

Supplementary Table 1. A simple hierarchical classification system for assignment of a sample to a BCA subtype. The genes listed at panels 1-5 classify a sample as follows: panel 1: (Luminal vs Basal vs HER2+), panel 2: (Luminal A vs Luminal B), panel 3: (LB1 vs LB2 vs LB3), panel 4: (BA1 vs BA2), panel 5: (HER2+I vs HER2+NI). Given an unknown BCA sample array, assuming that there is no variance bias (see **Appendix 4** for details), normalize it so that the median expression value across all genes is zero. Next, standardize each gene value using the mean and stdev values in the table by the rule $x \rightarrow (x - \text{mean}) / \text{stdev}$. Going down the panels as appropriate from 1 \rightarrow 5, compute the Spearman rank correlation coefficient between the sample values and the gene values (given in rows for each category) using all the genes in the panel under consideration. Thus, for panel 1, three coefficients are calculated using the genes in rows for Luminal A, Basal and HER2+; ie, over the 30 dimensional space of genes listed. The sample is assigned to the category with the largest Spearman rank correlation. If the classification in panel 1 is Luminal and a further stratification is desired, then panel 2 may be used to determine whether the assignment is LA or LB. If the latter, then panel 3 is used to decide whether it is LB1 or LB2 or LB3. If panel 1 assigns the samples to Basal or HER2+, then panels 4 or 5 respectively are used for further stratification. At any given panel of stratification, all previous panels are computed first. The final assignment is made to the leaf category which matches up to the root of the panel hierarchy. The classification accuracy of this classifier for the core subtype samples as well as the samples set aside (see text) is given as two pairs of numbers in each panel. The first number in each pair is the number of correctly assigned samples in the category and the second is the actual number of samples in the category. We emphasize that this classifier is presented to show that it is easy to get a reasonably accurate classifier using genes that are biologically meaningful. The more accurate way of classification of a set of unknown samples would be to combine them into the dataset after variance bias correction and redo the entire clustering exercise, using all gene expression values. The assignment would then follow from the subtype cluster to which the unknown samples were placed by the full procedure described in this paper.

Panel 1

Phenotype	Luminal (core samples:148/163 potential samples: 5/8)									
Gene Symbol	CA12	ESR1	FOXA1	GATA3	KRT18	SCUBE2	SLC39A6	TFF3	VAV3	XBP1
mean	5167.23	8024.70	1467.64	2195.40	13494.20	2897.92	14047.27	6879.17	3236.46	8846.58
stdev	4056.10	7315.66	882.45	1669.27	7697.40	3702.97	10062.58	10598.34	2232.81	5007.27
Luminal	0.475	0.565	0.479	0.561	0.411	0.400	0.495	0.252	0.398	0.425
Basal	-1.039	-1.015	-1.378	-1.126	-1.140	-0.691	-0.949	-0.621	-1.027	-1.249
HER2+	-0.315	-0.563	0.108	-0.427	0.127	-0.432	-0.460	0.027	-0.098	0.066
Phenotype	Basal (core samples: 48/49 potential samples: 13/14)									
Gene Symbol	CCNE1	CDH3	CRYAB	CX3CL1	DSC2	FOXC1	KRT17	KRT5	KRT6B	TTK
mean	347.26	883.79	1278.69	613.34	230.67	366.89	1852.38	1400.39	1978.06	311.76
stdev	387.89	1317.33	1919.36	480.39	207.44	441.92	4017.41	2866.41	4458.27	238.97
Luminal	-0.347	-0.421	-0.294	-0.358	-0.391	-0.356	-0.258	-0.250	-0.305	-0.416
Basal	0.931	1.215	0.946	0.940	1.279	1.126	0.792	0.646	1.026	1.226
HER2+	-0.069	0.055	-0.276	-0.047	-0.230	-0.385	-0.197	-0.236	-0.255	-0.004
Phenotype	HER2+ (core samples: 35/42 potential samples: 6/10)									
Gene Symbol	CAP1	CLCA2	ERBB2	GCHFR	GRB7	KMO	PPARBP	S100A9	STARD3	TCAP
mean	4665.34	178.43	47.21	440.09	569.37	640.76	217.18	1914.20	446.08	60.04
stdev	1518.52	643.40	59.06	293.36	840.72	776.70	379.37	4249.78	582.84	93.04
Luminal	-0.215	-0.076	-0.156	-0.143	-0.274	-0.171	-0.277	-0.313	-0.340	-0.148
Basal	0.212	-0.029	-0.173	-0.168	-0.311	-0.160	-0.196	0.383	-0.355	-0.277
HER2+	0.708	0.294	0.725	0.753	1.562	0.832	1.481	0.787	1.908	0.976

Panel 2

Phenotype	LA (core samples: 24/28 potential samples: 12/14)									
Gene Symbol	ACAD8	CAPZA2	CYR61	IFNGR2	KIAA0992	PARVA	RASA1	TCF7L2	TPM1	WIG1
mean	459.029	1156.458	2110.634	3369.000	5733.662	391.810	928.122	2255.399	7717.067	250.793
stdev	198.557	427.554	1561.638	985.558	2431.578	128.835	548.856	1192.947	3455.717	100.295
Luminal A	0.414	0.191	0.510	0.507	0.820	0.186	0.372	0.230	0.202	0.724
Luminal B	0.035	0.024	-0.133	-0.139	0.013	0.049	0.073	-0.268	0.106	0.030
Phenotype	LB (core samples: 79/104 potential samples: 14/17)									
Gene Symbol	BUB3	COX17	COX6C	DCI	FAM77C	GDF15	HAGH	MAGED2	NDUFA2	SEMA3F
mean	2717.263	2484.437	24610.518	1220.864	452.981	555.130	803.002	2400.845	3473.736	1993.503
stdev	1072.503	1303.784	12855.515	590.208	422.736	811.959	396.934	1417.721	1276.736	1018.110
Luminal A	-0.322	-0.354	-0.251	-0.093	0.005	-0.175	-0.115	-0.237	0.129	-0.008
Luminal B	0.400	0.425	0.656	0.555	0.520	0.246	0.553	0.573	0.471	0.325

Panel 3

Phenotype	LB ₁ (core samples: 44/44 potential samples: 4/4)									
Gene Symbol	AHCYL1	BECN1	CCNG2	DCTN4	GLUD2	JMJD2B	MCP	MPDU1	RBBP4	ZNF278
mean	791.949	723.138	540.219	298.113	457.487	255.866	860.639	167.880	996.992	178.671
stdev	355.818	378.121	489.643	158.883	224.190	232.429	581.523	208.696	531.038	99.980
Luminal B1	1.401	1.329	1.337	1.319	1.373	1.330	1.520	1.359	1.267	1.328
Luminal B2	-0.329	0.070	-0.399	-0.097	-0.311	0.056	-0.430	-0.382	-0.352	-0.225
Luminal B3	-0.528	-0.532	-0.677	-0.743	-0.728	-0.182	-0.687	-0.608	-0.888	-0.388
Phenotype	LB ₂ (core samples: 16/22 potential samples: 2/4)									
Gene Symbol	BAIAP3	C16orf53	E2F1	HAGH	HSPB1	NIBP	PGS1	SS18L1	TPD52	TUBG1
mean	339.878	882.910	179.500	803.002	8997.352	1188.854	547.619	660.438	5601.876	720.698
stdev	190.043	310.899	139.587	396.934	6908.955	374.975	235.573	297.948	3365.370	376.888
Luminal B1	0.386	0.381	-0.286	0.116	0.302	-0.007	-0.387	-0.285	0.056	0.182
Luminal B2	1.145	1.330	1.051	1.625	1.231	1.257	1.180	1.191	1.073	1.045
Luminal B3	0.268	0.315	-0.423	0.439	0.172	0.023	0.163	0.034	0.288	-0.156
Phenotype	LB ₃ (core samples: 35/38 potential samples: 9/9)									
Gene Symbol	SEP2	FBXL5	GNS	ITGB5	MATR3	MAX	NDUFB8	RAP1B	TCF12	ZNF410
mean	6383.697	606.837	2192.491	3165.829	8235.690	2383.070	3954.063	3778.036	1863.060	2870.853
stdev	1712.830	294.367	972.371	1630.431	2023.743	809.800	1241.342	1146.607	566.223	949.872
Luminal B1	-0.481	-0.219	-0.551	-0.033	-0.325	-0.321	-0.713	-0.223	-0.367	-0.396
Luminal B2	0.013	0.318	0.209	-0.193	0.303	-0.158	0.073	-0.379	-0.068	0.272
Luminal B3	0.993	1.148	1.195	0.889	0.985	0.935	1.092	0.801	1.197	1.145

Panel 4

Phenotype	BA ₁ (core samples: 15/15 potential samples: 9/9)									
Gene Symbol	CCNB1	CDC42EP1	CTSS	CXCL11	FZD9	HSPA14	IFRG28	PLK1	SERBP1	SLC43A3
mean	701.699	285.785	869.380	329.916	40.985	1005.707	507.332	254.573	1867.652	154.032
stdev	385.483	181.836	711.831	423.784	51.726	556.076	516.910	144.337	916.302	124.631
BA1	1.445	2.274	1.479	2.042	1.673	1.474	2.282	1.798	1.722	2.567
BA2	0.181	0.246	-0.275	0.022	0.945	0.492	0.010	0.715	1.065	0.174
Phenotype	BA ₂ (core samples 20/22 potential samples 2/3)									
Gene Symbol	BAX	CALD1	COL5A3	HOXA11	NGFB	PDGFA	RAB6B	RASL12	SGK2	TGFB11
mean	65.512	201.567	607.137	176.337	71.495	400.138	353.405	354.337	273.093	817.494
stdev	50.461	99.990	287.502	135.677	46.291	265.249	128.142	135.843	94.519	401.746
BA1	-0.614	-0.232	-0.649	-0.639	-0.622	-0.408	-0.543	-0.902	-0.780	-0.787
BA2	0.905	1.614	1.045	0.973	0.977	1.391	1.225	1.095	0.833	0.935

Panel 5

Phenotype	HER2 ₊ (core samples: 14/14 potential samples: 4/7)									
Gene Symbol	DTNB	HLA-DQB1	IGHM	IGKC	IGLC2	IL18RAP	LCK	TNFRSF17	TRBV19 /// TRBC1	XCL1 /// XCL2
mean	81.199	2213.799	3908.121	5518.757	52.564	128.840	172.754	109.526	1187.965	137.258
stdev	111.258	1765.683	6786.195	7906.736	62.179	84.251	199.221	174.876	1469.899	201.309
HER2+_I	0.586	2.066	1.072	1.356	0.909	1.791	1.699	0.859	1.745	2.028
HER2+_NI	-0.503	-0.291	-0.420	-0.455	-0.371	-0.665	-0.458	-0.510	-0.372	-0.406
Phenotype	HER2 _{NI} (core samples: 15/17 potential samples: 3/4)									
Gene Symbol	ACTN1	ATP6V1D	ITGB5	KCNS3	KRT10	P4HA2	RAB1B	SLC24A3	SMARCE1	TPM1
mean	1981.816	1203.167	912.726	745.155	1947.358	1293.722	1602.200	543.061	2126.037	1620.066
stdev	1220.126	438.508	363.583	500.602	1094.575	551.905	855.463	459.661	1605.859	1275.302
HER2+_I	-0.285	-0.096	0.022	-0.410	-0.452	-0.668	-0.086	-0.215	-0.044	-0.413
HER2+_NI	0.365	0.289	0.321	0.232	1.088	1.100	1.227	0.301	1.433	0.165