A Molecular Signature of the Nottingham Prognostic Index in Breast Cancer

(Supplementary Information)

Yu Kun[1], Lee Chee How[1], Tan Puay Hoon[2], Hong Ga Sze[1], Wee Siew Bok[1], Wong Chow Yin[1], and Patrick Tan[1,3,]*

[1] National Cancer Centre / [3] Defence Medical Research Institute

11 Hospital Drive

Singapore 169610

Republic of Singapore

[2] Department of Pathology

11 Hospital Drive

Singapore 169610

Republic of Singapore

* Address correspondence to cmrtan@nccs.com.sg

Tel : 65-6-436-8345

Fax : 65-6-226-5694

Supplementary Table S1. Histopathology of Breast Tumors*

	Age	Size (mm)	Grade	Node	NPI	ER	PR	Subtype	LVI	DCIS
ER+										
2000220	52	60	3	30 of 34	7.2	pos	neg	ductal	yes	minimal
980278	64	40	3	14 of 20	6.8	pos	neg	ductal/ micropap	yes	minimal
2000597	57	40	2	0 of 12	3.8	pos	neg	ductal	possible	extensive
2000609	62	70	2	17 of 17	6.4	pos	pos	ductal	yes	none
20020071	58	28	3	0 of 16	4.56	pos	pos	ductal	no	none
20020160	86	120	3	0 of 10	6.4	pos	pos	lobular	no	none
2000787	57	60	3	0 of 9	5.2	pos	pos	ductal	yes	none
2000818	52	10	2	0 of 11	3.2	pos	neg	ductal	no	minimal
20020051	38	50	3	1 of 25	6	pos	pos	ductal	no	none
20020056	71	20	1	2 of 17	3.4	pos	neg	ductal	no	minimal
980197	55	30	3	2 of 4	5.6	pos	pos	ductal	yes	minimal
980261	60	15	2	0 of 9	3.3	pos	neg	ductal	no	minimal
980391	56	20	2	0 of 7	3.4	pos	pos	ductal	no	none
2000768	39	40	3	0 of 17	4.8	pos	pos	ductal	no	minimal
2000779	48	55	3	0 of 14	5.1	pos	neg	ductal	no	none
990123	54	55	3	7 of 11	7.1	pos	pos	ductal	no	none
2000422	51	63	3	3 of 7	6.26	pos	pos	ductal	no	minimal
2000683	72	35	2	0 of 17	3.7	pos	pos	ductal	no	minimal
2000775	51	25	2	0 of 12	3.5	pos	neg	ductal	no	minimal
2000804	39	40	3	5 of 21	6.8	pos	pos	ductal	yes	minimal
980346	52	20	3	0 of 4	4.4	pos	pos	ductal	possible	minimal
980383	64	30	2	0 of 16	3.6	pos	pos	ductal	no	minimal
990082	49	34	2	3 of 16	4.68	pos	pos	ductal	no	minimal
980177	75	26	2	6 of 13	5.52	pos	pos	ductal	yes	none
980178	69	32	3	2 of 15	5.74	pos	neg	ductal	no	minimal
980403	73	30	3	0 of 9	4.6	pos	pos	ductal	possible	minimal
980434	73	30	3	0 of 16	4.6	pos	pos	ductal	no	minimal
990075	66	25	3	5 of 21	6.5	pos	pos	ductal	yes	none
990113	70	90	3	11 of 15	7.8	pos	pos	ductal	no	minimal
990107	50	40	1	1 of 18	3.8	pos	neg	tub-mixed	yes	minimal
980208	42	25	3	5 of 20	6.5	pos	pos	ductal	no	none
980220	40	37	2	0 of 5	3.74	pos	pos	ductal	yes	minimal
980221	33	65	3	1 of 13	6.3	pos	pos	ductal	no	none
990375	38	15	1	0 of 10	2.3	pos	neg	ductal	no	extensive
ER-										
980193	49	25	3	3 of 23	5.5	neg	neg	ductal	no	minimal
980216	65	45	2	5 of 20	5.9	neg	neg	ductal	no	none
980256	46	36	3	1 of 12	5.72	neg	neg	ductal	no	none
980285	49	40	3	1 of 7	5.8	neg	neg	ductal	yes	minimal
980338	55	30	3	0 of 7	4.6	neg	neg	ductal	no	none

980353	58	45	3	0 of 25	4.9	neg	neg	metaplastic	no	none
980411	69	30	2	0 of 9	3.6	neg	neg	ductal	no	none
980441	66	30	3	4 of 14	6.6	neg	neg	ductal	yes	none
990174	55	45	2	3 of 24	5.9	neg	neg	ductal	yes	minimal
2000320	67	20	3	20 of 21	6.4	neg	neg	ductal	yes	none
2000500	44	75	3	6 of 6	7.5	neg	neg	ductal	yes	none
980247	35	45	3	1 of 19	5.9	neg	neg	ductal	yes	minimal
990299	58	55	3	7 of 17	7.1	neg	neg	ductal	possible	minimal
2000593	60	41	3	0 of 15	4.82	neg	neg	ductal	no	none
2000638	60	40	1	0 of 15	2.8	pos	neg	lobular	no	none
2000731	68	51	3	1 of 29	6.02	pos	neg	ductal	no	minimal
2000880	55	15	2	0 of 26	3.3	neg	neg	ductal	no	none
ERBB2										
980194	58	50	3	25 of 32	7	neg	neg	ductal	yes	none
980214	49	60	2	5 of 13	6.2	pos	neg	ductal	no	extensive
980238	62	20	3	7 of 21	6.4	neg	neg	ductal	no	extensive
980288	45	60	3	13 of 15	7.2	pos	neg	ductal	yes	extensive
980335	33	3	3	3 of 7	5.06	neg	neg	ductal	yes	extensive
980373	77	30	3	0 of 14	4.6	neg	neg	ductal	no	minimal
980380	56			0 of 6		neg	neg			
980395	68	30	3	1 of 10	5.6	neg	neg	ductal	yes	none
980396	66	35	3	10 of 12	6.7	neg	neg	ductal	yes	extensive
990115	38	28	3	9 of 10	6.56	pos	pos	ductal	yes	extensive
990134	43	40	3	0 of 19	4.8	neg	neg	ductal	no	none
990148	60	40	2	6 of 19	5.8	pos	neg	ductal	yes	minimal
990223	52	5	3	1 of 21	5.1	pos	neg	ductal	no	extensive
2000104	59					pos	neg	ductal		
2000171	50	25	2	0 of 9	3.5	neg	neg	ductal	no	none
2000209	58	50	3	0 of 7	5	pos	neg	ductal	no	none
2000210	50	40	3	3 of 6	5.8	neg	neg	ductal	yes	none
2000237	43	47	3	23 of 40	6.94	pos	pos	ductal	yes	minimal
2000287	53	40	3	0 of 8	4.8	neg	neg	ductal	possible	none
2000399	44	40	2	0 of 8	3.8	neg	neg	ductal	no	minimal
2000641	47	60	3	16 of 24	5.2	neg	neg	ductal	yes	minimal
2000652	56	25	3	6 of 21	6.5	neg	neg	ductal	no	minimal
2000675	78	55	3	16 of 16	7.1	neg	neg	ductal	yes	minimal
2000709	45	30	3	0 of 16	4.6	neg	neg	ductal	no	none
2000759	57	7	3	0 of 12	4.14	neg	neg	ductal	no	extensive
2000813	60	23	3	16 of 17	6.46	neg	neg	ductal	yes	extensive
2000829	51	45	2	10 of 10	5.9	neg	neg	ductal	yes	extensive
20020090	60	45	3	19 of 27	6.9	neg	neg	ductal	yes	minimal

* This list contains clinical information for 79 out of 98 tumors used in this study. Clinical information for the remaining 19 tumors was incomplete and not included in this list. Only the 79 samples with complete clinical information was used for subsequent NPI-ES analysis.

Supplementary Information S2 : Descriptions of Weighted Voting (WV) and Leave-One-Out Cross Validation (LOOCV) Assays

***Weighted Voting (WV)* :** The weighted voting algorithm utilizes a signal-to-noise (S2N) metric to perform binary classifications. Each gene belonging to a predictor set is assigned a 'vote', expressed as the weighted difference between the gene expression level in the sample to be classified and the average class mean expression level. Weighting is determined using the correlation metric $P(g,c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$ (μ and σ denotes means and standard deviations of expression levels of the gene in each of the two classes). The ultimate vote for a particular class assignment is computed by summing all weighted votes made by each gene used in the class discrimination. The "prediction strength" (PS) is defined as $PS = \frac{V_{win} - V_{lose}}{V_{win} + V_{lose}}$, where V_{win} and V_{lose} are the vote totals for the winning and losing classes, respectively. PS reflects the relative margin of victory and hence provides a quantitative reflection of prediction certainty.

Leave-One-Out Cross Validation (LOOCV) : We used a standard leave-one-out cross-validation (LOOCV) approach to assess classification accuracy in the training set. In LOOCV, one sample in the training set is initially 'left out', and the classifier operations (eg gene selection and classifier training) are performed on the remaining samples. The 'left out' sample is then classified using the trained algorithm, and this process is then repeated for all samples in the training set.

Supplementary Data S3 : Derivation of a NPI Expression Signature Using All Tumors, Regardless of Subtype

In the main text, we defined the NPI-ES using a two-step methodology. First, unsupervised clustering was used to cluster tumors according to their respective 'molecular subtype' (ie ER+, ER-, ERBB2+, see main text). Second, tumors within each subtype were analyzed for expression signatures that might be correlated to the NPI. Here, we show that performing the first step (definition of distinct molecular subtypes) is essential in the identification of the NPI-ES. We assembled a data set consisting of ALL 79 tumors, regardless of molecular subtype, and performed a moving NPI threshold analysis to define an 'appropriate' NPI threshold, as in the main text (Figure 2a in main text). We found that using an NPI threshold of 4 yielded a total of 44 differentially expressed genes. Of this 44 gene set, 16 (35%) also belong to the NPI-ES (which was derived from ER+ samples).

We then used Weighted Voting (WV) and cross-validation (LOOCV) assays to assess the ability of this 44 gene set to confidently classify the tumor samples into discrete groups. As can be seen in Figure S3, the number of low-confidence (PS<0.3, red area) samples, as well as the misclassification rate (9% for the 44 gene set) are both significantly increased compared to Figure 2c in the main text. This result indicates that the 44-gene set, based upon all 79 tumors, is less effective in predicting the NPI status of a tumor than the NPI-ES (on ER+ tumors).

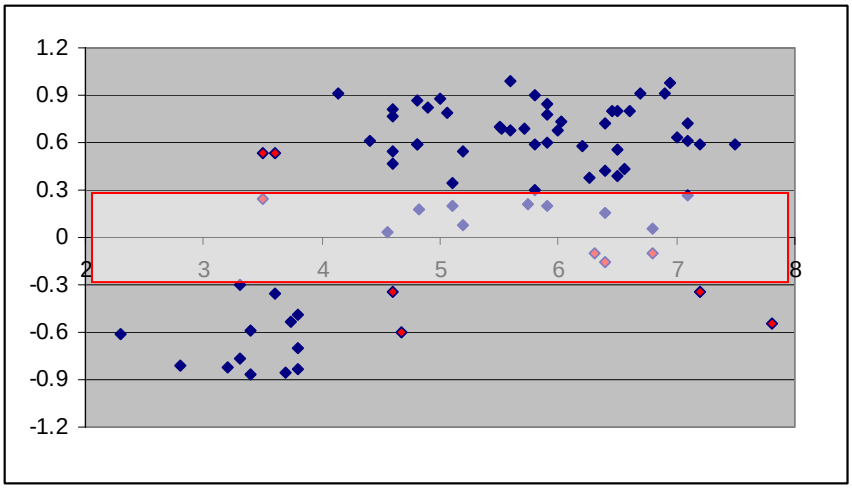

Figure S3 : Classification and prediction confidence of tumor samples using the 44-gene set based on ALL tumors regardless of subtype. Samples are sorted by their NPI value (X-axis). Weighted voting was used to classify the samples and the prediction strengths of each sample (Y-axis) calculated based upon Golub et al., (ref. 13). Sample classifications with a prediction strength of <0.3 are considered 'uncertain' or 'low-confidence' (gray area). A higher number of 'uncertain' (low PS) samples and misclassified samples are observed compared to Figure 2c (main text).

The 44 gene set derived from ALL tumors regardless of subtype is also not as effective as the NPI-ES at predicting NPI status in an independent data set. Using the Rosetta data set as a blinded test set, we applied the 44 gene set to the 49 ER+ tumors found in the Rosetta data set, and used Student's t-test to determine the significance of association between a ER+ tumors expressing high levels of the 44 gene set and possessing a high NPI. We obtained a p-value of 0.29 for the 44 gene set, which was much less significant compared to a p-value of 0.0004 for the NPI-ES (main text).

Interestingly, the NPI-ES, despite being derived from an analysis of ER+ tumors, outperforms the 44 gene set even when applied across all 78 tumors in the Rosetta data set. To illustrate this, the 78 Rosetta tumors were divided into two groups of NPI<3.4 (good prognosis) and >3.4 respectively (moderate prognosis). Weighted voting was then used to classify the Rosetta tumors by the NPI-ES or the 44 gene set. As can be seen in

Table S3, the NPI-ES delivered a classification accuracy of 80%, compared to the 44 gene set which delivered a 70% classification accuracy.

Table S3 : Classification accuracy of the NPI-ES or 44 gene set on 78 Rosetta Tumors

	NPI classification (<3.4 or >3.4)
	No. of misclassifications (Accuracy)
44 Genes	23 (70%)
NPI-ES	15 (80%)

Supplementary Table S5 : List of top 50 Significantly Regulated Genes in ER+, ER- and ERBB2+ Molecular Subtypes

This list represents the top 50 genes identified by SAM to be significantly regulated in each molecular subtype (ER+, ER-, ERBB2+). The genes are ranked by their S2N correlation ratio, which reflects the extent of the expression perturbation observed among different groups. There is good overlap between these genes and similar lists reported by other studies (ref. 8-11) (main text).

Gene description	Unigene	Chromosome
ER+ Molecular Subtype		
estrogen receptor 1	Hs.1657	Chr:6q25.1
GATA binding protein 3	Hs.169946	Chr:10p15
annexin A9	Hs.279928	Chr:1q21
KIAA0882 protein	Hs.90419	Chr:4q31.1
carbonic anhydrase XII	Hs.5338	Chr:15q22
cytochrome P450, subfamily IIB (phenobarbital-inducible), polypeptide 6	Hs.1360	Chr:19q13.2
dynein, axonemal, light intermediate polypeptide 1	Hs.406050	Chr:1p35.1
sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B	Hs.82222	Chr:3p21.3
N-acetyltransferase 1 (arylamine N-acetyltransferase)	Hs.155956	Chr:8p23.1-p21.3
serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5	Hs.76353	Chr:14q32.1
cytochrome c oxidase subunit VIc	Hs.351875	Chr:8q22-q23
Homo sapiens mRNA; cDNA DKFZp564F053 (from clone DKFZp564F053), mRNA sequence	Hs.71968	---
LIV-1 protein, estrogen regulated	Hs.79136	Chr:18q12.1
troponin T1, skeletal, slow	Hs.73980	Chr:19q13.4
hypothetical protein FLJ20151	Hs.279916	Chr:15q21.3
calsyntenin 2	Hs.12079	Chr:3q23-q24
B-cell CLL/lymphoma 2	Hs.79241	Chr:18q21.3
guanidinoacetate N-methyltransferase	Hs.81131	Chr:19p13.3
microtubule-associated protein tau	Hs.101174	Chr:17q21.1
hypothetical protein FLJ12910	Hs.15929	Chr:6q25.1
WW domain-containing protein 1	Hs.355977	Chr:8q21
UDP-glucose ceramide glucosyltransferase	Hs.432605	Chr:9q31
GREB1 protein	Hs.193914	Chr:2p25.1
RNB6	Hs.241471	Chr:14q32.32
Human insulin-like growth factor 1 receptor mRNA, 3' sequence, mRNA sequence	Hs.405998	---
interleukin 6 signal transducer (gp130, oncostatin M receptor)	Hs.82065	Chr:5q11
LAG1 longevity assurance homolog 2 (S. cerevisiae)	Hs.285976	Chr:1q21.2
cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila)	Hs.57652	Chr:1p21
paired basic amino acid cleaving system 4	Hs.170414	Chr:15q26
regulator of G-protein signalling 11	Hs.65756	Chr:16p13.3

UDP-glucose ceramide glucosyltransferase	Hs.432605	Chr:9q31
NPD009 protein	Hs.283675	Chr:16p13.2
v-myb myeloblastosis viral oncogene homolog (avian)	Hs.1334	Chr:6q22-q23
interleukin 6 signal transducer (gp130, oncostatin M receptor)	Hs.82065	Chr:5q11
discs, large (Drosophila) homolog 5	Hs.170290	Chr:10q23
Homo sapiens mRNA; cDNA DKFZp434E082 (from clone DKFZp434E082), mRNA sequence	Hs.432587	---
cytochrome P450, subfamily IIB (phenobarbital-inducible), polypeptide 7	Hs.330780	Chr:19q13.2
HSPC009 protein	Hs.16059	Chr:17q21
KIAA1025 protein	Hs.4084	Chr:12q24.22
protein tyrosine phosphatase type IVA, member 2	Hs.82911	Chr:1p35
CGI-49 protein	Hs.238126	Chr:1q44
chromosome 20 open reading frame 35	Hs.256086	Chr:20q13.11
phorbol-12-myristate-13-acetate-induced protein 1	Hs.96	Chr:18q21.31
KIAA0876 protein	Hs.301011	Chr:19p13.3
hypothetical protein FLJ20152	Hs.82273	Chr:5p15.1
hypothetical protein FLJ22318	Hs.22753	Chr:5q35.3
trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)	Hs.350470	Chr:21q22.3
polymerase (DNA-directed), delta 4	Hs.82520	Chr:11q13
putative proline 4-hydroxylase	Hs.348198	Chr:3p21.31
GDNF family receptor alpha 1	Hs.105445	Chr:10q26

ERBB2+ Molecular Subtype

chloride channel, calcium activated, family member 2	Hs.241551	Chr:1p31-p22
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	Hs.323910	Chr:17q11.2-q12
growth factor receptor-bound protein 7	Hs.86859	Chr:17q21.1
dual specificity phosphatase 6	Hs.180383	Chr:12q22-q23
START domain containing 3	Hs.77628	Chr:17q11-q12
transient receptor potential cation channel, subfamily V, member 6	Hs.302740	Chr:7q33-q34
S100 calcium binding protein A8 (calgranulin A)	Hs.100000	Chr:1q21
protein phosphatase 1, regulatory (inhibitor) subunit 1A	Hs.76780	Chr:12q13.13
fibroblast growth factor receptor 4	Hs.165950	Chr:5q35.1-qter
SRY (sex determining region Y)-box 11	Hs.32964	Chr:2p25
Unknown protein [Homo sapiens], mRNA sequence	Hs.106642	---
transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila)	Hs.28935	Chr:9q21.32
hypothetical gene MGC9753	Hs.91668	Chr:17q21.1
mitogen-activated protein kinase kinase kinase 5	Hs.151988	Chr:6q22.33
KIAA1102 protein	Hs.202949	Chr:4p13
fatty acid hydroxylase	Hs.249163	Chr:16q23
transcription factor AP-2 beta (activating enhancer binding protein 2 beta)	Hs.33102	Chr:6p12
S100 calcium binding protein A9 (calgranulin B)	Hs.112405	Chr:1q21
fatty-acid-Coenzyme A ligase, long-chain 2	Hs.154890	Chr:4q34-q35
hypothetical protein FLJ22671	Hs.193745	Chr:2q37.3
kynurenine 3-monooxygenase (kynurenine 3-hydroxylase)	Hs.107318	Chr:1q42-q44

KIAA0644 gene product	Hs.21572	Chr:7p15.1
aspartate beta-hydroxylase	Hs.283664	Chr:8q12.1
electron-transfer-flavoprotein, alpha polypeptide (glutaric aciduria II)	Hs.169919	Chr:15q23-q25
secretory leukocyte protease inhibitor (antileukoproteinase)	Hs.251754	Chr:20q12
isocitrate dehydrogenase 1 (NADP+), soluble	Hs.11223	Chr:2q33.3
phenylethanolamine N-methyltransferase	Hs.1892	Chr:17q21-q22
hypothetical protein FLJ14146	Hs.103395	Chr:1q42.11
fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase, Lewis blood group included)	Hs.169238	Chr:19p13.3
keratin, hair, basic, 1	Hs.32952	Chr:12q13
PDZ domain containing 2	Hs.173035	Chr:5p13.3
argininosuccinate synthetase	Hs.160786	Chr:9q34.1
specific granule protein (28 kDa)	Hs.54431	Chr:6p12.3
Homo sapiens cDNA: FLJ21521 fis, clone COL05880, mRNA sequence	Hs.306777	---
kynureninase (L-kynurenine hydrolase)	Hs.169139	Chr:2q22.1
hypothetical protein FLJ20539	Hs.118552	Chr:11q12.1
proline dehydrogenase (oxidase) 1	Hs.343874	Chr:22q11.21
v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)	Hs.25960	Chr:2p24.1
integrin, beta 6	Hs.57664	Chr:2q24.2
hypothetical protein MGC3077	Hs.433404	Chr:7p15-p14
uncoupling protein 2 (mitochondrial, proton carrier)	Hs.80658	Chr:11q13
myosin X	Hs.61638	Chr:5p15.1-p14.3
keratin 7	Hs.23881	Chr:12q12-q21
steroid sulfatase (microsomal), arylsulfatase C, isozyme S	Hs.79876	Chr:Xp22.32
formin homology 2 domain containing 1	Hs.95231	Chr:16q22
ATP-binding cassette, sub-family C (CFTR/MRP), member 3	Hs.90786	Chr:17q22
chondroitin beta1,4 N-acetylgalactosaminyltransferase	Hs.11260	Chr:8p21.3
KIAA0485 protein	Hs.89121	---
kraken-like	Hs.301947	Chr:22q13
collagen, type XIII, alpha 1	Hs.211933	Chr:10q22

ER- Molecular Subtype

keratin 16 (focal non-epidermolytic palmoplantar keratoderma)	Hs.432448	Chr:17q12-q21
gamma-aminobutyric acid (GABA) A receptor, pi	Hs.70725	Chr:5q33-q34
TONDU	Hs.9030	Chr:Xq26.3
keratin 6B	Hs.432677	Chr:12q12-q13
serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 5	Hs.55279	Chr:18q21.3
keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types)	Hs.433845	Chr:12q12-q13
SRY (sex determining region Y)-box 10	Hs.44317	Chr:22q13.1
melanoma inhibitory activity	Hs.279651	Chr:19q13.32-q13.33
matrix metalloproteinase 7 (matrilysin, uterine)	Hs.2256	Chr:11q21-q22
secreted frizzled-related protein 1	Hs.7306	Chr:8p12-p11.1
B-cell CLL/lymphoma 11A (zinc finger protein)	Hs.130881	Chr:2p15

Homo sapiens cDNA FLJ11796 fis, clone HEMBA1006158, highly similar to Homo sapiens transcription factor forkhead-like 7 (FKHL7) gene, mRNA sequence	Hs.284186	---
solute carrier family 6 (neurotransmitter transporter), member 14	Hs.162211	Chr:Xq23-q24
desmuslin	Hs.10587	Chr:15q26.3
engrailed homolog 1	Hs.271977	Chr:2q13-q21
ribosomal protein, large P2	Hs.153179	Chr:11p15.5-p15.4
tripartite motif-containing 29	Hs.82237	Chr:11q22-q23
calmodulin-like skin protein	Hs.180142	Chr:10p15.1
desmocollin 2	Hs.239727	Chr:18q12.1
ropporin, rhophilin associated protein	Hs.194093	Chr:3q21.1
crystallin, alpha B	Hs.391270	Chr:11q22.3-q23.1
tripartite motif-containing 2	Hs.12372	Chr:4q31.23
epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	Hs.77432	Chr:7p12
leucine-rich acidic nuclear protein like	Hs.71331	Chr:1q21.2
potassium channel, subfamily K, member 5	Hs.127007	Chr:6p21
kallikrein 5	Hs.50915	Chr:19q13.3-q13.4
procollagen C-endopeptidase enhancer 2	Hs.8944	Chr:3q21-q24
Hypothetical protein [Homo sapiens], mRNA sequence	Hs.66762	---
LIM domain only 4	Hs.3844	Chr:1p22.3
keratin 17	Hs.2785	Chr:17q12-q21
desmoglein 3 (pemphigus vulgaris antigen)	Hs.1925	Chr:18q12.1-q12.2
keratin 6A	Hs.367762	Chr:12q12-q13
sialyltransferase 8A (alpha-N-acetylneuraminate: alpha-2,8-sialytransferase, GD3 synthase)	Hs.82527	Chr:12p12.1-p11.2
Kruppel-like factor 5 (intestinal)	Hs.84728	Chr:13q21.32
Rho guanine nucleotide exchange factor (GEF) 4	Hs.6066	Chr:2q22
kallikrein 6 (neurosin, zyme)	Hs.79361	Chr:19q13.3
prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	Hs.196384	Chr:1q25.2-q25.3
chromosome 20 open reading frame 42	Hs.180479	Chr:20p12.3
glycoprotein M6B	Hs.5422	Chr:Xp22.2
uridine phosphorylase	Hs.77573	Chr:7
ladinin 1	Hs.18141	Chr:1q25.1-q32.3
pleiomorphic adenoma gene-like 1	Hs.75825	Chr:6q24-q25
desmocollin 3	Hs.41690	Chr:18q12.1
Homo sapiens cDNA FLJ30869 fis, clone FEBRA2004224, mRNA sequence	Hs.349611	---
HRAS-like suppressor	Hs.36761	Chr:3q29
cysteine and glycine-rich protein 2	Hs.10526	Chr:12q21.1
scrapie responsive protein 1	Hs.7122	Chr:4q31-q32
amyloid beta (A4) precursor protein-binding, family A, member 2 (X11-like)	Hs.26468	Chr:15q11-q12
jerky homolog-like (mouse)	Hs.105940	Chr:11q21
transforming growth factor, alpha	Hs.170009	Chr:2p13

Supplementary Table S6 : Genes Belonging to the NPI-ES (62 Genes)

DC13 protein is the only gene of NPI-ES that can be matched in Rosetta 70-gene 'prognosis' signature (PES, see main text), out of which 42 are present in the Affymetrix U133A chip.

Gene Description	Unigene	Biological Process (GO)
Positive genes (60) (Highly Expressed in High NPI Tumors)		
adenine phosphoribosyltransferase	Hs.28914	9116 // nucleoside metabolism // extended:inferred from electronic annotation; Pribosyltran; 5e-44
MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)	Hs.154443	6260 // DNA replication // predicted/computed
exonuclease 1	Hs.47504	6310 // DNA recombination // experimental evidence /// 6281 // DNA repair // experimental evidence /// 6298 // mismatch repair // predicted/computed
Metallothionein 1H-like protein [Homo sapiens], mRNA sequence	Hs.367850	---
Homo sapiens, clone IMAGE:5270727, mRNA, mRNA sequence	Hs.319215	---
DC13 protein	Hs.6879	---
HSPC037 protein	Hs.433180	---
H2A histone family, member Z	Hs.119192	---
discs, large homolog 7 (Drosophila)	Hs.77695	7267 // cell-cell signaling // extended:Unknown; GKAP; 2.1e-05
RNA helicase-related protein [Homo sapiens], mRNA sequence	Hs.381097	---
kinesin-like 1	Hs.8878	7067 // mitosis // experimental evidence /// 7052 // mitotic spindle assembly // experimental evidence
chromosome 20 open reading frame 1	Hs.9329	7067 // mitosis // predicted/computed /// 8283 // cell proliferation // predicted/computed
KIAA0095 gene product	Hs.155314	---
helicase, lymphoid-specific	Hs.203963	---
homeo box HB9	Hs.37035	6959 // humoral immune response // experimental evidence /// 6357 // regulation of transcription from Pol II promoter // predicted/computed /// 7345 // embryogenesis and morphogenesis // experimental evidence
DNA segment on chromosome X (unique) 9879 expressed sequence	Hs.18212	---
MAD2 mitotic arrest deficient-like 1 (yeast)	Hs.79078	7067 // mitosis // predicted/computed /// 7093 // mitotic checkpoint // experimental evidence
eukaryotic translation initiation factor 4E binding protein 1	Hs.433317	6445 // regulation of translation // predicted/computed
cathepsin C	Hs.10029	6508 // proteolysis and peptidolysis // not recorded /// 6955 // immune response // experimental evidence
H2B histone family, member J	Hs.249216	---
proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional protease 7)	Hs.180062	6508 // proteolysis and peptidolysis // not recorded
hypothetical protein FLJ20105	Hs.89306	---
chromosome 10 open reading frame 3	Hs.14559	---
uncharacterized bone marrow protein BM039	Hs.283532	---
likely ortholog of mouse gene rich cluster, C8 gene	Hs.30114	---
cell division cycle 2, G1 to S and G2 to M	Hs.334562	74 // regulation of cell cycle // not recorded /// 7089 // start control point of mitotic cell cycle // not recorded
metallothionein 2A	Hs.118786	6878 // copper homeostasis // predicted/computed

geminin, DNA replication inhibitor	Hs.234896	7050 // cell cycle arrest // predicted/computed /// 8156 // negative regulation of DNA replication // predicted/computed
low density lipoprotein receptor-related protein 8, apolipoprotein e receptor	Hs.54481	7165 // signal transduction // predicted/computed /// 6629 // lipid metabolism // predicted/computed
hematological and neurological expressed 1	Hs.109706	---
H1 histone family, member 2	Hs.7644	---
nudix (nucleoside diphosphate linked moiety X)-type motif 1	Hs.388	6979 // response to oxidative stress // predicted/computed /// 6281 // DNA repair // not recorded
metallothionein 1X	Hs.374950	---
H2B histone family, member T	Hs.247817	---
tetraspan 1	Hs.38972	8283 // cell proliferation // not recorded /// 8583 // mystery cell fate differentiation (sensu Drosophila) // predicted/computed /// 7155 // cell adhesion // not recorded /// 6928 // cell motility // not recorded
metallothionein 1H	Hs.2667	---
H3 histone family, member K	Hs.70937	---
ribonucleotide reductase M2 polypeptide	Hs.75319	---
baculoviral IAP repeat-containing 5 (survivin)	Hs.1578	86 // G2/M transition of mitotic cell cycle // experimental evidence /// 7048 // oncogenesis // predicted/computed /// 6916 // anti-apoptosis // experimental evidence
F-box only protein 5	Hs.272027	6508 // proteolysis and peptidolysis // predicted/computed
serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	Hs.297681	---
lysosomal associated protein transmembrane 4 beta	Hs.296398	---
chemokine (C-X3-C motif) ligand 1	Hs.80420	7165 // signal transduction // experimental evidence /// 6954 // inflammatory response // not recorded /// 6935 // chemotaxis // experimental evidence /// 6955 // immune response // not recorded /// 7155 // cell adhesion // experimental evidence /// 7267 // cell-cell signaling // experimental evidence
CD27-binding (Siva) protein	Hs.112058	8624 // induction of apoptosis by extracellular signals // predicted/computed /// 6952 // defense response // predicted/computed
LGN protein	Hs.278338	7186 // G-protein coupled receptor protein signaling pathway // predicted/computed
Mouse Mammary Turmor Virus Receptor homolog 1	Hs.18686	---
forkhead box M1	Hs.239	6366 // transcription from Pol II promoter // experimental evidence /// 6979 // response to oxidative stress // experimental evidence
met proto-oncogene (hepatocyte growth factor receptor)	Hs.316752	7048 // oncogenesis // experimental evidence /// 8283 // cell proliferation // predicted/computed /// 7165 // signal transduction // predicted/computed
butyrophilin, subfamily 3, member A2	Hs.87497	---
SBBI26 protein	Hs.26481	---
likely ortholog of mouse Shc SH2-domain binding protein 1	Hs.123253	---
H3 histone family, member B	Hs.143042	---
trefoil factor 3 (intestinal)	Hs.82961	6952 // defense response // predicted/computed /// 7586 // digestion // predicted/computed
immunoglobulin lambda locus	Hs.405944	---
DNA replication factor	Hs.122908	---
Homo sapiens cDNA FLJ30781 fis, clone FEBRA2000874, mRNA sequence	Hs.301663	---

chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)	Hs.16530	7165 // signal transduction // experimental evidence /// 7154 // cell communication // predicted/computed /// 6935 // chemotaxis // experimental evidence /// 6955 // immune response // predicted/computed /// 6960 // antimicrobial humoral response (sensu Invertebrata) // predicted/computed /// 9607 // response to biotic stimulus // predicted/computed /// 7267 // cell-cell signaling // experimental evidence
immunoglobulin kappa constant	Hs.406565	---
suppressor of Ty 4 homolog 1 (S. cerevisiae)	Hs.79058	6355 // regulation of transcription, DNA-dependent // predicted/computed /// 6357 // regulation of transcription from Pol II promoter // predicted/computed /// 6338 // chromatin modeling // predicted/computed
paternally expressed 10	Hs.137476	---
Negative genes (2) (Highly Expressed in Low NPI Tumors)		
BTG family, member 2	Hs.75462	8285 // negative regulation of cell proliferation // predicted/computed /// 6281 // DNA repair // predicted/computed /// 6976 // DNA damage response, activation of p53 // predicted/computed
cytochrome P450, subfamily IVF, polypeptide 8	Hs.268554	6118 // electron transport // extended:Unknown; p450; 1.9e-142 /// 6693 // prostaglandin metabolism // predicted/computed

Supplementary Table S7. Genes associated with histological grade (1 & 2 vs. 3)

Since the classical NPI is a composite metric derived from tumor grade, tumor size, and lymph node status, we defined the contributions made by each of these individual elements to the molecular makeup of the NPI-ES. Using SAM to identify genes correlated to each of the three histopathological variables, we were unable to convincingly identify genes whose expression was significantly correlated to either tumor size or lymph node status. In contrast, in the case of histological grade, a significant number of genes were found to be differentially expressed between grade 1 or 2 and grade 3 tumors, and the genes in this grade-correlated gene set exhibited substantial overlap (66%) with the NPI-ES (Table S6). These results suggest that tumors exhibiting different histological grades may be biologically distinct, and that tumor grade is a key contributor to the NPI expression signature, with the remaining two parameters (tumor size and lymph node status) delivering comparatively lesser contributions.

Table S7. SAM was performed to identify 68 genes significantly associated with grade (FDR of 14%, >=2-fold change). 45 out of these genes (66%) are also belong to the NPI classifier, labeled as "YES" in the NPI-ES column.

Gene Name	NPI-ES
Genes upregulated in Grade 3 tumors	
RAD51-interacting protein	
DC13 protein	YES
HSPC037 protein	YES
homeo box HB9	YES
cyclin B2	
protein regulator of cytokinesis 1	
likely ortholog of mouse gene rich cluster, C8 gene	YES
kinesin-like 1	YES
H2A histone family, member Z	YES
DNA replication factor	YES
MCM4 minichromosome maintenance deficient 4 (S. cerevisiae)	YES
discs, large homolog 7 (Drosophila)	YES
ZW10 interactor	
MAD2 mitotic arrest deficient-like 1 (yeast)	YES
Metallothionein 1H-like protein [Homo sapiens], mRNA sequence	YES
chromosome 10 open reading frame 3	YES
ribonucleotide reductase M2 polypeptide	YES
cell division cycle 2, G1 to S and G2 to M	YES
forkhead box M1	YES
uncharacterized bone marrow protein BM039	YES
helicase, lymphoid-specific	YES
RNA helicase-related protein [Homo sapiens], mRNA sequence	YES
metallothionein 1X	YES
Homo sapiens, clone IMAGE:5270727, mRNA, mRNA sequence	YES
metallothionein 2A	YES
metallothionein 1H	YES
KIAA0095 gene product	YES
baculoviral IAP repeat-containing 5 (survivin)	YES
geminin, DNA replication inhibitor	YES
enhancer of zeste homolog 2 (Drosophila)	
cathepsin C	YES
nudix (nucleoside diphosphate linked moiety X)-type motif 1	YES
hypothetical protein FLJ10719	
chemokine (C-X3-C motif) ligand 1	YES
tetraspan 1	YES
proapoptotic caspase adaptor protein	
immunoglobulin lambda locus	YES
H2B histone family, member J	YES
trefoil factor 3 (intestinal)	YES
CD27-binding (Siva) protein	YES
topoisomerase (DNA) II alpha 170kDa	

immunoglobulin lambda joining 3	
eukaryotic translation initiation factor 4E binding protein 1	YES
H3 histone family, member K	YES
chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)	YES
lysosomal associated protein transmembrane 4 beta	YES
Mouse Mammary Turmor Virus Receptor homolog 1	YES
LGN protein	YES
immunoglobulin kappa constant	YES
carboxypeptidase B1 (tissue)	
met proto-oncogene (hepatocyte growth factor receptor)	YES
H2B histone family, member T	YES
RAB38, member RAS oncogene family	
H1 histone family, member 2	YES
hypothetical protein from EUROIMAGE 2021883	
apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B	
H3 histone family, member B	YES
immunoglobulin heavy constant gamma 3 (G3m marker)	
similar to bK246H3.1 (immunoglobulin lambda-like polypeptide 1, pre-B-cell specific)	
Immunoglobulin lambda light chain [Homo sapiens], mRNA sequence	
Immunoglobulin kappa light chain variable region [Homo sapiens], mRNA sequence	
serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	YES
proteolipid protein 1 (Pelizaeus-Merzbacher disease, spastic paraplegia 2, uncomplicated)	
sodium channel, nonvoltage-gated 1, beta (Liddle syndrome)	
H4 histone family, member H	
syndecan 2 (heparan sulfate proteoglycan 1, cell surface-associated, fibroglycan)	
neuropilin (NRP) and tolloid (TLL)-like 2	
Genes downregulated in Grade 3 tumors	
hypothetical protein FLJ22418	

Supplementary Data S8 : Identifying ER+ Tumors in the Rosetta Data Set using the SAM-409 Gene Set

Out of the 409 genes in the SAM-409 gene set used to define the ER+, ER-, and ERBB2+ molecular subtypes (main text), there are 276 matched genes found in the Rosetta data set. Use of these overlapping 276 genes to re-cluster our data set and successfully reproduced the three groups, with only one sample being assigned to a different group. This result indicates that this overlapping set of 276 genes contains sufficient information to distinguish between the three molecular subtypes. These 276 genes were then used to cluster the 78 breast primary tumors from the Rosetta data set.

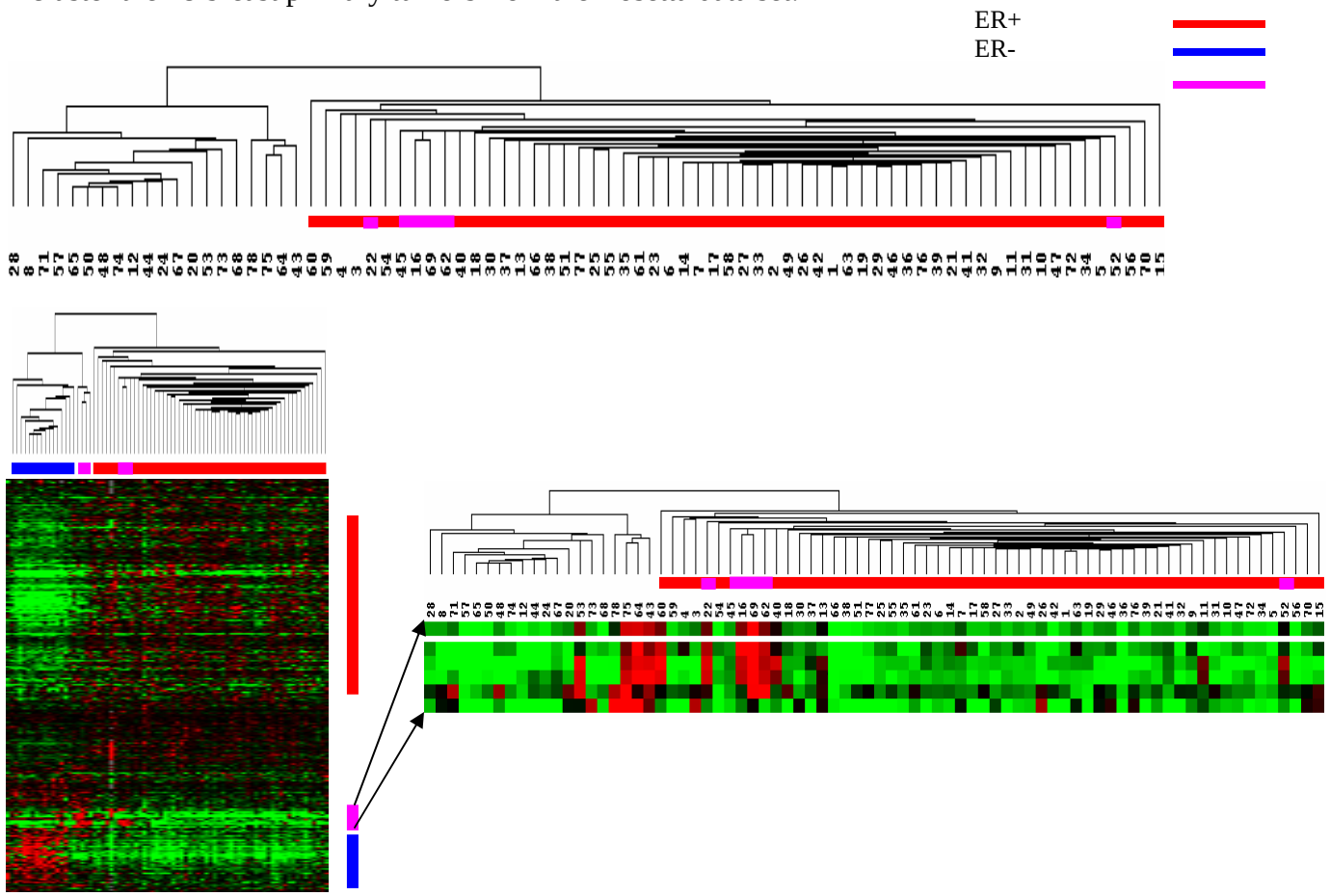

Figure S8. Hierarchical clustering of gene expression data from Rosetta data set. Top) Dendrogram displaying the similarities between tumors. The color-coded bar indicated the subtype to the corresponding gene signature. Left) The full cluster of 276 genes with three distinct gene clusters. Note that some ERBB2 tumors appeared to segregate with

ER+ tumors (red bar), but were identified as ERBB2+ upon close inspection of expression of ERBB2+-related genes (zoom up of clustergram). This is due to the Rosetta microarray possessing a much higher number of genes related to the ER+ subtype than the ERBB2 subtype.

Supplementary Data S9 : Using NPI-ES expression levels to Subtype the Rosetta ER+ samples

From Fig S8, we identified 49 ER+ tumors in the Rosetta data set. The NPI-ES clustered these 49 ER+ tumors into two distinct groups (Fig S9).

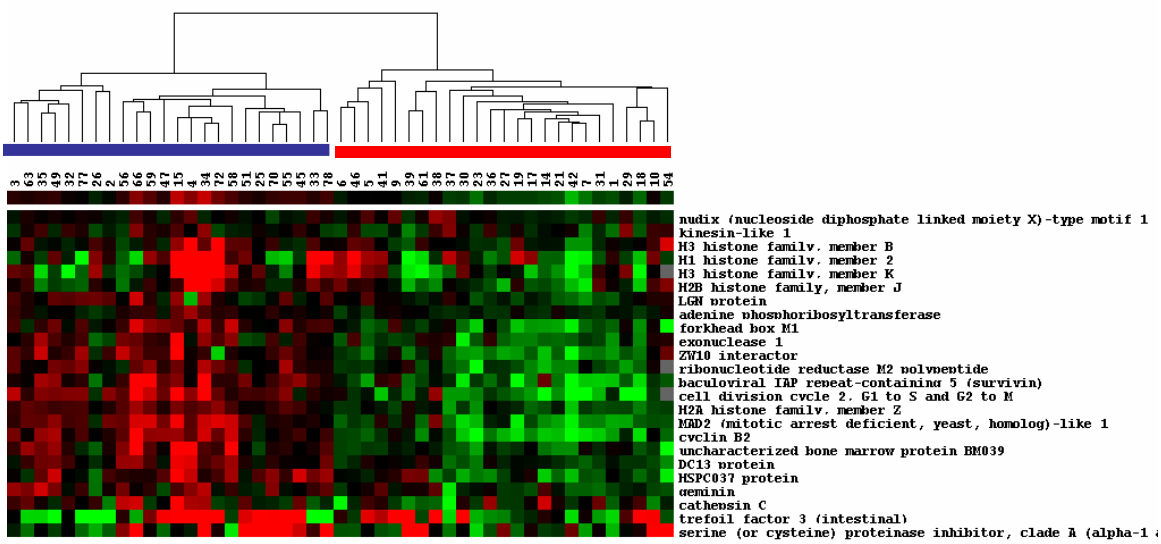

Figure S9. Hierarchical clustering Rosetta ER+ samples (49) based upon the expression level of the NPI-ES (46 matches found in Rosetta data out of 62 genes).The color bar is defined as Figure 2b) (main text).

Supplementary Data S10 : Identifying ER+ Tumors in the Stanford Data Set using the SAM-409 Gene Set

136 genes belonging to the SAM-409 gene set were found on the Stanford microarray. (http://genome-www5.stanford.edu/MicroArray/SMD/). This set of 136 overlapping genes was used to classify the Stanford data (72 tumors after discarding the normal-like tumor subgroup (6 tumors), as this latter group is likely to be due to the presence of contaminating non-malignant tissue). From Fig S8 we identified 46 ER+ samples, of which 45 belong to the 'Luminal' subtype reported by Sorlie et al., (ref 14).

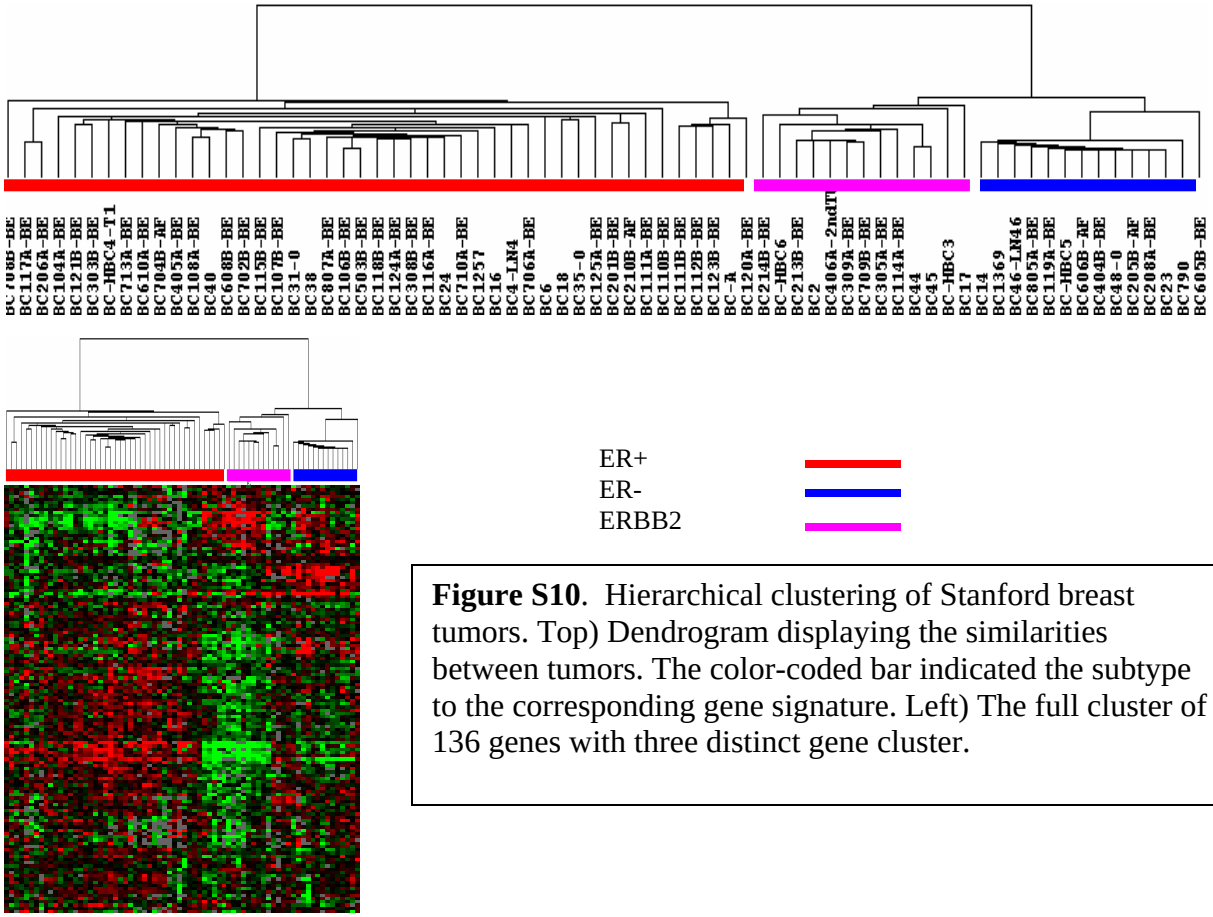

Figure S10. Hierarchical clustering of Stanford breast tumors. Top) Dendrogram displaying the similarities between tumors. The color-coded bar indicated the subtype to the corresponding gene signature. Left) The full cluster of 136 genes with three distinct gene cluster.

Supplementary Data S11 : Using NPI-ES expression levels to Subtype the Stanford ER+ samples

From Figure S10 we identified 46 ER+ tumors. The NPI-ES clustered these 46 ER+ tumors into two distinct groups (Fig S11).

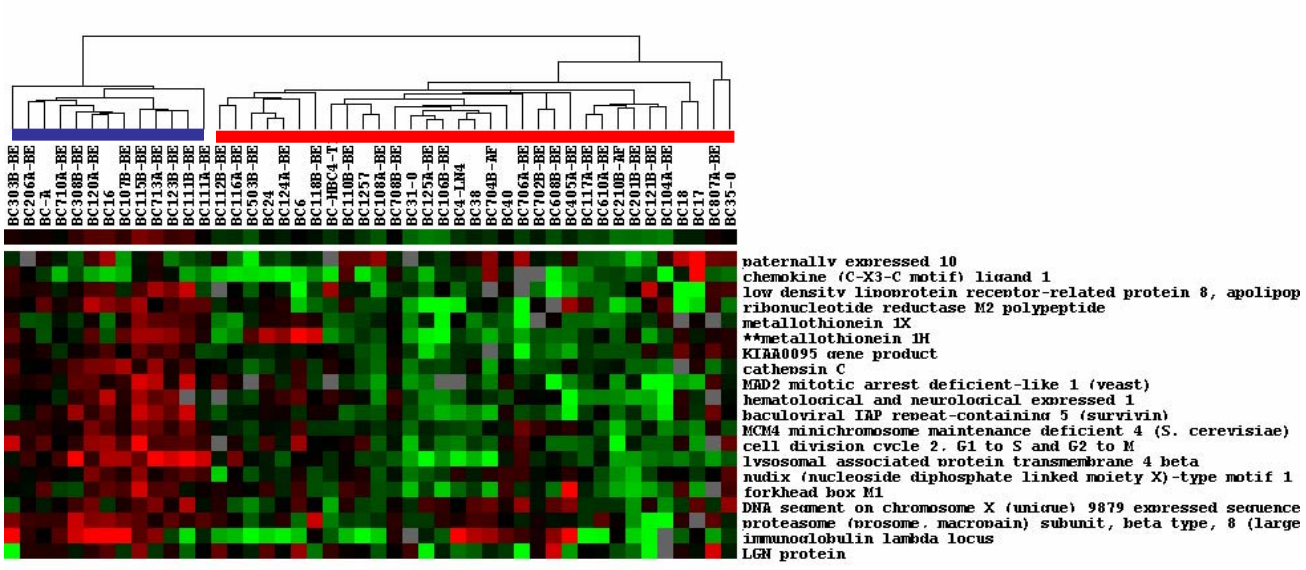

Figure S11. Hierarchical clustering of Stanford 46 ER+ samples using NPI-ES (31 matches out of 62 genes). The color bar is defined as Figure 2b) (main text).

Interestingly, Sorlie et al., (ref. 14), previously reported the identification of a “Luminal C” subtype based upon an ‘intrinsic’ set of 500 genes. Although none of the 62 genes belonging to the NPI-ES are found in this ‘intrinsic’ set, there appears to be a strong overlap (96%) between ‘Luminal C’ tumors and tumors expressing high levels of the NPI-ES. This is illustrated in Table S11.

	Luminal A	Luminal C
Low NPI-ES	30	0
High NPI-ES	2	10

Table S11 : Correlation of Luminal A and Luminal C Tumors with High and Low NPI-ES Expression (Luminal Tumors were identified based upon results of Sorlie et al., (2001))

Supplementary Data S12 : NPI-ES Expression Does not Correlate to NPI Status in the ER- and ERBB2+ Molecular Subtypes

The NPI status of ER- and ERBB2 tumors is in general higher than ER+ tumors. Unlike the case for ER+ tumors, we were unable to identify by SAM genes that were differentially regulated in high vs low NPI tumors for the ER- and ERBB2+ subtypes.

Also, NPI-ES does not appear to be correlated as well to NPI values associated with the other molecular subtypes (Fig. S12).

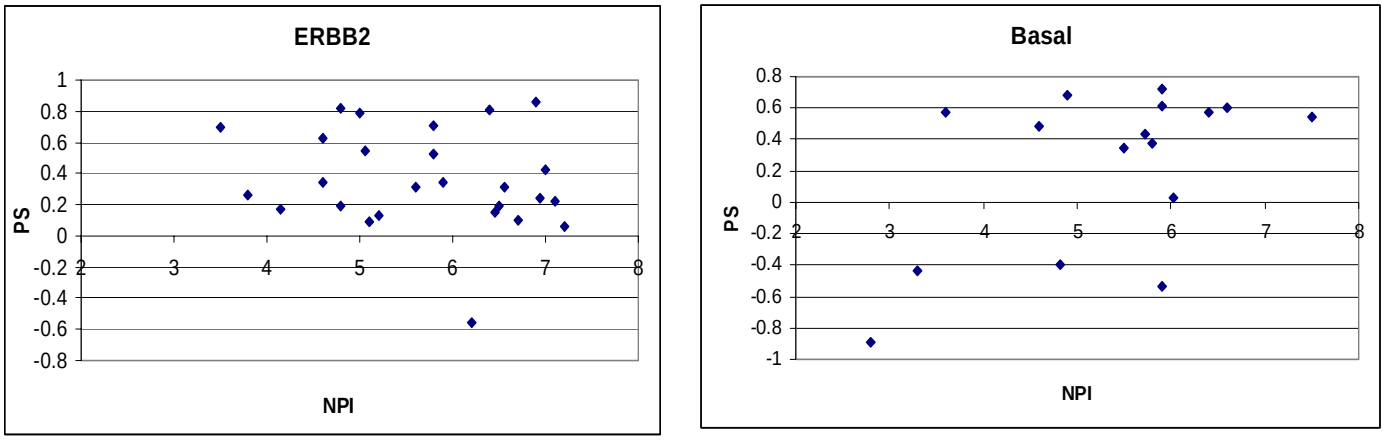

Figure S12. Weighted voting, trained by ER+ samples, is used to predict NPI status of ERBB2 and Basal samples. There are no distinct binary groups (ie, below vs. above X-axis) observed in both results.