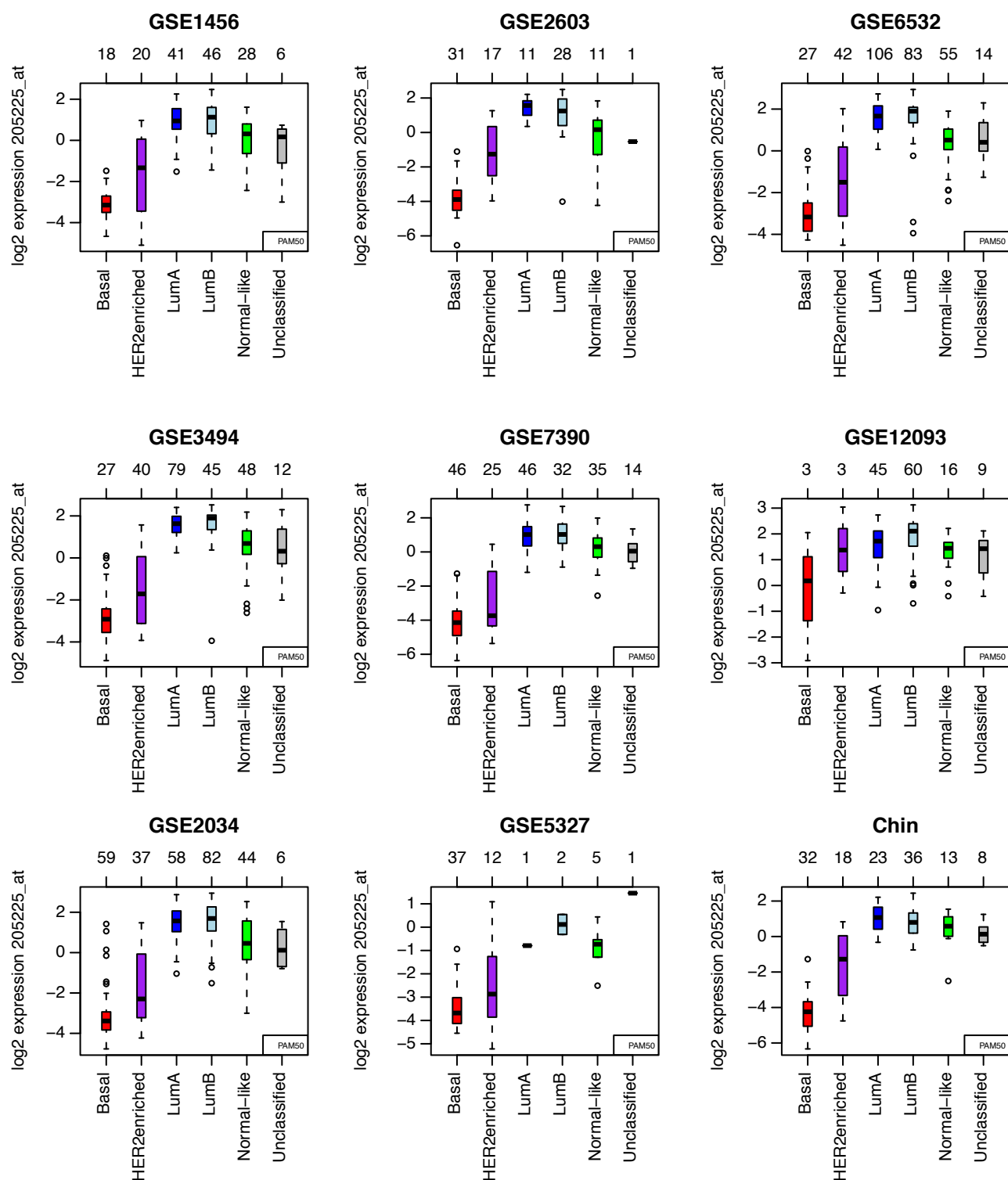


Figure S3. Log2 gene expression values for Affymetrix probe 205225_at mapping to *ESR1* across nine different data sets included in the pooled 1881-data set. Tumors are grouped according to PAM50 molecular subtype classification [26]. Affymetrix probe 205225_at have been reported to have the highest concordance with biochemical ER-status. (Karn et al. [21]).

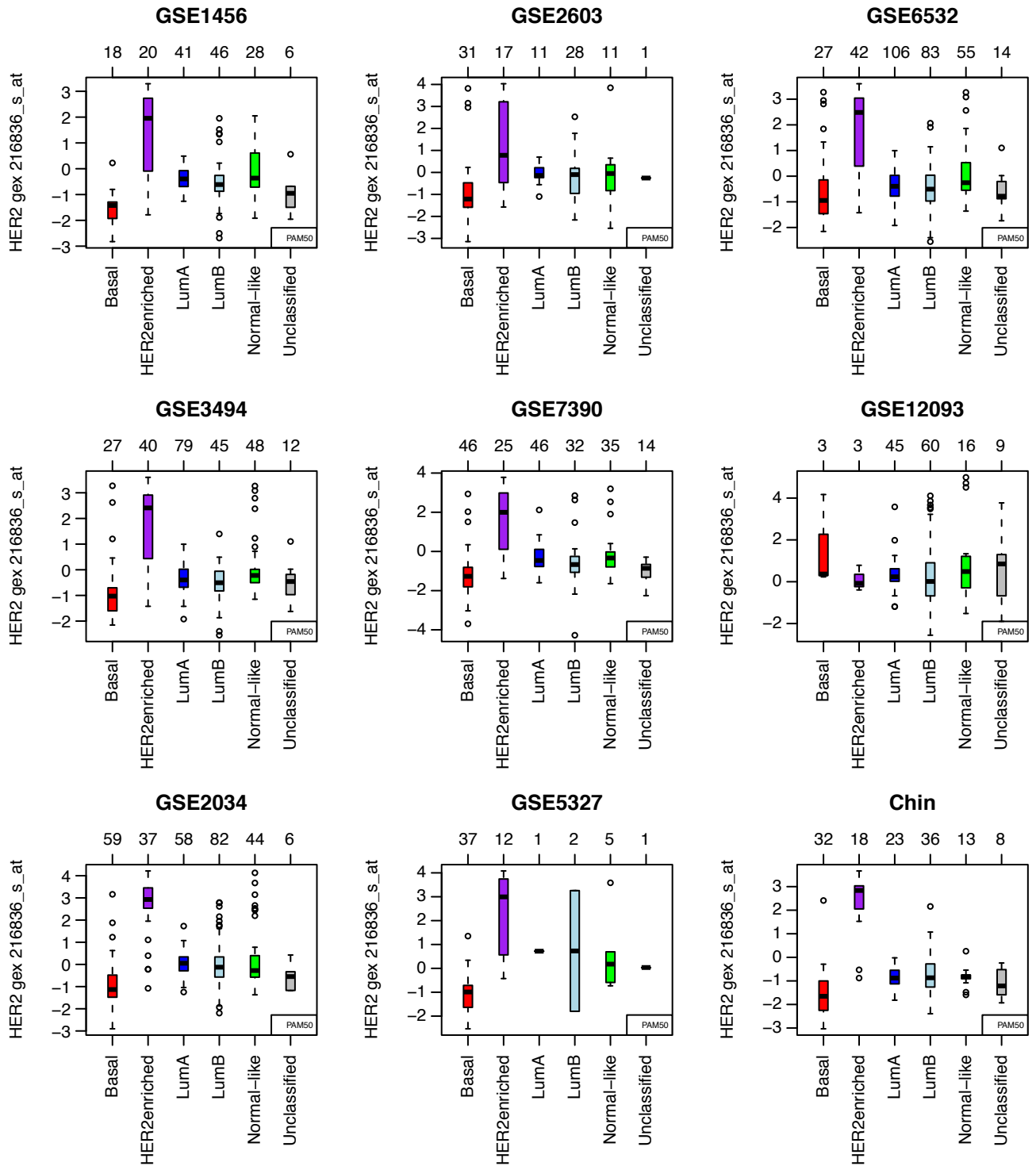


Figure S4. Log2 gene expression values for Affymetrix probe 216836_s_at mapping to *HER2* across nine different data sets included in the pooled 1881-data set. Tumors are grouped according to PAM50 molecular subtype classification [26]. Affymetrix probe 216836_s_at was used by Karn et al. [21] to derive data driven cut-offs from gene expression data for clinical subtype classification.

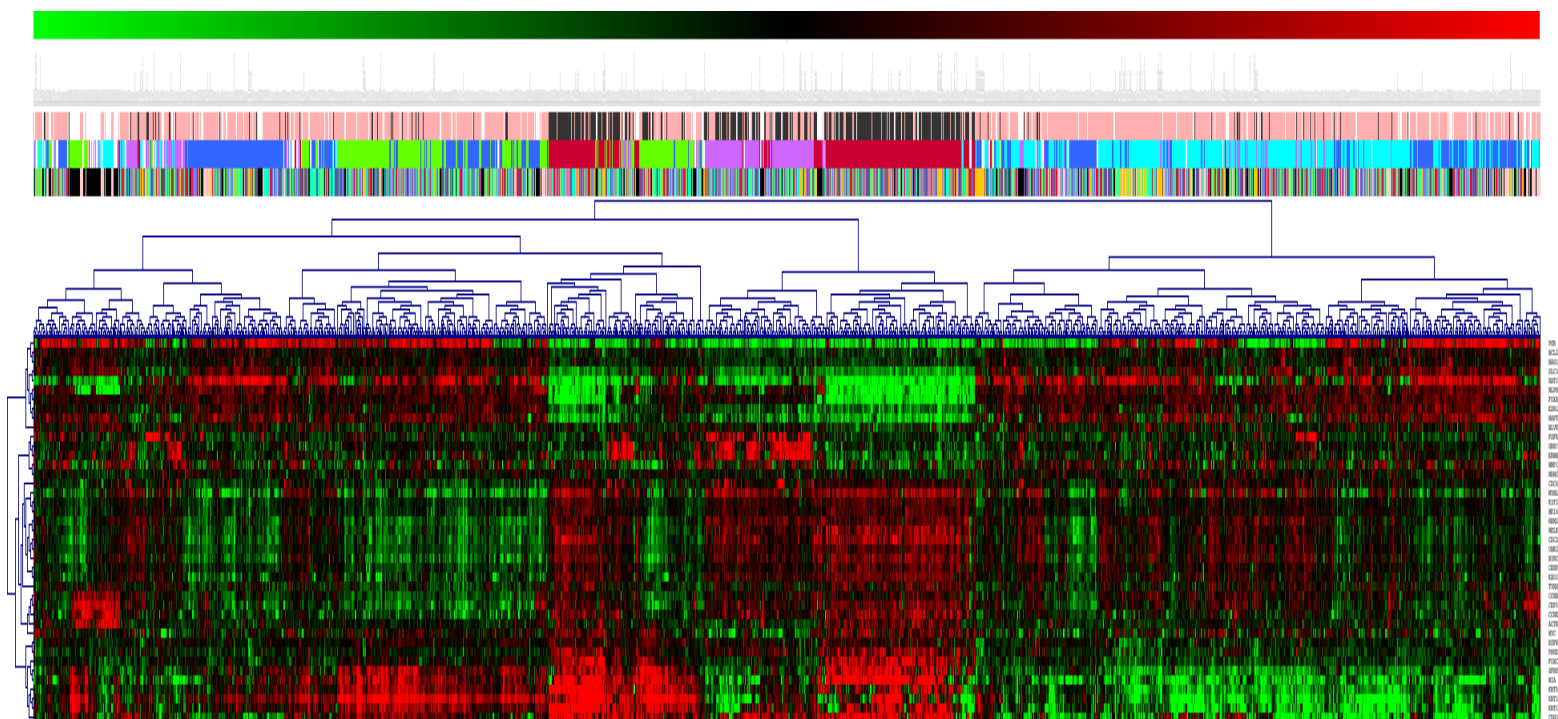


Figure S5. Clustering of PAM50 genes in the 1881-sample data set. The lower annotation bar indicates data set affiliation for each sample (11 different data sets present). The center annotation bar shows PAM50 centroid classification, basal-like (red), HER2 enriched (purple), luminal A (blue), luminal B (light blue) and normal-like (green). Top annotation bar shows ER-status for samples with available annotations; ER-negative (black) and ER-positive (pink). Clustering performed using Pearson correlation and complete linkage.