

S1 Table. Label distribution of molecular subtypes in each dataset

Dataset	Luminal A	Luminal B	HER-2	TNBC	Total
TCGA	616 (57.46)	243 (22.67)	39 (3.64)	174 (16.23)	1,072
KG	149 (24.67)	380 (62.91)	21 (3.48)	54 (8.94)	604
Total	765 (45.64)	623 (37.18)	60 (3.58)	228 (13.60)	1,676

The value in the bracket indicates the proportion of each class. HER-2, human epidermal growth factor receptor 2; KG, breast cancer cases from Korea University Guro Hospital; TCGA, breast cancer cases from The Cancer Genomic Atlas; TNBC, triple-negative breast cancer.