Figure S2. A schematic framework for machine learning. We used the downloaded TCGA RNA-Seq data for gene selection. In machine learning, we need to split the total samples (N= 646) into training (3/4) and test (1/4) subsets. Subtype assignment of the training samples was based on the PAM50 method (PARKER et al. 2009). The gene panel used for gene search includes: 1) a set of well annotated 1,400 transcriptional factor genes(VAQUERIZAS et al. 2009); and 2) a set of 138 cancer driver genes (VOGELSTEIN et al. 2013). Once we have identified a small ensemble of genes that can best discriminate the four subtypes of breast tumors, we cross-validated the classifier using both the TCGA test data and the data from independent cohorts.