A Modular Analysis of Breast Cancer Reveals A Novel Low-Grade Molecular Signature in Estrogen Receptor Positive Tumors

(Supplementary Information)

Kun Yu[1], Kumaresan Ganesan[2], Lance D. Miller[3], and Patrick Tan[1,2,3,*]

[1] National Cancer Centre / [2] Agenica Research

11 Hospital Drive

Singapore 169610

Republic of Singapore

[3] Genome Institute of Singapore

60 Biopolis Street

Singapore 138672

Republic of Singapore

* Address correspondence to cmrtan@nccs.com.sg

Tel : 65-6-436-8345

Fax : 65-6-226-5694

Supplementary Information S1 : Patient and Tissue Sample Information

Table S1. Clinical information for Breast Tumors

Sample ID	Age	Size (mm)	Grade	LN	ER	PR	LVI
980058	72	45	3	0 of 12	pos	pos	no
980177	75	26	2	6 of 13	pos	pos	yes
980178	69	32	3	2 of 15	pos	neg	no
980193	49	25	3	3 of 23	neg	neg	no
980194	58	50	3	25 of 32	neg	neg	yes
980197	55	30	3	2 of 4	pos	pos	yes
980203	44	15	1	0 of 11	pos	pos	no
980208	42	25	3	5 of 20	pos	pos	no
980214	49	60	2	5 of 13	pos	neg	no
980215	50	30	2	8 of 23	pos	neg	no
980216	65	45	2	5 of 20	neg	neg	no
980217	50	30	2	7 of 12	pos	neg	yes
980220	40	37	2	0 of 5	pos	pos	yes
980221	33	65	3	1 of 13	pos	pos	no
980238	62	20	3	7 of 21	neg	neg	no
980247	35	45	3	1 of 19	neg	neg	yes
980256	46	36	3	1 of 12	neg	neg	no
980261	60	15	2	0 of 9	pos	neg	no
980278	64	40	3	14 of 20	pos	neg	yes
980285	49	40	3	1 of 7	neg	neg	yes
980288	45	60	3	13 of 15	pos	neg	yes
980315	59	45	3	0 of 19	neg	neg	yes
980333	51	40	3	2 of 7	pos	pos	no
980335	33	3	3	3 of 7	neg	neg	yes
980338	55	30	3	0 of 7	neg	neg	no
980346	52	20	3	0 of 4	pos	pos	possible
980353	58	45	3	0 of 25	neg	neg	no
980373	77	30	3	0 of 14	neg	neg	no
980380	56			0 of 6	neg	neg	
980383	64	30	2	0 of 16	pos	pos	no
980391	56	20	2	0 of 7	pos	pos	no
980395	68	30	3	1 of 10	neg	neg	yes
980396	66	35	3	10 of 12	neg	neg	yes
980403	73	30	3	0 of 9	pos	pos	possible
980404	46	30	2	1 of 5	pos	pos	yes
980409	48	15	2	0 of 19	pos	neg	no
980411	69	30	2	0 of 9	neg	neg	no
980434	73	30	3	0 of 16	pos	pos	no
980441	66	30	3	4 of 14	neg	neg	yes
990075	66	25	3	5 of 21	pos	pos	yes

990082	49	34	2	3 of 16	pos	pos	no
990107	50	40	1	1 of 18	pos	neg	yes
990113	70	90	3	11 of 15	pos	pos	no
990115	38	28	3	9 of 10	pos	pos	yes
990123	54	55	3	7 of 11	pos	pos	no
990134	43	40	3	0 of 19	neg	neg	no
990148	60	40	2	6 of 19	pos	neg	yes
990174	55	45	2	3 of 24	neg	neg	yes
990223	52	5	3	1 of 21	pos	neg	no
990262	68	40	3	4 of 14	neg	neg	no
990299	58	55	3	7 of 17	neg	neg	possible
990375	38	15	1	0 of 10	pos	neg	no
2000104	59				pos	neg	
2000171	50	25	2	0 of 9	neg	neg	no
2000209	58	50	3	0 of 7	pos	neg	no
2000210	50	40	3	3 of 6	neg	neg	yes
2000215	50	15	2	1 of 21	pos	pos	no
2000220	52	60	3	30 of 34	pos	neg	yes
2000237	43	47	3	23 of 40	pos	pos	yes
2000272	49	30	3	1 of 16	pos	neg	yes
2000274	40	35	3	10 of 23	pos	pos	yes
2000287	53	40	3	0 of 8	neg	neg	possible
2000320	67	20	3	20 of 21	neg	neg	yes
2000376	65		3	8 of 23	neg	neg	yes
2000399	44	40	2	0 of 8	neg	neg	no
2000401	51	50	3	2 of 6	neg	pos	no
2000422	51	63	3	3 of 7	pos	pos	no
2000500	44	75	3	6 of 6	neg	neg	yes
2000593	60	41	3	0 of 15	neg	neg	no
2000597	57	40	2	0 of 12	pos	neg	possible
2000609	62	70	2	17 of 17	pos	pos	yes
2000638	60	40	1	0 of 15	pos	neg	no
2000641	47	60	3	16 of 24	neg	neg	yes
2000651	45	41	2	3 of 5	pos	pos	yes
2000652	56	25	3	6 of 21	neg	neg	no
2000675	78	55	3	16 of 16	neg	neg	yes
2000683	72	35	2	0 of 17	pos	pos	no
2000709	45	30	3	0 of 16	neg	neg	no
2000731	68	51	3	1 of 29	pos	neg	no
2000759	57	7	3	0 of 12	neg	neg	no
2000768	39	40	3	0 of 17	pos	pos	no
2000775	51	25	2	0 of 12	pos	neg	no
2000779	48	55	3	0 of 14	pos	neg	no
2000787	57	60	3	0 of 9	pos	pos	yes
2000804	39	40	3	5 of 21	pos	pos	yes
2000813	60	23	3	16 of 17	neg	neg	yes
2000818	52	10	2	0 of 11	pos	neg	no
2000829	51	45	2	10 of 10	neg	neg	yes

2000880	55	15	2	0 of 26	neg	neg	no
2000948	56	35	3	4 of 22	pos	neg	yes
20020021	64	38	3	0 of 13	pos	neg	yes
20020051	38	50	3	1 of 25	pos	pos	no
20020056	71	20	1	2 of 17	pos	neg	no
20020071	58	28	3	0 of 16	pos	pos	no
20020090	60	45	3	19 of 27	neg	neg	yes
20020160	86	120	3	0 of 10	pos	pos	no

LN: lymph node; ER: estrogen receptor; PR: progesterone receptor; LVI: lymphovascular invasion

Histopathological Techniques

Human breast tissues were obtained from the NCC Tissue Repository, after appropriate approvals from the NCC Repository and Ethics Committees. Samples were grossly dissected in the operating theater immediately after surgical excision, and flash-frozen in liquid N2. Samples had not been treated with pre-operative chemotherapy. For histological assessment of tumors and axillary lymph nodes, formalin-fixed, paraffin-embedded tumor tissue was used to determine tumor subtype (WHO classification), histologic grade, and lymphovascular invasion. Tumor size, based only on the invasive component, was assessed macroscopically and confirmed microscopically. For small tumors, the size was measured on this histologic section. ER status was determined by immnohistochemistry, with a positive result being >10% of carcinoma cells showing nuclear reactivity of at least +2 intensity. For ERBB2 immunohistochemistry, the Dako classification system was used with scores of 0 and 1+ considered negative while 2+ and 3+ were positive. An indeterminate conclusion was made when benign breast epithelium was immunoreactive. Profiled samples contained at least 50% tumor content. The independent collection of second data set which includes 86 breast tumors (see Result) are short of clinical data.

Supplementary Information S2 : ISA work scheme

The Iterative signature algorithm (ISA) is an extension of the basic signature algorithm that can be used to globally decompose gene expression data. In general, the ISA is a self-feed system that operates as follows: 1) Initial generation of a sufficiently large sample of randomly assembled input seeds (gene sets); 2) Identification of robust modules corresponding to each seed through multiple iterations using the recurrence criterion, similar to SA (see Materials & Methods). Figure S2 depicts the ISA schema. A detailed technical report of the ISA can be found in Bergmann et al., 2003. The parameters used are shown in Table S2.

Figure S2. The workflow of the Iterative Signature Algorithm (ISA).

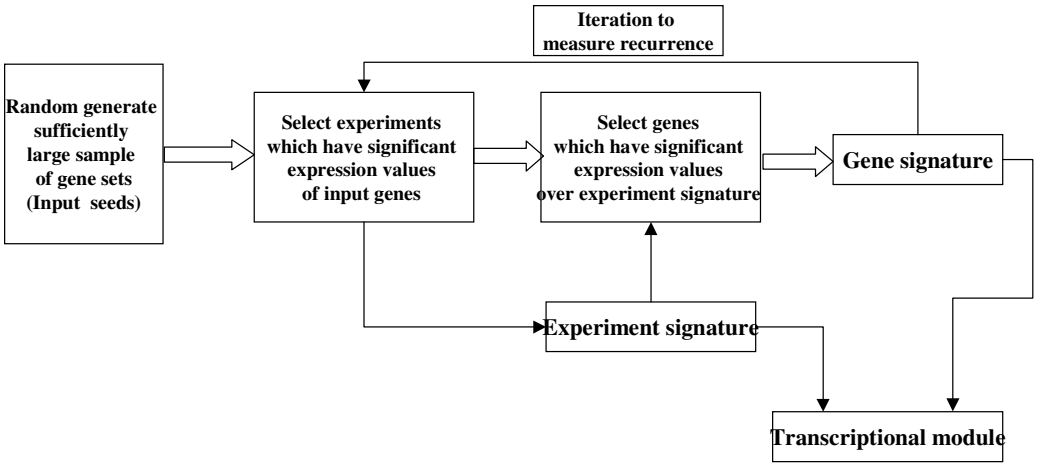


Table S2. Parameter Settings of ISA. Definitions of each parameter can be found in: http://barkai-serv.weizmann.ac.il/GroupPage/software.htm.

Parameter settings of ISA				
Condition Threshold	Gene Threshold range	minRecurrence	minNoGenes	randomSizes
3	[1.8, 4]	2	10	[5, 10:1:20]

Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys. 2003 Mar;67(3 Pt 1):031902.

Supplementary Information S3: Summary of TuMs derived from breast cancer

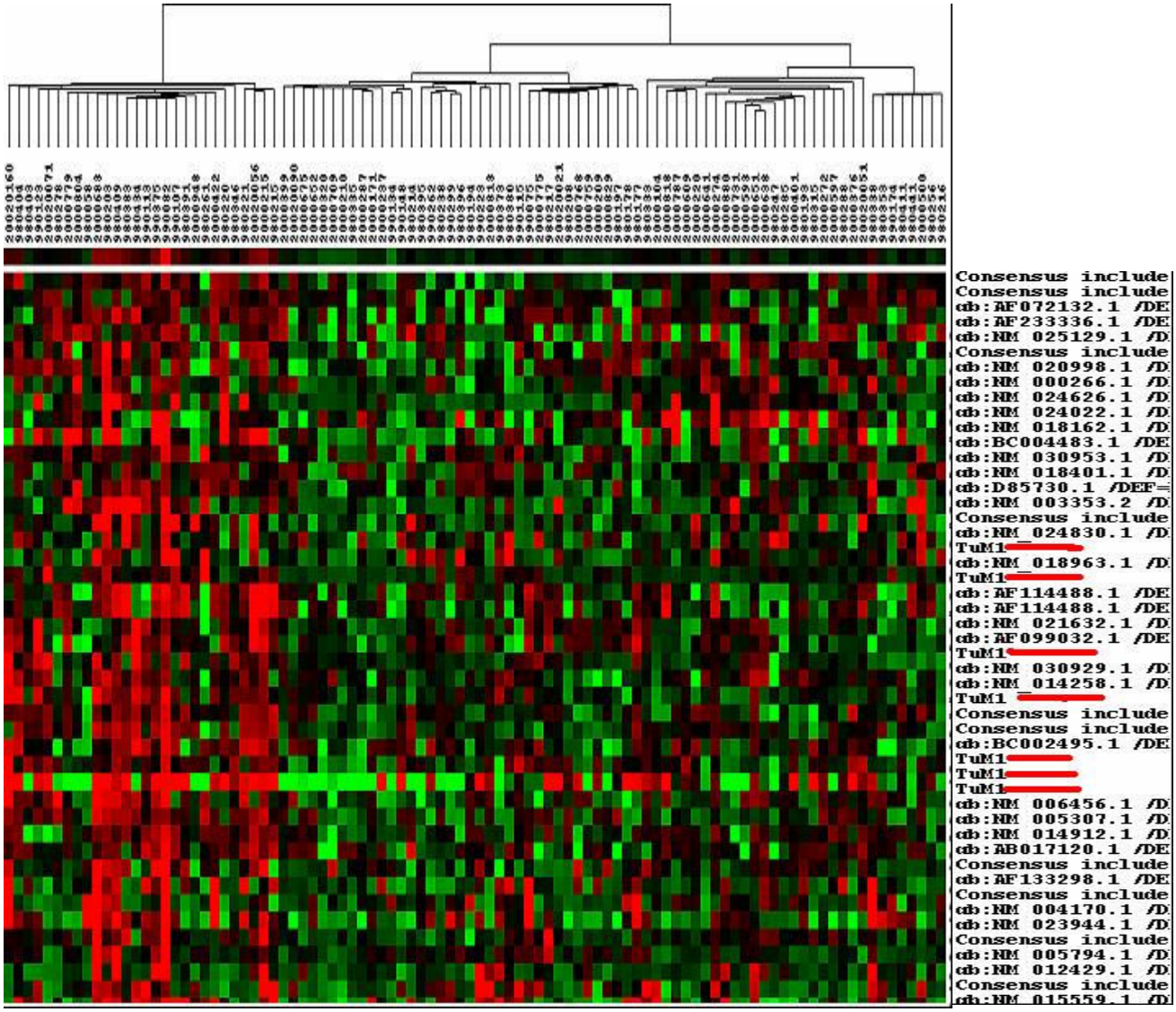
In addition to TuMs 1-3, the gene content of TuMs 4-8 could also be mapped to many previously defined gene expression signatures in breast cancer (ref 3-6 in the main paper): TuM4 consists of a large set of genes involved in immune function, including immunoglobulin genes, T cell receptor subunits, and TNF family members (3), while TuM5, containing *FBLN1*, *SPARC* and various collagens, are likely to represent contributions from the stromal cell population (3). TuM6, containing *Keratin 5*, *Keratin 17*, and *SFRP1*, corresponds to the expression signatures of Basal/ER- cancers (3–6), and TuM7 contains a significant number of genes ($p<10^{-4}$) belonging to the NPI-ES expression signature (ref. 33 in the main text), previously identified as a molecular surrogate of the Nottingham Prognostic Index, as well as genes involved in cellular proliferation (eg, *MAD2L1*, *CDC2*). Finally, TuM8 contained several genes physically linked to the 17q21 locus (eg, *v-erb-b2*, *GRB7*, *PNMT*), corresponding to a previously reported ERBB2 cluster (1–4). We also performed Gene Ontology (GO) statistics analysis to identify which GO terms/Pathways are enriched in each TuMs. As expected, the results showed a good agreement with 'known functions' of gene clusters. For example, we found out that cell cycle genes is significantly overexpressed ($p=4.08*10^{-16}$) in TuM7 (cell proliferation); meanwhile extracellular matrix (ECM) ($p=2.85*10^{-6}$) and collagen binding ($p=8.72*10^{-6}$) are enriched in TuM5 (stromal). The full details of GOStat analysis are available at http://www.omniarray.com/Breast_TuM/.

Taken collectively, these results validate the biological consistency of the TuMs. More generally, they demonstrate that despite being an entirely unsupervised analytical approach, the ISA appears to be remarkably efficient at re-discovering many, if not all, of the major gene expression signatures previously reported for breast cancer identified by conventional analysis.

Supplementary Information S4 : HC on the entire ISA-input gene set

The hierarchical clustering was also performed on the entire ISA-input gene set of 9116 probes. The similar results were obtained to that using the 1500 probe set (see Figure 2 in the Main Text). Specifically, the TuM1 genes are scattered across the estrogen receptor (ER) cluster and do not form a distinct sub-cluster. The TuM1 genes fall into several sub-cluster and did not form an obvious module within it (Figure S4). Furthermore, one TuM1 member gene segregated outside the ER cluster. This result supports our hypothesis that the TuM1 module cannot be observed using conventional hierarchical clustering.

Figure S4. A sub-cluster is cut from the entire ER cluster. Only 7 (out of 33) TuM1 members (marked in red line) lied in this sub-cluster. And they did not form a tight group under the subcluster. The complete tree view file can be downloaded via: http://www.omniarray.com/Breast_TuM/**.**



Supplementary Information S5 : Gene and tumor content of TuMs

Table S5 provides a list of the TuM1 genes. The six apoptosis-related genes reported by the Ingenuity system (see Results in the main text) are highlighted in bold. The full list of gene and tumor contents for all eight TuMs are available via: http://www.omniarray.com/Breast_TuM/.

Table S5. Co-regulated genes in TuM1.

Probe	Gene Name	Unigene
218613_at	hypothetical protein DKFZp761K1423	Hs.236438
203355_s_at	ADP-ribosylation factor guanine nucleotide factor 6	Hs.408177
202731_at	**programmed cell death 4 (neoplastic transformation inhibitor)**	**Hs.257697**
214440_at	N-acetyltransferase 1 (arylamine N-acetyltransferase)	Hs.458430
203404_at	armadillo repeat protein ALEX2	Hs.48924
202174_s_at	pericentriolar material 1	Hs.348501
217838_s_at	Enah/Vasp-like	Hs.241471
219455_at	hypothetical protein FLJ21062	Hs.276466
221946_at	hypothetical protein MGC29761	Hs.414028
222314_x_at	Homo sapiens, clone IMAGE:5759947, mRNA	Hs.437867
211596_s_at	**leucine-rich repeats and immunoglobulin-like domains 1**	**Hs.166697**
211538_s_at	**heat shock 70kDa protein 2**	**Hs.432648**
214705_at	InaD-like protein	Hs.436450
218398_at	**mitochondrial ribosomal protein S30**	**Hs.124165**
201667_at	**gap junction protein, alpha 1, 43kDa (connexin 43)**	**Hs.74471**
215300_s_at	flavin containing monooxygenase 5	Hs.396595
209884_s_at	solute carrier family 4, sodium bicarbonate cotransporter, member 7	Hs.250072
212196_at	**interleukin 6 signal transducer (gp130, oncostatin M receptor)**	**Hs.71968**
200648_s_at	glutamate-ammonia ligase (glutamine synthase)	Hs.442669
214519_s_at	relaxin 2 (H2)	Hs.127032
219114_at	g20 protein	Hs.21050
206081_at	solute carrier family 24 (sodium/potassium/calcium exchanger), member 1	Hs.173092
214430_at	galactosidase, alpha	Hs.69089
221562_s_at	sirtuin (silent mating type information regulation 2 homolog) 3 (S. cerevisiae)	Hs.511950
218149_s_at	hypothetical protein DKFZp434K1210	Hs.32352
214087_s_at	myosin binding protein C, slow type	Hs.169849
213933_at	*prostaglandin E receptor 3 (subtype EP3)*	*Hs.27860*
215014_at	Homo sapiens mRNA; cDNA DKFZp547P042 (from clone DKFZp547P042)	Hs.232127
203143_s_at	KIAA0040 gene product	Hs.368916
204901_at	beta-transducin repeat containing	Hs.226434
209123_at	quinoid dihydropteridine reductase	Hs.75438
213832_at	Homo sapiens clone 24405 mRNA sequence	Hs.23729
207519_at	solute carrier family 6 (neurotransmitter transporter, serotonin), member 4	Hs.448453

Supplementary Information S6: Association between TuMs and Clinical Parameters

Every tumor module is associated with a set of tumors. The significance of each tumor is characterized by a score (the "tumor score"). A positive or negative score indicates that in this tumor the genes are upregulated or downregulated, in comparison to the rest of the tumor population. In the present study, we have only studied tumors with positive scores because tumors with negative scores are very rare (only three modules had tumors with negative scores; see Fig. S6). We found that certain tumors with low tumor score are clearly apart from others (those in red rectangle). These tumors were treated as "low confidence" samples and removed them from subsequent correlation analysis.

Statistical approaches were then used to discover the clinical significance of these transcriptional modules. The results revealed a number of significant associations between modules and clinical characteristics. At $p<0.001$, only ER/PR status and tumor grade are likely to be associated with gene expression data, which was also observed by ref. 6 (Table S6). TuM4, the immune cluster, was negatively correlated with ER positivity and marginally positively correlated with high grade ($p=0.02$). This result is consistent with the report that Immunoglobin genes comprise the majority of 'ER-' genes (Iwko et al., 2002). TuM5, the predominantly stromal cell cluster, was not associated with any clinical parameters. As expected, TuM6 and TuM8, representing the ER-/Basal and ERBB2+ molecular subtypes respectively, were significantly negatively correlated with ER ($p<0.001$). TuM7, the cell proliferation cluster, is significantly correlated with high histological grade but not correlated with ER status.

Figure S6. Tumor Scores of transcriptional modules. Tumors are sorted by their tumor score. Y-axis is the tumor score. X-axis is the index of the tumor, which varies in different modules.

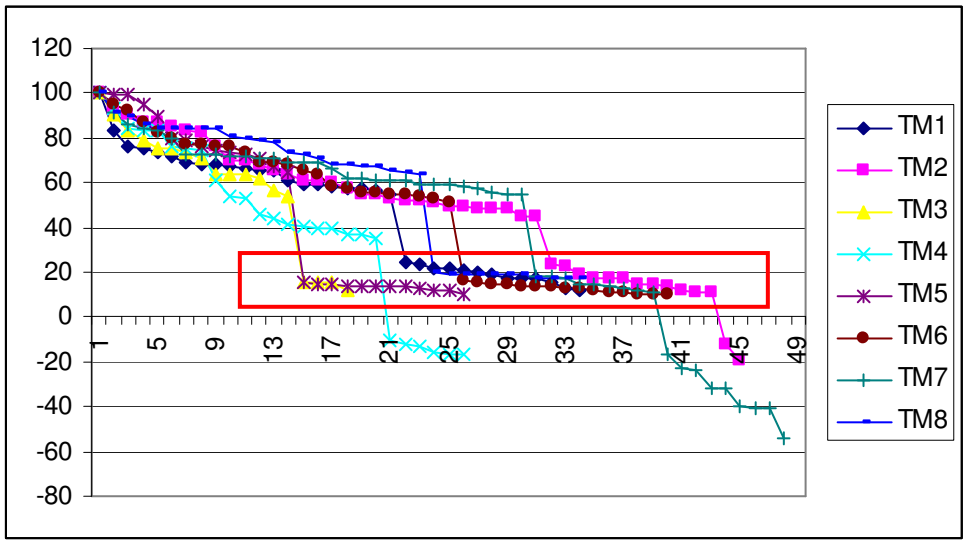


Table S6. Correlations between TuMs and Clinical Characteristics. Only association with P<0.05 are displayed.

	Age (≤/>55)	**Size** (≤/>3 cm)	**Grade**	**LN**	**ER**	**PR**	**LVI**
TuM1 (Low Grade)		0.0152 (≤3*)	<0.001		<0.001 (+)	0.0107 (+)	0.0152 (-)
TuM2 (ER+/Luminal)			0.005		<0.001 (+)	0.0021 (+)	
TuM3 (ER+ II)					<0.001 (+)	0.0015 (+)	
TuM4 (Immune)					0.0044 (-)		
TuM5 (Stroma)							
TuM6 (ER-/Basal)				0.0236 (+)	<0.001 (-)	0.0098 (-)	
TuM7 (Cell Proliferation)			<0.001				
TuM8 (ERBB2+)					<0.001 (-)	<0.001 (-)	

Iwao K, Matoba R, Ueno N, Ando A, Miyoshi Y, Matsubara K, Noguchi S, Kato K. Molecular classification of primary breast tumors possessing distinct prognostic properties. Hum Mol Genet. 2002 Jan 15;11(2):199-206.

Supplementary Information S7 : Associations between grade and TuM1

Univariate association by using ER+ breast tumors only

We also repeated the univariate association studies, this time using a sample set of *only* ER positive tumors (unlike the previous analysis where all tumors were used). In this "ER positive only" data set, we found that TuM1 still remained significantly correlated with ***low*** tumor grade ($p<0.001$). In contrast, TuM2 and TuM3, which both contain several ER-related genes, failed to exhibit a significant correlation with tumor grade when the ER negative tumors were removed from the analysis ($p=0.16$ and $p=0.34$ respectively) (Table S7a). This association is thus generally applicable and observed in four independent data sets (Table S7b).

Table S7a. Correlation between TuMs 1, 2, 3 and grade within ER positive tumors. In this analysis, only the ER positive tumors were used in the sample population. 1st column for each module indicated the number of TuM-overexpressed tumors.

	TuM1		**TuM2**		**TuM3**	
Grade	**P<0.001**		**P= 0.155**		**P= 0.341**	
1	5	0	4	1	1	4
2	7	12	8	11	7	12
3	2	30	11	21	6	26

Table S7b. Correlation between TuM1 and grade within ER+ tumors in four public data sets. 1st column represents TuM1 overexpressed tumors.

	Stanford		**Rosetta**		**Ma**		**Uppsala**	
Grade	**0.0002**		**<0.001**		**0.023**		**0.005**	
1	6	3	3	9	3	0	7	5
2	26	14	15	9	30	9	20	22
3	6	26	3	33	8	10	0	12

Multivariate analysis on four public data set

Multivariate analysis is performed by using linear regression analysis (SPSS). In all the four data sets, TuM1 is significantly correlated with grade set (Table S7c), independent of other clinical characters such as size, age. ER did not show a significant association with grade in Rosetta and Stanford data set (Table S7c). Ma data set (ref. 28) and Uppsala data set (ref. 29) consist of ER+ tumors only, which obviously means ER is not correlated with grade in these two cases. Taken together, our results show that TuM1 is directly associated with grade, not subject to ER.

Table S7c. Multivariate analysis of associations between grade and TuM1, as well as various clinical characters. The independent significant association ($p<0.05$) is displayed in bold text. The positive regression coefficient means the variable is associated with low grade.

Ma	P-Value	Regression Coefficient	95% Confidence Interval for Regression Coefficient	
Variable			Lower Bound	Upper Bound
TUM1	**0.015**	0.395	0.082	0.707
SIZE	0.497	0.045	-0.087	0.176
NODE	0.929	0.013	-0.281	0.307
AGE	0.704	0.003	-0.014	0.021

Rosetta	P-Value	Regression Coefficient	95% Confidence Interval for Regression Coefficient	
Variable			Lower Bound	Upper Bound
METAST	**0.022**	0.323	0.048	0.597
TUM1	**0.014**	0.414	0.085	0.744
ANGIOINV	0.197	0.178	-0.094	0.449
ER	0.269	-0.003	-0.007	0.002
LVI	0.242	0.209	-0.143	0.560
AGE	0.859	-0.002	-0.021	0.017
SIZE	0.138	0.012	-0.004	0.028
PR	0.688	-0.001	-0.005	0.003

Stanford	P-Value	Regression Coefficient	95% Confidence Interval for Regression Coefficient	
			Lower Bound	Upper Bound
TUM1	**<0.001**	0.499	0.230	0.768
AGE	0.270	-0.005	-0.013	0.004
SIZE	0.894	0.010	-0.143	0.163
NODE	0.634	0.037	-0.116	0.190
ER	0.119	-0.229	-0.518	0.060

Uppsala	P-Value	Regression Coefficient	95% Confidence Interval for Regression Coefficient	
			Lower Bound	Upper Bound
P53	**0.008**	0.469	0.126	0.812
AGE	0.202	0.008	-0.004	0.021
SIZE	0.100	0.012	-0.002	0.026
TUM1	**0.017**	0.330	0.062	0.599

Supplementary Information S8 : Common sets with public data sets successfully recapture TuM1 tumors

We have stated the number of overlapping TuM1 genes in the different data sets in the Main Text : 13 genes in Stanford data, 21 genes in Rosetta, 21 genes in Ma and all 33 TuM1 genes in Veridex and Uppsala (the latter two utilizing the same microarray platform as our original cohort). As shown in latter (S11 and S15), the smaller overlapping gene sets in the first three cohorts (Stanford, Rosetta, and Ma) are still sufficient to distinctly segregate the tumors into distinct groups on the basis of their overall expression ratios. To further show that the smaller overlapping gene sets are indeed sufficient to identify the TuM1 expression tumor population, we have now re-tested the overlapping sets used in the Rosetta, Stanford, and Ma data sets in our original tumor cohort, and confirmed that all three overlapping sets can successfully re-identified almost all 14 TuM1 tumors (Table S8). This result showed that these smaller subsets of TuM1 can also function as accurate surrogates of the entire TuM1 tumor module.

Table S8: Common gene sets successfully identify TUM1-overexpressed tumors.

#common genes		TuM1	Non-Tum1
Stanford	13	14	0
Rosetta	21	13	0
Ma	21	12	0

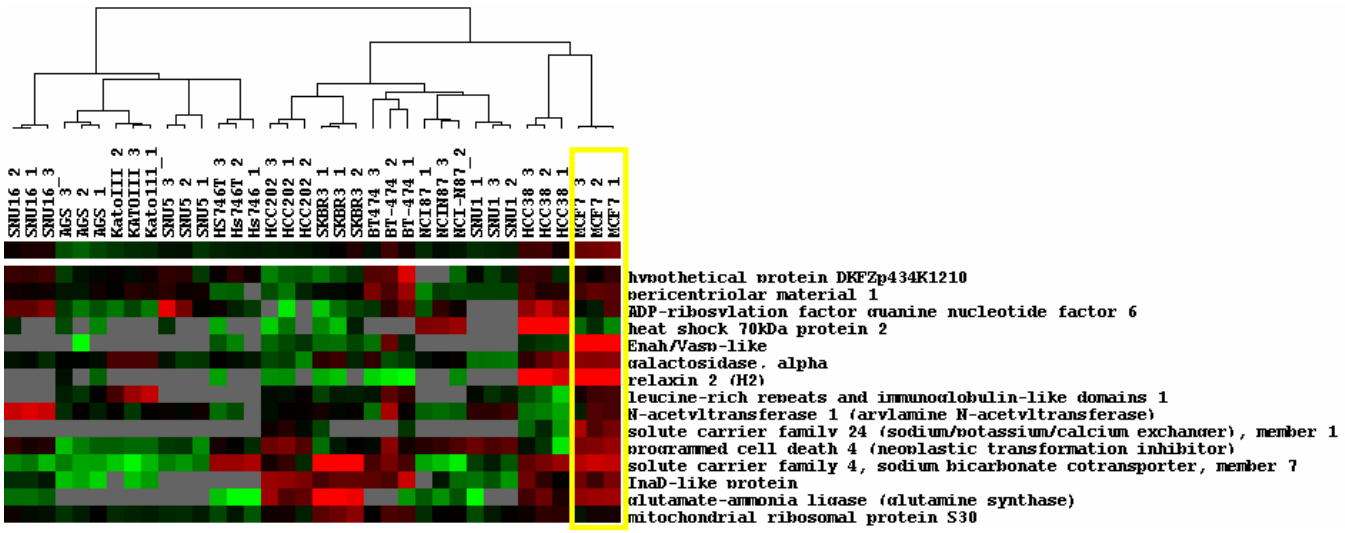
Supplementary Information S9 : TuM1 expression levels in Breast Cancer Cell Lines

We profiled MCF7, HCC38, BT474, SKBR3 and HCC202 (Table S9) on HG-U133 plus gene chip (Affymetrix Inc., Santa Clara, CA). 10 out of 33 TuM1 genes were filtered out because of insufficient valid values. Almost half of the remaining 23 TuM1 genes are overexpressed in MCF7 cell line. Seven gastric cell lines: AGS, SUN1, SUN5, SUN16, KatoIII, Hs746, N87, which are profiled on U133A chip (Aggarwal et al, 2005), have also been included as a reference. Figure S9 showed that the overexpression of TuM1 genes is dominant in MCF7 cell line.

Table S9. The ER and HER2/neu status of the breast cancer cell lines. The data are derived from ATCC (www.atcc.org).

Breast Cell Line	ER	HER2
HCC202	ER-/PR-	HER2/neu +
BT474	N/A	N/A
HCC38	ER-/PR-	HER2/neu -
MCF7	ER+	N/A
SKBR3	N/A	overexpresses the HER2/c-erb-2 gene product

Figure S9. Hierarchical clustering of various cell lines on the basis of expression profiling of TuM1 genes. Average-linkage hierarchical clustering employing a Pearson correlation metric was used in this analysis. The overexpression of TuM1 genes in MCF7 is highlighted in a yellow rectangle.



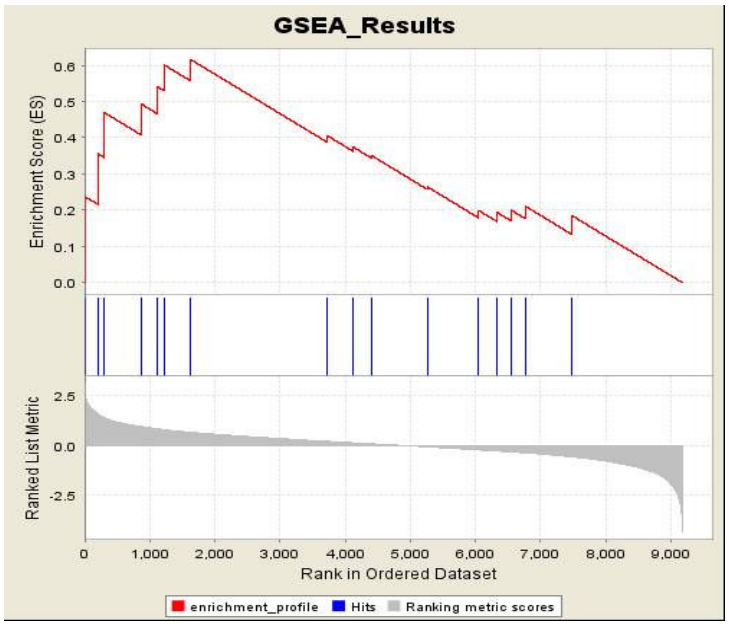
Aggarwal A, Leong SH, Lee C, Kon OL, Tan P. Wavelet transformations of tumor expression profiles reveals a pervasive genome-wide imprinting of aneuploidy on the cancer transcriptome. Cancer Res. 2005, 65(1):186-94.

Supplementary Information S10 : Downregulation of TuM1 in MCF7 cell line with tam-treatment

Gene set enrichment analysis (GSEA, ref. 16) to ask if expression of the tumor module genes might be affected by tamoxifen treatment. Four control samples and two post-treatment samples (See Materials and Methods) were used for GSEA analysis. Three modules (TuM4, 5 and 6) were filtered out due to insufficient number of genes (<10) expressed in MCF7 cell lines. TuM1 is the sole module showed a significant correlation with control samples (ie, downregulated in treated MCF7 cell line; see table and figure below).

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
downregulated in treated MCF7 cells						
TuM1	16	0.616471	1.6929	0	**0.05**	0
TuM2	16	0.727534	1.426171	0	0.19	0.15
TuM7	33	0.797655	1.320043	0.159574	0.216667	0.37
TuM3	10	0.588948	1.24243	0.146341	0.266667	0.45
upregulated in treated MCF7 cells						
TuM8	25	-0.51	-1.18	0.429	0.34	0.38

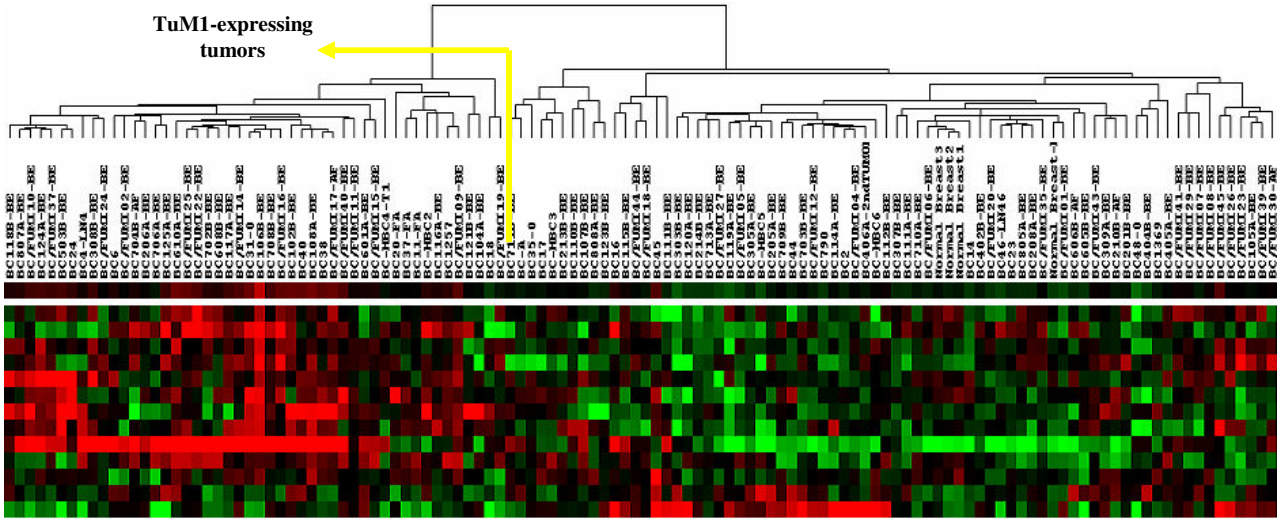
Figure S10. Genes are ranked by the signal-to-noise (S2N) ratio on control vs. treated cell line. The higher S2N ratio (rank), the lower expression values in treated cell line compared to control.



Supplementary Information S11 : TuM1 in Stanford Data Set

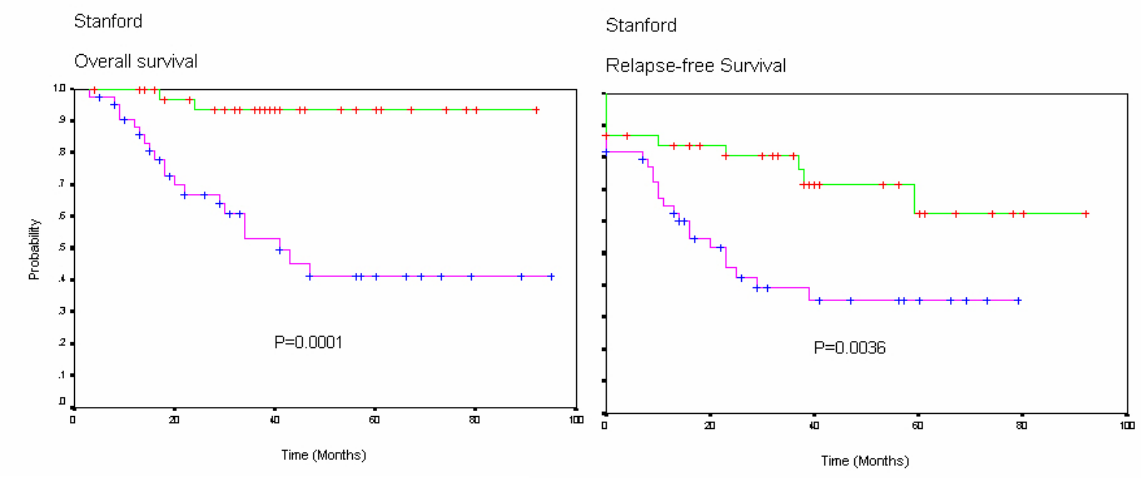
In the Stanford patient cohort, the Norwegian patients had presurgical chemotherapy and all patients expressing ESR1 received adjuvant endocrine treatment (reference 5 in the main text). The details of chemotherapy can be found at the website: http://genome-www.stanford.edu/breast_cancer/robustness/data/SupplText.html.

Hierarchical clustering was applied to classify Stanford ER+ based on the expression level of TuM1 genes. As shown in the figure, 81 ER+ tumors were formed into two clearly distinct groups. The distributions of the three different patient cohorts (with different treatment regimens) provided in the following tables, where the 1st column represents the number of ER+ tumors exhibiting TuM1 overexpression. Importantly, there is no apparent bias between TuM1-expressing tumors and TuM1-nonexpressing tumors, in terms of their proportions in the different patient cohorts. Thus, different treatment regimens are unlikely to confound this analysis. Kaplan-Meier analysis shows that patients with TuM1-expressing ER+ tumors exhibited better survival outcomes compared to patients with ER+ tumors where TuM1 was not expressed (p=0.0001 for overall survival; p=0.0036 for relapse-free survival).



ER+ tumors	TuM1	
cohort 1	21	26
cohort 2	11	12
cohort 3	6	5

ER+ tumors	TuM1	
Norwegian	31	39
Stanford	7	4



Supplementary Information S12 : Multivariate analysis of prognosis significance on the three independent data sets

Multivariate analysis using Cox regression was performed using the software package SPSS. These results show that TuM1 is an independent predictor of survival outcome (Table S12); while grade is not. In the Ma data set, tumor grade is controlled by the experiment design and hence is not correlated with tamoxifen response. TuM1 remains significantly associated with survival in Uppsala data set; while ER, PR, grade, tumor size, and lymph node status did not (Table S12). Notably, there are no clinical characters consistently correlated with survival outcome. Taken together, TuM1 showed independent prognostic capability in multiple patient cohorts that received anti-hormonal treatment.

Table S12: Multivariate analysis of risk factors for death (Uppsala and Stanford) or metastasis (Ma) as the first event. Parameters found to be significant (P<0.05) in the COX proportional hazard model are shown in bold.

Ma	**p-value**	Hazard ratio (95% CI)
TUM1	**0.030**	0.4 (0.175-0.913)
SIZE	0.150	1.307 (0.908-1.883)
NODE(2)	0.532	1.308 (0.564-3.037)
NODE(1)	0.709	1.321 (0.306-5.704)
NODE	0.809	
GRADE	0.568	1.274 (0.555-2.923)
AGE	0.309	0.977 (0.935-1.021)

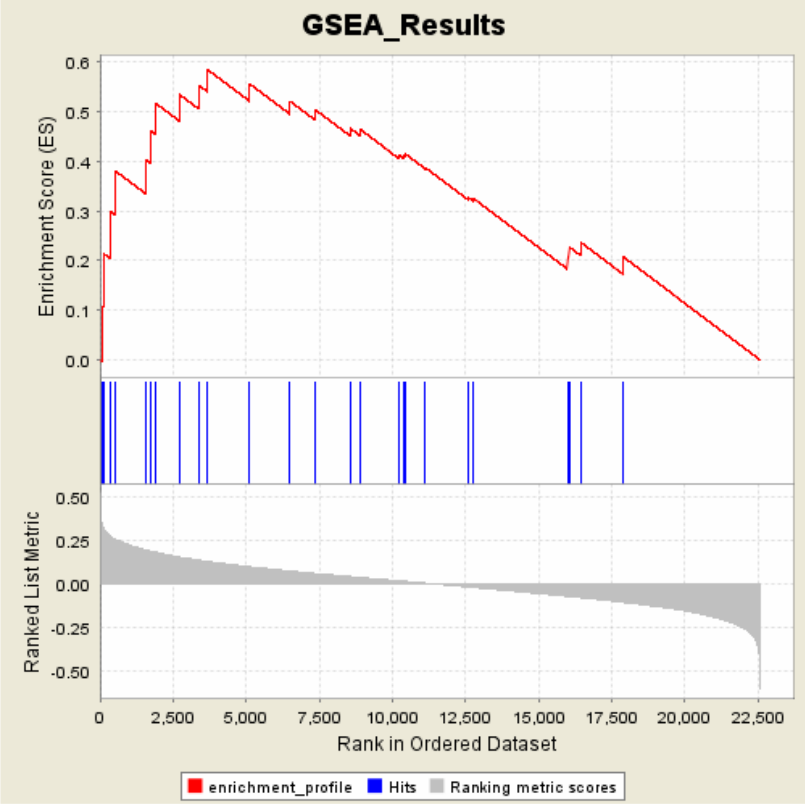
Uppsala	**p-value**	Hazard ratio (95% CI)
TUM1	**0.024**	0.27 (0.087-0.838)
SIZE	0.534	1.016 (0.967-1.067)
P53	0.998	0.999 (0.307-3.243)
NODE(2)	0.065	0.307 (0.088-1.075)
NODE(1)	0.983	
NODE	0.181	
GRADE	0.853	0.92 (0.38-2.226)
AGE	**0.016**	1.058 (1.01-1.108)

Stanford	**p-value**	Hazard ratio (95% CI)
TUM1	**0.003**	0.067 (0.012-0.388)
SIZE	0.113	2.206 (0.83-5.868)
NODE	0.439	0.801 (0.456-1.406)
METATASIS	**0.007**	5.822 (1.633-20.75)
GRADE	0.090	2.094 (0.892-4.917)
AGE	0.577	0.989 (0.951-1.028)

Supplementary Information S13 : GSEA on Ma data set

We ranked the genes in the Ma data set (22575 genes in total) according to their correlation to tamoxifen response, and used Gene Set Enrichment Analysis (GSEA) to assess the distribution of the TuM1 genes (25 common genes) in this ranking. To confirm the specificity of association between TuM1 genes and tamoxifen response status, we repeated the GSEA under conditions where the class labels of the samples (tamoxifen responsive or resistant) were randomly shuffled to generate two random sample groups, and the distribution of the TuM1 genes within these 'random' groupings was then compared to the distribution found in the original data. After 100 random trials, we found that TuM1 remained significantly associated with treatment response (p=0.024). There is no multiple test correction because it is a single test.

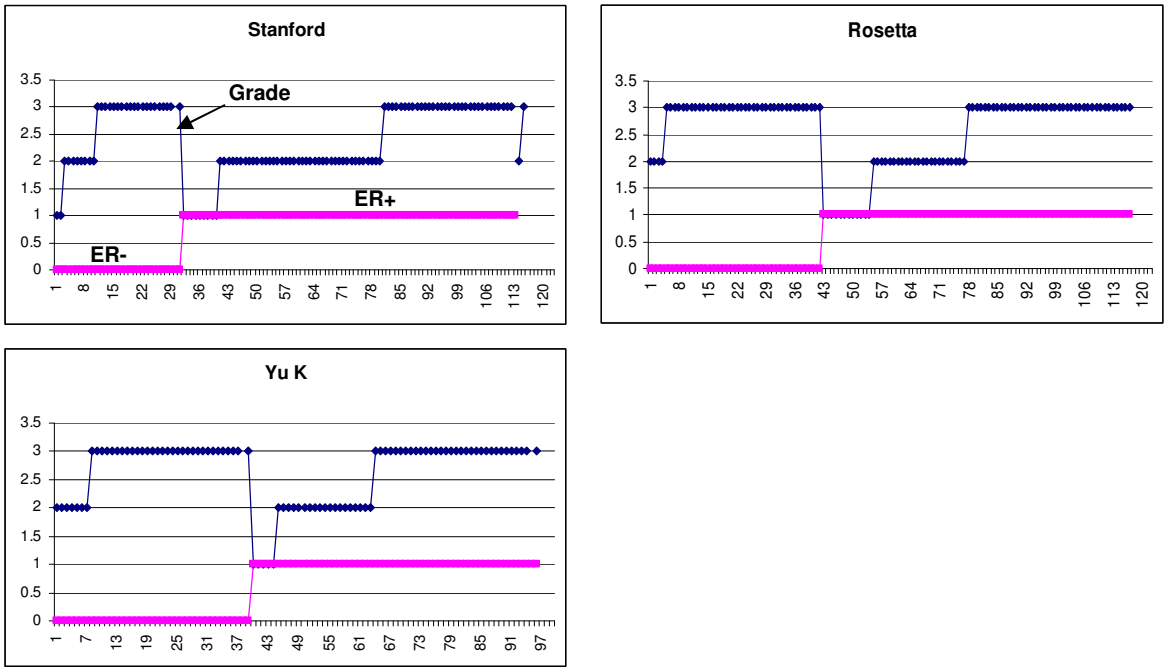
Figure S13. GSEA to determine the distribution of TuM1 genes in the list of genes from the Ma data set (19), which were ordered by their correlation with tamoxifen response status (X-axis). The running sum (Y-axis) of consecutive values of the Vector V indicates the distribution of TuM1 gene within the ordered genes.



Supplementary Information S14 : Correlation between Grade and ER Status

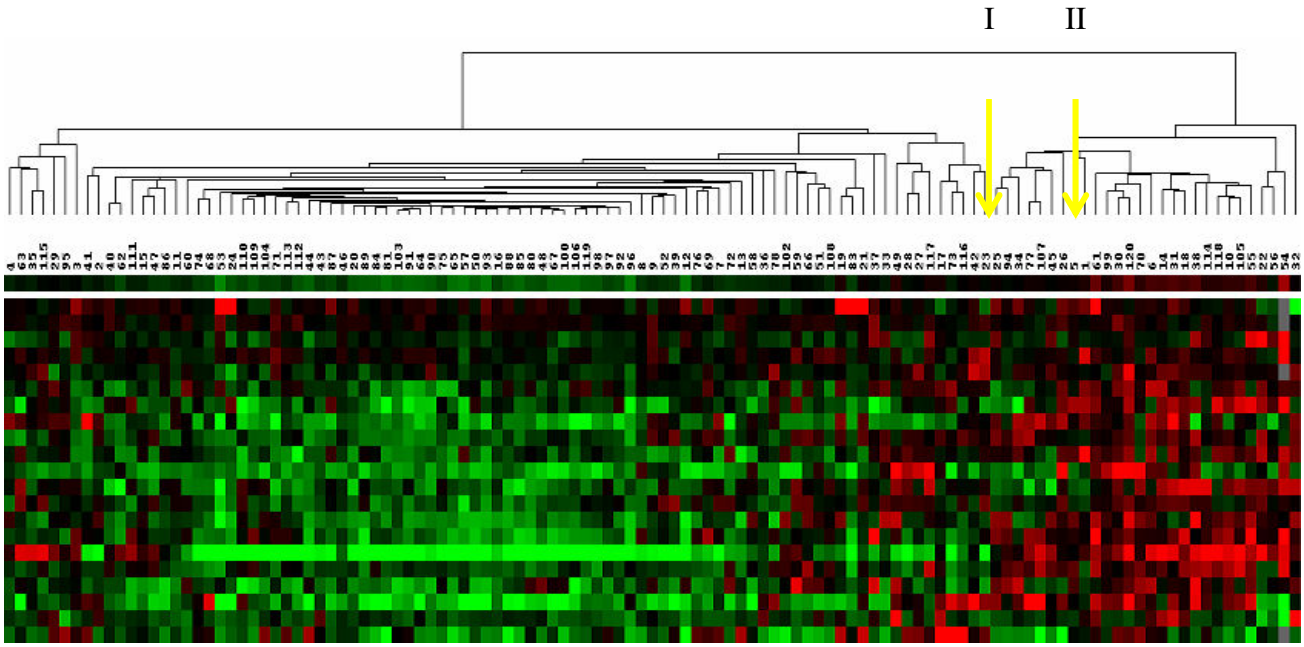
To study the relationship between grade and ER status, we surveyed three breast cancer data sets: 1) Standford data set (ref. 5); 2) Rosetta data set (ref. 27); and 3) our in-house data set; Fig S14 shows the relationship between grade and ER status for each breast tumor. In all three data sets, we observe that most of the ER negative tumors are high-grade (grade 3).

Figure S14: The distribution of grade and ER status in various breast cancer data sets. The dark blue line is grade; the pink line is ER status. Y-axis showed the grade (1-3). ER-positive was assigned as 1; while ER-negative was 0. The samples were sorted by ER, and by grade subsequently.

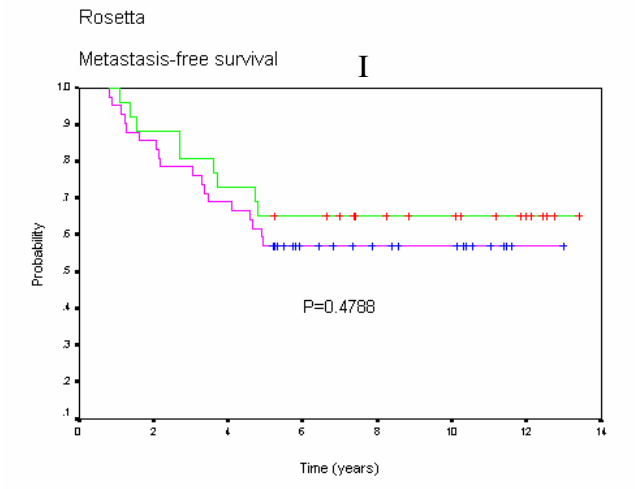


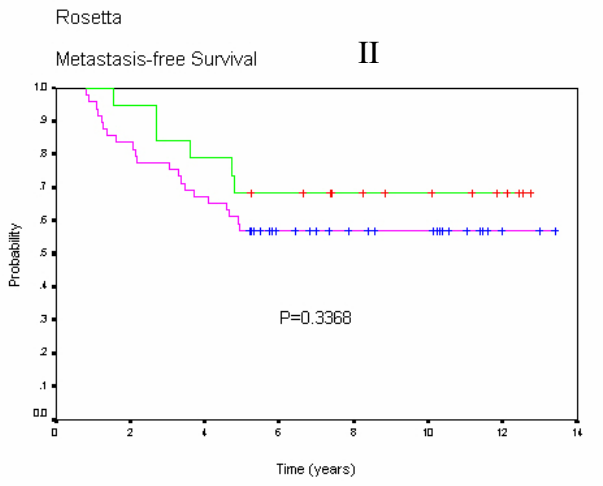
Supplementary Information S15 : TuM1 in Other Breast Cancer Data Sets that patients did not receive any treatment

Rosetta data set (ER positive tumors only)



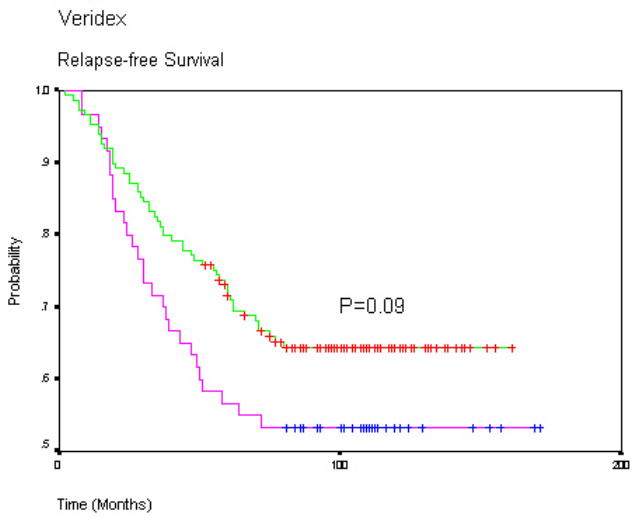
Rosetta Data set	I		II	
	Low-grade (1 &2)	High-grade (3)	Low-grade (1 &2)	High-grade (3)
	P=0.012 (Chi-square test)		**P<0.0001** (Chi-square test)	
TM1-expresed	19	9	18	3
Other ER+	16	27	17	33





Veridex data set

Another public data set (the "Veridex data set", Wang et al), comprising a cohort of lymph-node-negative breast cancer patients who did not receive systematic neoadjuvant or adjuvant therapy, was also studied. In the Veridex series, patients with TuM1-expressing ER+ tumors (209 samples) also failed to exhibit an improved clinical outcome compared to patients with ER+ tumors where TuM1 was not expressed (p=0.09 for disease-free survival). Grade information is not available for Veridex data set.



It is worth noting that in both the Rosetta and Veridex data set, patients with TuM1-overexpressing show a trend (Green Lines) to have better survival outcome than other patients with ER+ breast cancer, although this trend lacks statistical significance by the Kaplan Meier analysis.

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.
Lancet. 2005, 365(9460):671-9.