Supplementary Results

Pratyaksha J. Wirapati

April 19, 2007

Contents

1 Survival data

1.1 Data collection and preprocessing

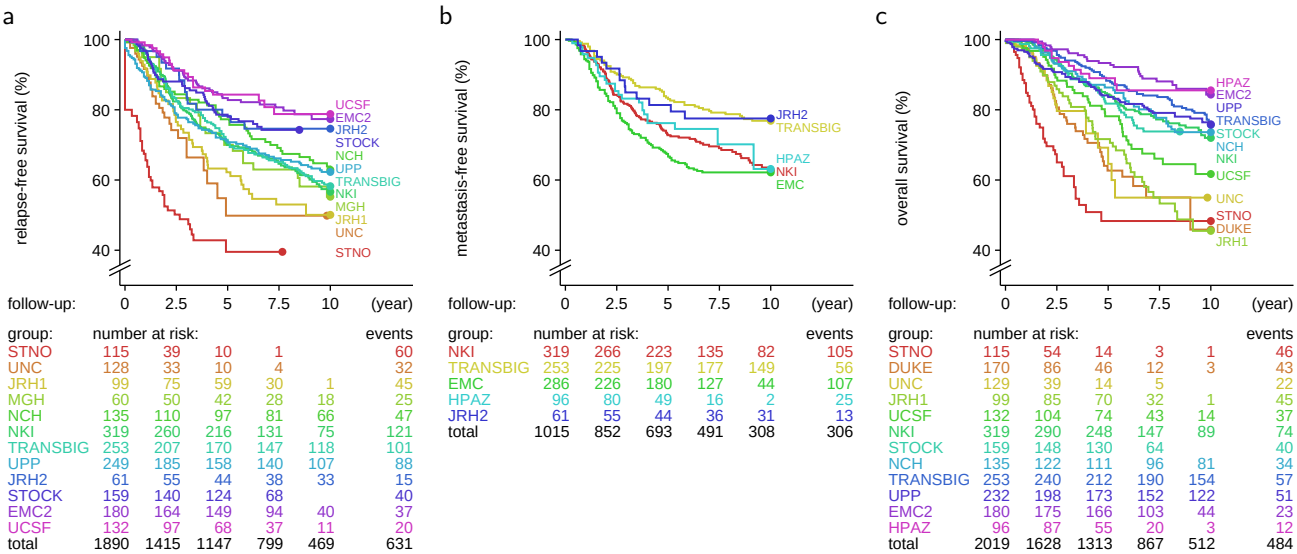
Survival endpoints The original studies often do not give precise description of their endpoints. The following are the definitions we used:

- **Relapse-free survival (RFS)** The endpoints are any events including local (in the same breast or the other side) or regional relapse, and distant metastasis (which is often implied by death). We also consider as RFS endpoints described as disease-free survival (DFS) in the original studies.
- **Metastasis-free survival (MFS)** The endpoints are relapses to other organs, such as liver or brain.
- **Overall survival (OS)** The endpoints are patient's deaths. Few studies specify "disease-specific survival" to distinguish death due to breast cancer, as opposed to any death. When not specified, we assume all deaths are disease related.

Note that these endpoint types are "nested". Death due to breast cancer implies distant metastasis shortly before (and often the time to metastasis is identical to the time to death). RFS typically includes MFS and OS. The events are therefore correlated. Although they obviously have different baseline hazard (RFS drops faster than MFS, which in turn drops faster than OS) (See Supplementary Result 1.2 below), most of our conclusions are concerned with relative survival between groups, which are valid for any type of events (Supplementary Result 1.2, 5.1, 5.2, 5.3, 6, 8). To save space (without affecting the conclusions), Figure 4c,d and Figure 6a,b,c in the main text are based on combined event types. Furthermore, we can not detect genes specifically associated with one type of endpoint but not the others (see Supplementary Result 7).

Time-unit conversion and modifications All time units were converted to "days". One year is considered to be 365.25 days. "Months" and "'weeks" were converted to "years" first. Patients with missing time were excluded, but those with missing event indicator (very few) were considered censored. Median follow-up time can vary greatly between studies (e.g. very short in STNO and very long in TRANSBIG; see the numbers at risk in Supplementary Result 1.2). Therefore all survival data were truncated to 10 years (i.e. patients with longer follow-up are considered to be censored at this time point) in Cox regression analysis to ensure that the meta-analytical results pertain to similar time range in all datasets. The truncation also deals with problems with non-proportional hazard in long follow-up time (see Hilsenbeck *et al.* [1998] Time-dependence of hazard ratios for prognostic factors in primary breast cancer. *Breast Cancer Res. Treat.* **52**:227-237).

1.2 Cohort-specific survival for all endpoints



group:	number at risk:					events
STNO	115	39	10	1		60
UNC	128	33	10	4		32
JRH1	99	75	59	30	1	45
MGH	60	50	42	28	18	25
NCH	135	110	97	81	66	47
NKI	319	260	216	131	75	121
TRANSBIG	253	207	170	147	118	101
UPP	249	185	158	140	107	88
JRH2	61	55	44	38	33	15
STOCK	159	140	124	68		40
EMC2	180	164	149	94	40	37
UCSF	132	97	68	37	11	20
total	1890	1415	1147	799	469	631

group:	number at risk:					events
NKI	319	266	223	135	82	105
TRANSBIG	253	225	197	177	149	56
EMC	286	226	180	127	44	107
HPAZ	96	80	49	16	2	25
JRH2	61	55	44	36	31	13
total	1015	852	693	491	308	306

group:	number at risk:					events
STNO	115	54	14	3	1	46
DUKE	170	86	46	12	3	43
UNC	129	39	14	5		22
JRH1	99	85	70	32	1	45
UCSF	132	104	74	43	14	37
NKI	319	290	248	147	89	74
STOCK	159	148	130	64		40
NCH	135	122	111	96	81	34
TRANSBIG	253	240	212	190	154	57
UPP	232	198	173	152	122	51
EMC2	180	175	166	103	44	23
HPAZ	96	87	55	20	3	12
total	2019	1628	1313	867	512	484

Cohort-specific survival curves for relapse-free survival (a), metastasis-free survival (b) and overall survival (c). Panel c is the same as figure 1b in the main text.

2 Coexpression modules

2.1 Gene Tables

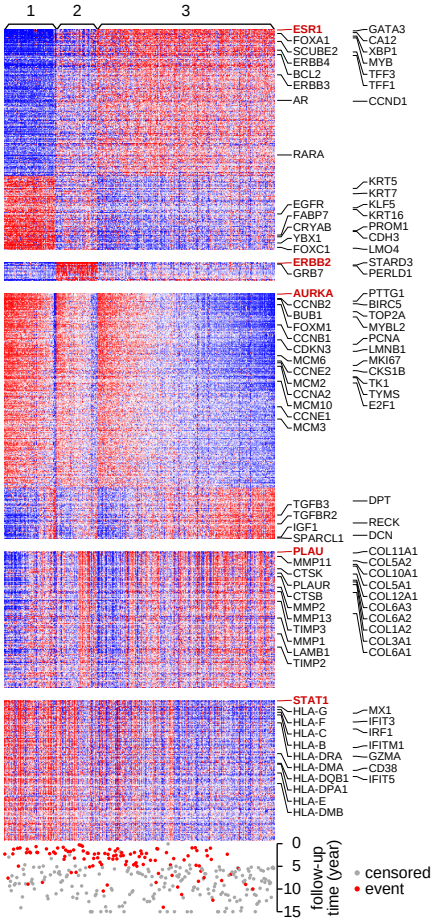
The genes identified by prototype-based coexpression analysis are listed in the file `modules.txt`, in a tab-delimited text file. For convenient browsing, open the file using spreadsheet program (e.g. MS Excel) and format the column widths to fit the contents automatically. We identified 909 genes under the specified criteria. The ordering of the genes in the table is the same as that in Figure 2b of the main text.

2.2 Number of genes selected and mapped to each dataset

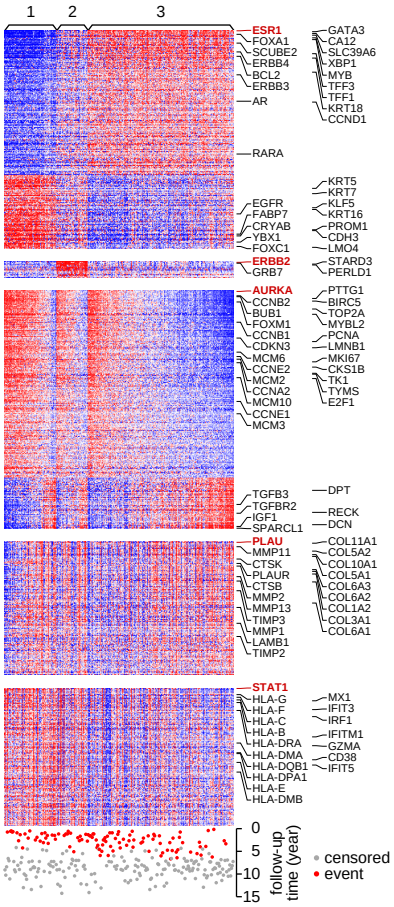
	modules				
datasets	ESR1	ERBB2	AURKA	PLAU	STAT1
NKI	250	20	278	154	159
EMC	248	18	270	151	156
UPP	268	21	294	160	166
STOCK	268	21	294	160	166
DUKE	182	14	204	130	134
UCSF	143	11	176	109	116
UNC	235	19	270	156	161
NCH	235	19	270	156	161
STNO	133	8	156	100	102
JRH1	100	8	135	90	99
JRH2	248	18	270	151	156
MGH	196	11	198	123	108
expO	268	21	294	160	166
TGIF1	248	18	270	151	156
BWH	268	21	294	160	166
TRANSBIG	11	1	50	2	2
EMC2	1	1	10	1	0
HPAZ	2	1	13	1	0
Gene union	268	21	294	160	166

2.3 Heatmaps for all datasets

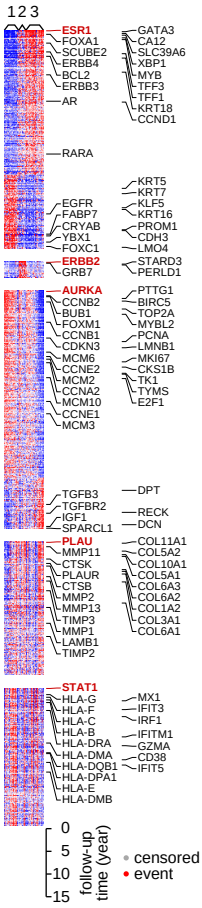
NKI



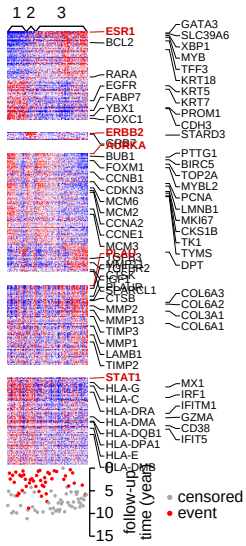
EMC



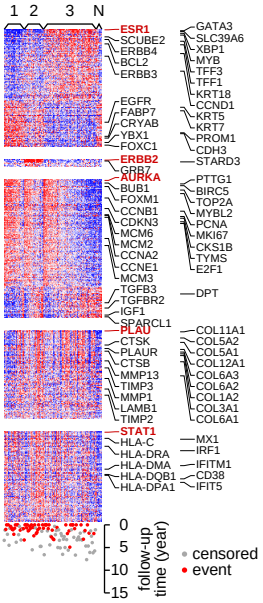
TGIF1



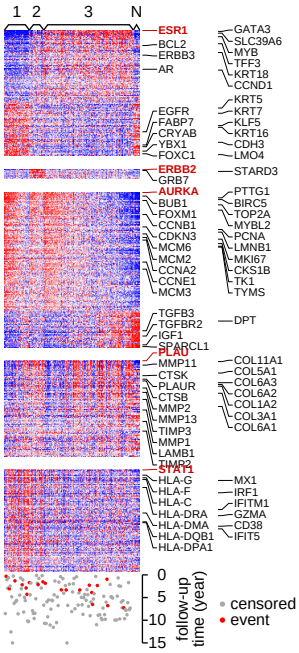
JRH1

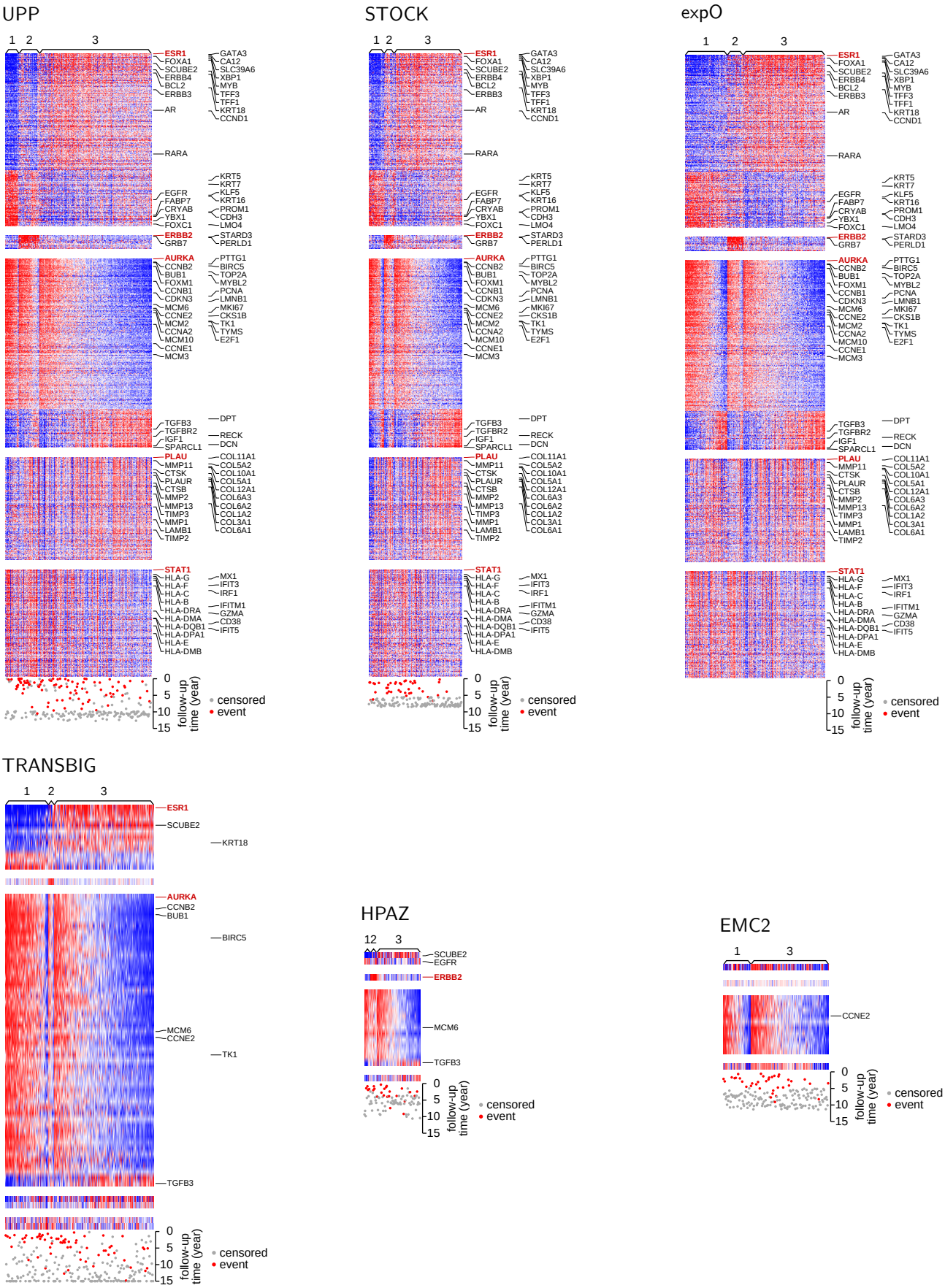


STNO

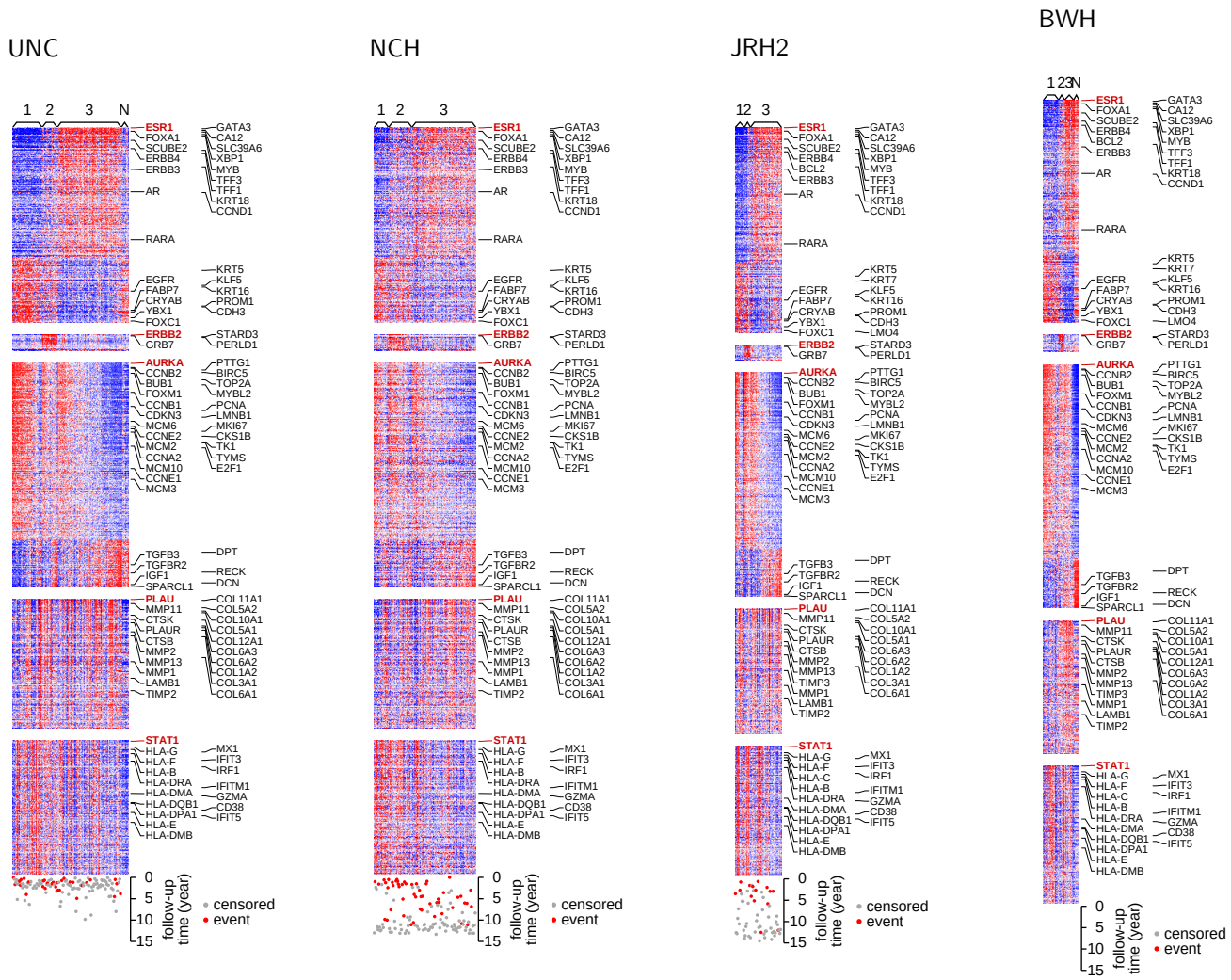


UCSF





Note Because of the small number of genes mapped, the vertical scale of the heatmaps for TRANSBIG, HPAZ and EMC2 is larger than that of other datasets.


UNC
NCH
JRH2
BWH
DUKE
MGH
1 2 3 N
ESR1
FOXA1
SCUBE2
ERBB4
ERBB3
AR
RARA
GATA3
CA12
SLC39A6
XBP1
MYB
TFF3
TFF1
KRT18
CCND1
EGFR
FABP7
CRYAB
YBX1
FOXC1
KRT5
KLF5
KRT16
PROM1
CDH3
ERBB2
GRB7
STARD3
PERLD1
AURKA
CCNB2
BUB1
FOXM1
CCNB1
CDKN3
MCM6
CCNE2
MCM2
CCNA2
MCM10
CCNE1
MCM3
PTTG1
BIRC5
TOP2A
MYBL2
PCNA
LMNB1
MKI67
CKS1B
TK1
TYMS
E2F1
TGFB3
TGFBR2
IGF1
SPARCL1
DPT
RECK
DCN
PLAU
MMP11
CTSK
PLAUR
CTSB
MMP2
MMP13
MMP1
LAMB1
TIMP2
COL11A1
COL5A2
COL10A1
COL5A1
COL12A1
COL6A3
COL6A2
COL1A2
COL3A1
COL6A1
STAT1
HLA-G
HLA-F
HLA-C
HLA-B
HLA-DRA
HLA-DMA
HLA-DQB1
HLA-DPA1
HLA-E
HLA-DMB
MX1
IFIT3
IRF1
IFITM1
GZMA
CD38
IFIT5
follow-up time (year)
censored
event

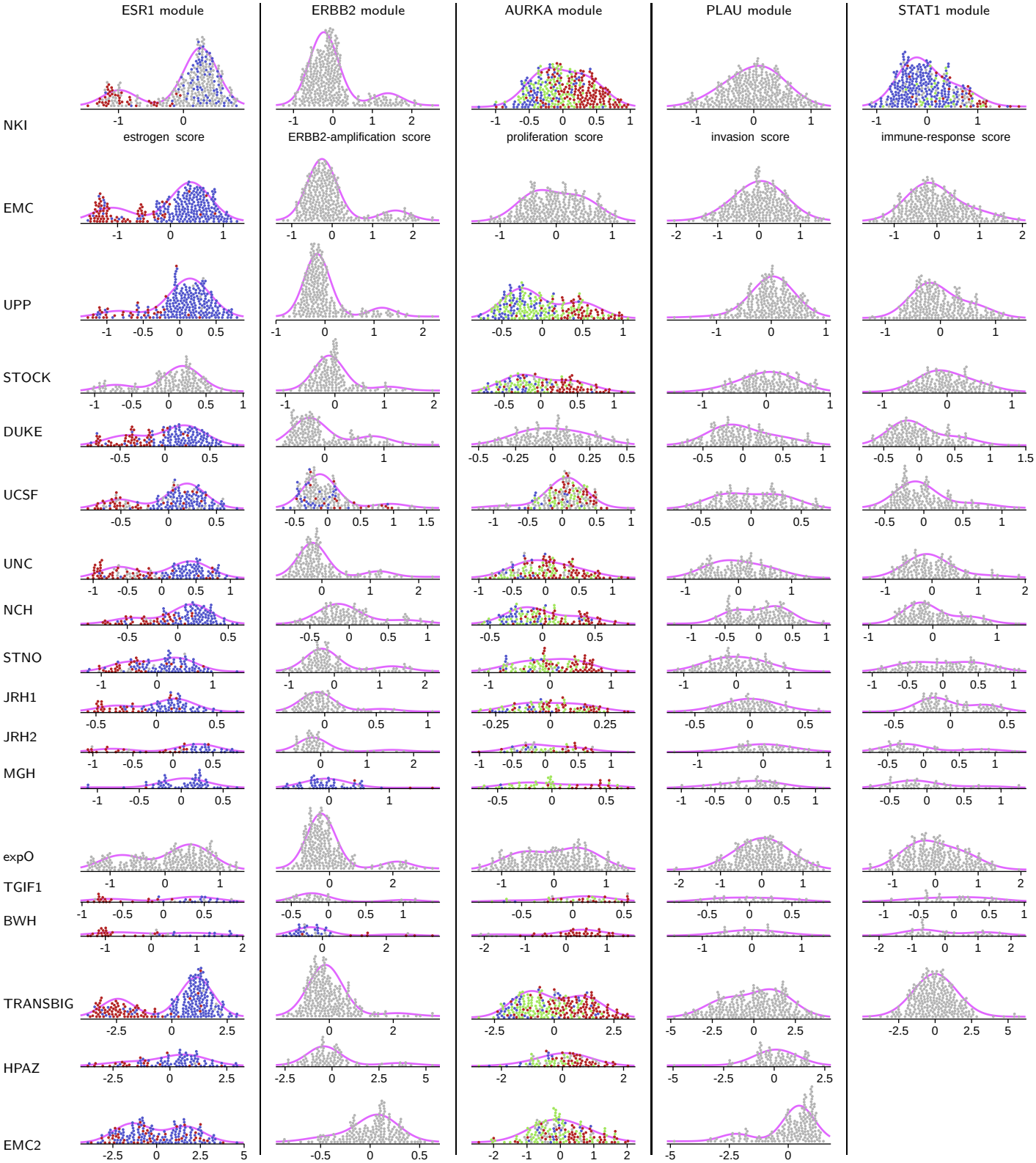
2.4 Biological annotation of the modules

The genes were divided into groups according to the module and the direction of the correlation (i.e., those correlated and anticorrelated with ESR1 were grouped separately; denoted by ESR1+ and ESR1-). Only ESR1 and AURKA modules have negatively correlated genes.

Each of these grouped was compared against the genesets in GO ontology database (based on annotation in Entrez Gene database tables from `ftp://ftp.ncbi.nih.gov/gene/`; version 21 January 2007) and Molecular Signature Database (MSigDB) version 2 from `http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html`, Fisher's exact test was used to score the association between the modules and a gene set. The top-ranking genesets were shown in the file `go.txt` and `msigdb.txt`. The tables are in tab-delimited text format that can be opened from spreadsheet programs. The files contain Fisher's exact test p-values, odd-ratio of enrichment, geneset identifiers, and brief description.

3 Module score distributions

3.1 Dot histograms



Note In datasets TRANSBIG, EMC2 and HPAZ (produced on custom diagnostic platforms), few genes were found for modules other than proliferation (AURKA), and therefore the module score might be unreliable. None can be found for STAT1-module for EMC2 and HPAZ.

3.2 Tests of bimodality

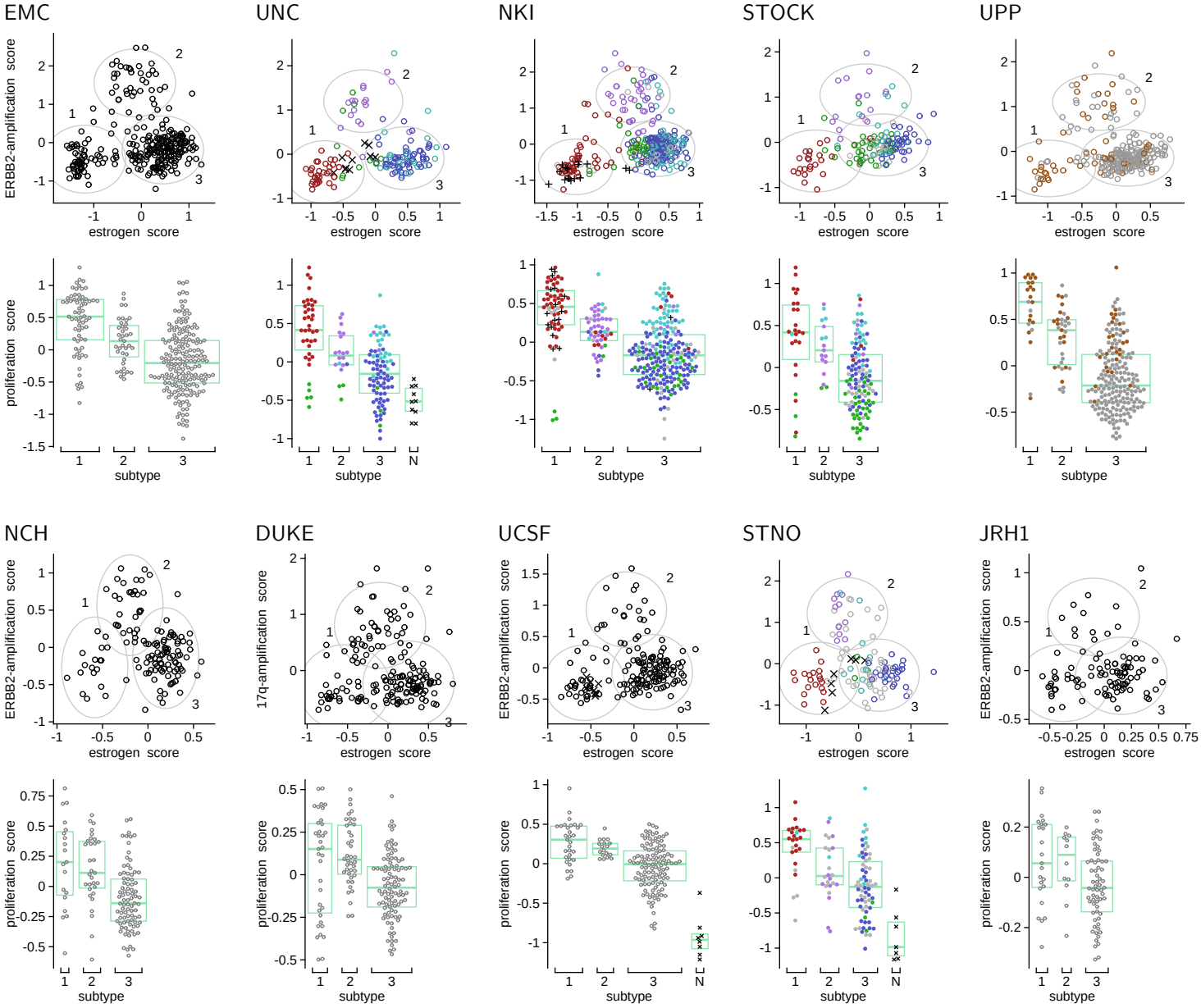
Log likelihood ratio between two- and one-component Gaussian mixture models were used as the test statistics, (as suggested in McLachlan and Peel [2000] *Finite Mixture Modules, Wiley, New York*). However, the parametric bootstrap null distribution was generated from uniform distribution instead of unimodal normal distribution (following Hartigan [1985] "The dip test of unimodality." *Ann. Statist.* **13**:30–84). This was done because the p-values based on unimodal-normal null distribution were too sensitive to non-normal, but still unimodal, densities.

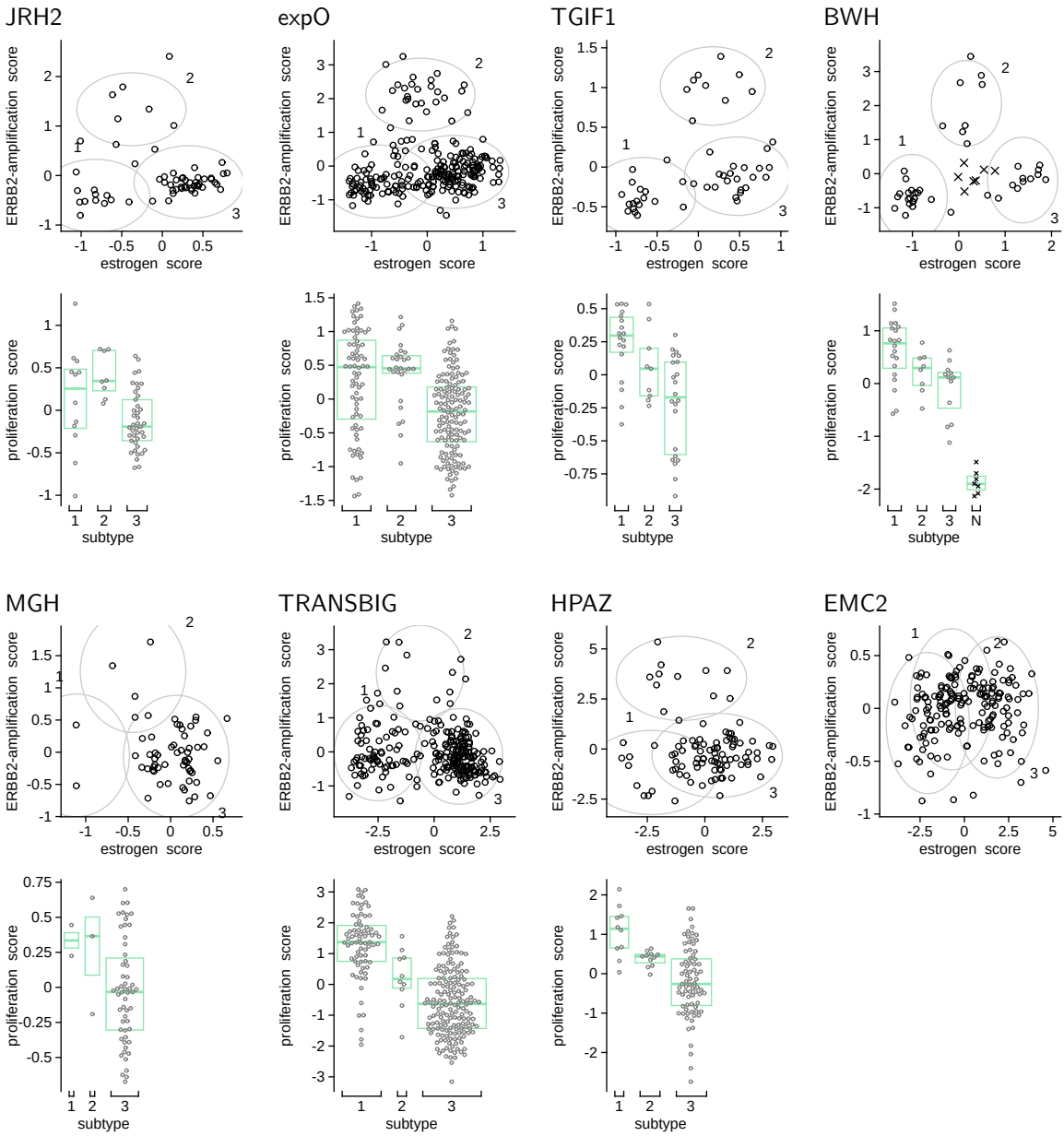
	p-values of bimodality test of module scores				
dataset	ESR1	ERBB2	AURKA	PLAU	STAT1
NKI	0.001 *	0.001 *	0.560	0.990	0.068
EMC	0.001 *	0.001 *	0.438	0.996	0.064
UPP	0.002 *	0.001 *	0.047	0.061	0.061
STOCK	0.001 *	0.001 *	0.083	0.083	0.083
DUKE	0.069	0.001 *	0.630	0.082	0.078
UCSF	0.003 *	0.001 *	0.055	0.473	0.054
UNC	0.005 *	0.001 *	0.146	0.087	0.085
NCH	0.040	0.047	0.097	0.295	0.083
STNO	0.689	0.001 *	0.412	0.573	0.571
JRH1	0.113	0.003 *	0.724	0.804	0.731
JRH2	0.003 *	0.001 *	0.119	0.531	0.010 *
MGH	0.064	0.076	0.929	0.965	0.070
expO	0.021	0.001 *	0.673	0.574	0.481
TGIF1	0.877	0.001 *	0.088	0.875	0.919
BWH	0.927	0.001 *	0.020	0.926	0.889
TRANSBIG	0.001 *	0.001 *	0.715	0.108	0.072
EMC2	0.692	0.081	0.753	0.001 *	nan
HPAZ	0.108	0.001 *	0.109	0.097	nan

* p <= 0.01

4 Three subtypes

4.1 Scatter plots of ERBB2 vs ESR1 module scores





Notes

1. Dataset BWH contains entirely of high-grade tumors according to pathological data. It does not have patient outcome data and therefore the difficulty in deciding high- or low-proliferation does not affect survival analysis.

2. Dataset MGH contains only one ER-negative, 3 ERBB2+ and 3 low-grade tumors (out of 60) according to pathological data.

3. Dataset EMC2 does not have enough genes to reliably determine the ER and ERBB2 module scores. The clusters can not be fitted well. In subsequent analysis, the pathological ER-status are used to assign the subtype (ER+ is considered type 3, ER- is type 1).

4.2 Tests of trimodality

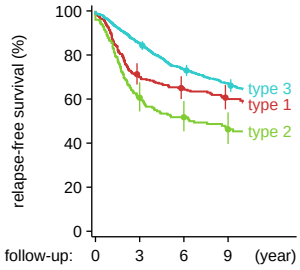
Log likelihood ratio between k- and $(k-1)$-component Gaussian mixture models were used as the test statistics, (as suggested in McLachlan and Peel [2000] *Finite Mixture Modules, Wiley, New York*). Parametric bootstrap were used to general the null distribution, using the the parameters from the fitted null model.

	p-values of tests for the number of components	
dataset	3 vs 2	4 vs 3
NKI	0.001 *	0.978
EMC	0.001 *	0.024
UPP	0.001 *	0.325
STOCK	0.001 *	0.014
DUKE	0.002 *	0.007 *
UCSF	0.001 *	0.012
UNC	0.001 *	0.054
NCH	0.001 *	0.589
STNO	0.001 *	0.026
JRH1	0.002 *	0.030
JRH2	0.002 *	0.088
MGH	0.932	0.671
expO	0.001 *	0.002 *
TGIF1	0.001 *	0.268
BWH	0.001 *	0.025
TRANSBIG	0.001 *	0.035
EMC2	0.091	0.007 *
HPAZ	0.001 *	0.905

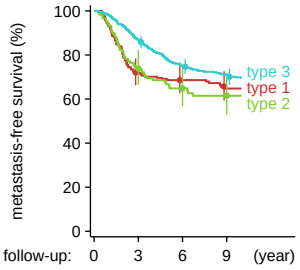
* p <= 0.01

5 Additional survival analyses

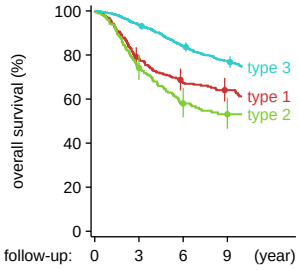
5.1 Three subtypes



group:	number at risk:				events
type 1	361	219	172	88	129
type 2	224	119	87	57	110
type 3	1305	1027	751	439	392
total	1890	1365	1010	584	631

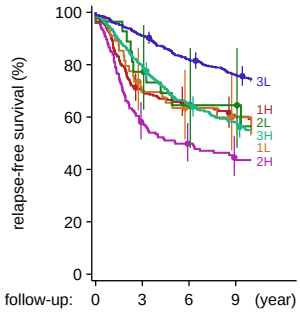


group:	number at risk:				events
type 1	211	151	125	70	71
type 2	119	83	60	34	44
type 3	685	587	435	271	191
total	1015	821	620	375	306

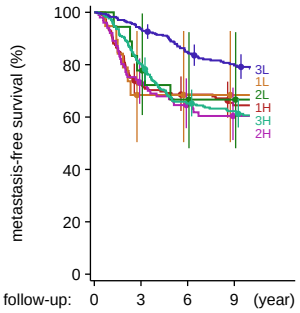


group:	number at risk:				events
type 1	389	260	189	101	122
type 2	263	160	99	59	100
type 3	1367	1149	834	484	262
total	2019	1569	1122	644	484

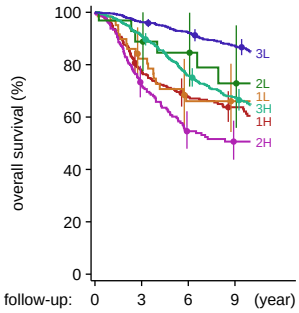
5.2 Three subtypes combined with high/low proliferation



1H	305	184	147	76	108
1L	56	35	25	12	21
2H	196	99	75	47	100
2L	28	20	12	10	10
3H	654	477	335	202	253
3L	651	550	416	237	139
total	1890	1365	1010	584	631



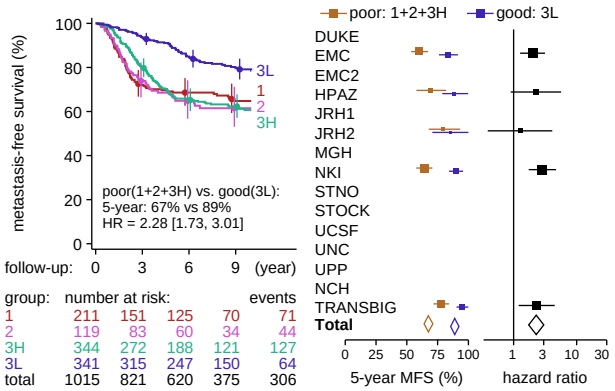
1H	192	138	114	65	65
1L	19	13	11	5	6
2H	101	69	49	26	38
2L	18	14	11	8	6
3H	344	272	188	121	127
3L	341	315	247	150	64
total	1015	821	620	375	306



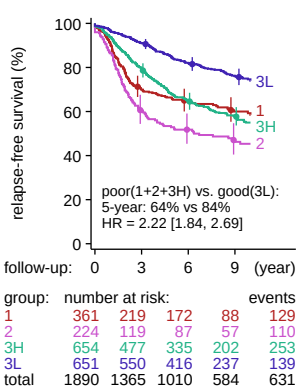
1H	324	216	162	86	104
1L	65	44	27	15	18
2H	231	138	83	47	94
2L	32	22	16	12	6
3H	686	559	385	229	189
3L	681	590	449	255	73
total	2019	1569	1122	644	484

5.3 Separate metastasis-free survival and relapse-free survival

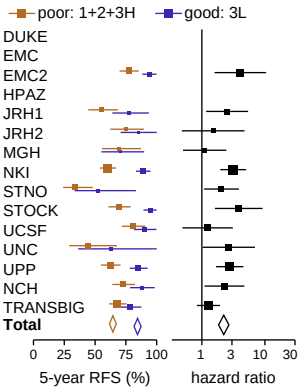
Figure 4c,d use metastasis-free survival, if available, or relapse-free survival otherwise. Below are the results for the two types of endpoints separately.



group:	number at risk:				events
1	211	151	125	70	71
2	119	83	60	34	44
3H	344	272	188	121	127
3L	341	315	247	150	64
total	1015	821	620	375	306

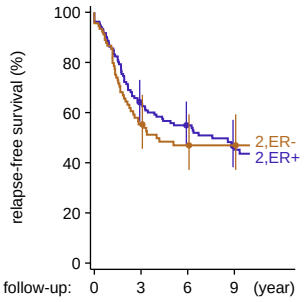


group:	number at risk:				events
1	361	219	172	88	129
2	224	119	87	57	110
3H	654	477	335	202	253
3L	651	550	416	237	139
total	1890	1365	1010	584	631

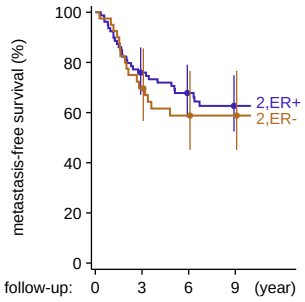


6 Prognostic value of ER-status within type-2 (ERBB2+) tumors

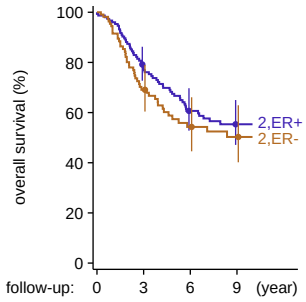
Because pathological ER status is not available for all patients, here ER-status is based on estrogen scores (using as cutoff the midpoint between the cluster center of type 1 and type 3).



2,ER+	133	78	56	30	66
2,ER-	91	41	31	27	44
total	224	119	87	57	110



2,ER+	79	57	41	23	28
2,ER-	40	26	19	11	16
total	119	83	60	34	44



2,ER+	152	101	66	36	57
2,ER-	111	59	33	23	43
total	263	160	99	59	100

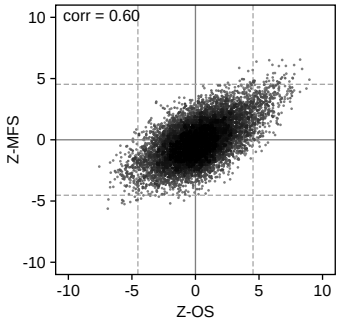
7 Gene-by-gene survival analysis

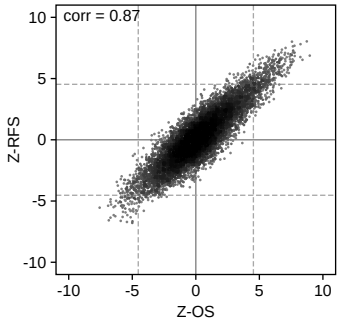
7.1 Gene Tables

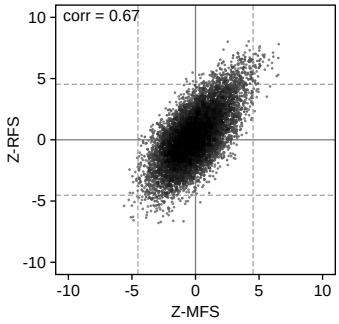
The top-ranking genes identified by meta-analytical gene-by-gene Cox regression are are listed in the file `zsurvival.txt`, in a tab-delimited text file. For convenient browsing, open in spreadsheet program (e.g. MS Excel) and format the column widths to fit the contents automatically.

7.2 Agreement of gene associations with different survival endpoints

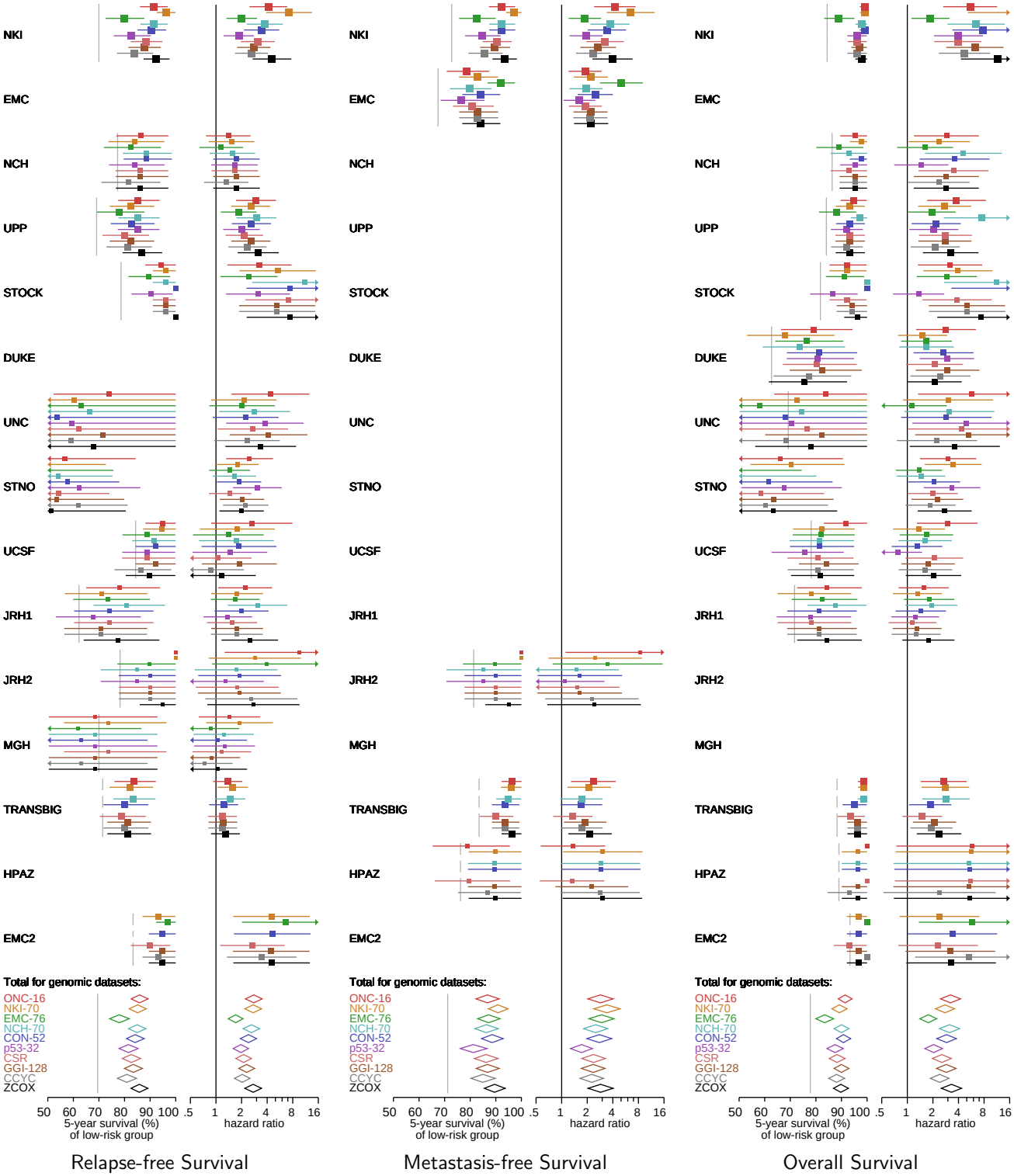
Here we show that the choice of endpoints (OS: overall survival, MFS: metastasis-free survival, RFS: relapse-free survival) do not substantially affect gene-by-gene survival analysis. The lower Z-scores of MFS is due mainly to smaller sample size (MFS data are available only in 5 datasets, see Figure 1a of the main text).







8 Forest plots of signature performance for all survival endpoints



9 Concordance in risk classifications

9.1 Tables of pairwise concordance

Whole signatures:

	NKI-70	ONC-16	EMC-76	NCH-70	CON-52	p53-32	GGI-128	CSR	CCYC	ZCOX
NKI-70	*	79.97	69.37	77.54	81.02	76.07	82.18	75.12	81.34	83.66
ONC-16	79.97	*	70.22	82.50	83.87	74.06	86.51	79.02	84.08	86.72
EMC-76	69.37	70.22	*	70.11	73.59	61.04	73.38	72.11	73.27	73.48
NCH-70	77.54	82.50	70.11	*	85.45	67.84	86.29	82.39	83.13	85.66
CON-52	81.02	83.87	73.59	85.45	*	69.11	91.04	82.29	88.82	91.14
p53-32	76.07	74.06	61.04	67.84	69.11	*	71.64	67.21	69.74	72.69
GGI-128	82.18	86.51	73.38	86.29	91.04	71.64	*	83.34	92.62	93.46
CSR	75.12	79.02	72.11	82.39	82.29	67.21	83.34	*	82.81	83.87
CCYC	81.34	84.08	73.27	83.13	88.82	69.74	92.62	82.81	*	91.25
ZCOX	83.66	86.72	73.48	85.66	91.14	72.69	93.46	83.87	91.25	*

Average pairwise concordance: 79.4% (84.5% if EMC-76 and p53-32 are excluded)

Proliferation-gene partial signatures:

	ONC-16	NKI-70	EMC-76	NCH-70	CON-52	p53-32	CSR	GGI-128	CCYC	ZCOX
ONC-16	*	84.71	82.39	85.35	89.46	84.92	84.40	90.72	89.77	90.20
NKI-70	84.71	*	83.24	83.45	88.09	80.50	83.34	88.51	88.72	89.67
EMC-76	82.39	83.24	*	84.61	86.93	78.39	83.13	86.29	86.72	85.56
NCH-70	85.35	83.45	84.61	*	87.45	81.13	87.66	87.56	86.51	87.24
CON-52	89.46	88.09	86.93	87.45	*	83.66	86.72	93.15	91.78	92.30
p53-32	84.92	80.50	78.39	81.13	83.66	*	81.34	85.98	84.29	84.40
CSR	84.40	83.34	83.13	87.66	86.72	81.34	*	86.61	86.19	87.45
GGI-128	90.72	88.51	86.29	87.56	93.15	85.98	86.61	*	94.83	94.52
CCYC	89.77	88.72	86.72	86.51	91.78	84.29	86.19	94.83	*	93.67
ZCOX	90.20	89.67	85.56	87.24	92.30	84.40	87.45	94.52	93.67	*

Average pairwise concordance: 86.7%

Non-proliferation-gene partial signatures:

	NKI-70	ONC-16	EMC-76	NCH-70	CON-52	p53-32	GGI-128	CSR	CCYC	ZCOX
NKI-70	*	67.00	59.30	65.21	62.78	69.43	69.30	65.42	62.05	73.22
ONC-16	67.00	*	58.36	63.73	60.15	68.90	65.29	62.68	59.20	70.48
EMC-76	59.30	58.36	*	59.52	60.46	54.98	61.85	64.47	56.88	57.62
NCH-70	65.21	63.73	59.52	*	66.16	58.78	64.18	65.74	56.04	65.00
CON-52	62.78	60.15	60.46	66.16	*	58.99	61.40	60.46	58.88	62.99
p53-32	69.43	68.90	54.98	58.78	58.99	*	65.07	61.52	58.67	69.74
GGI-128	69.30	65.29	61.85	64.18	61.40	65.07	*	65.74	57.40	69.97
CSR	65.42	62.68	64.47	65.74	60.46	61.52	65.74	*	62.05	67.53
CCYC	62.05	59.20	56.88	56.04	58.88	58.67	57.40	62.05	*	63.42
ZCOX	73.22	70.48	57.62	65.00	62.99	69.74	69.97	67.53	63.42	*

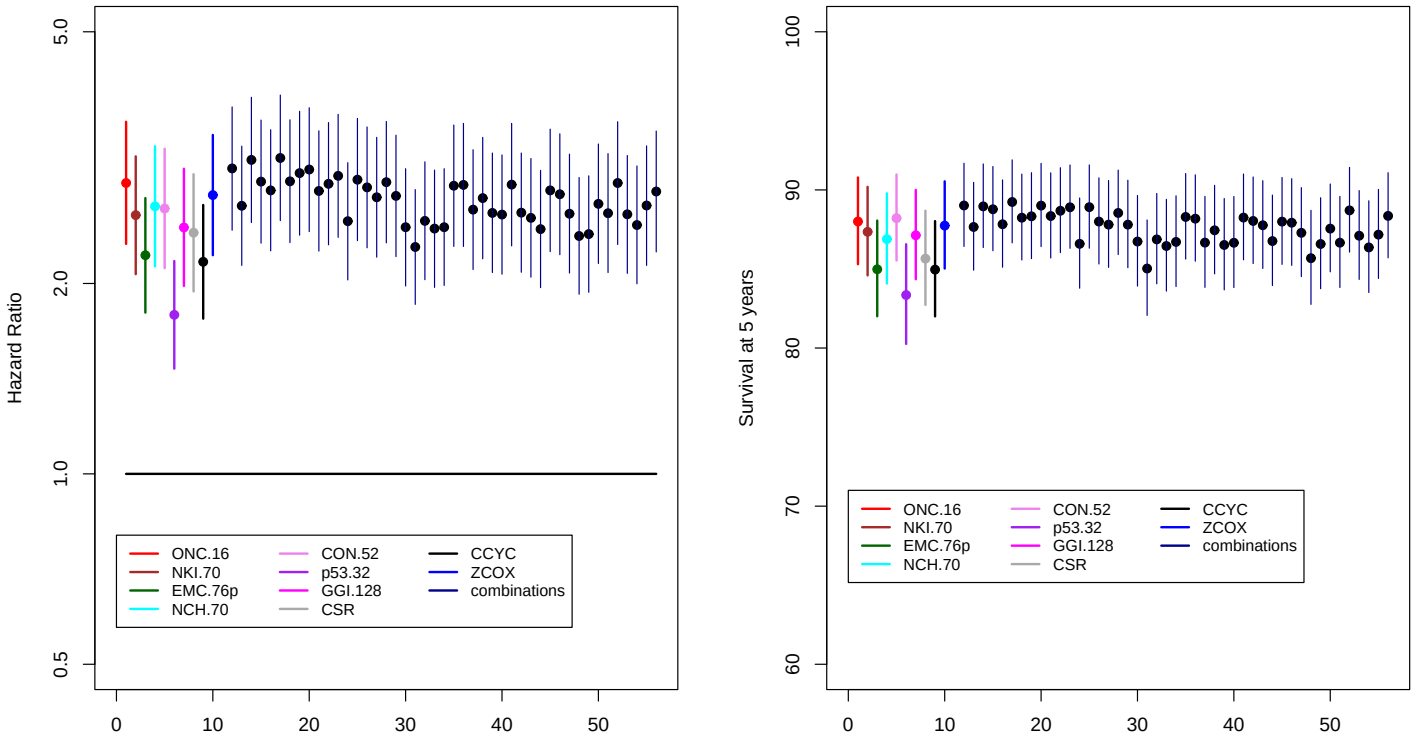
Average pairwise concordance: 63.1%

9.2 Combined prediction by pairs of signatures

Every distinct pair of signatures were combined in the following way:

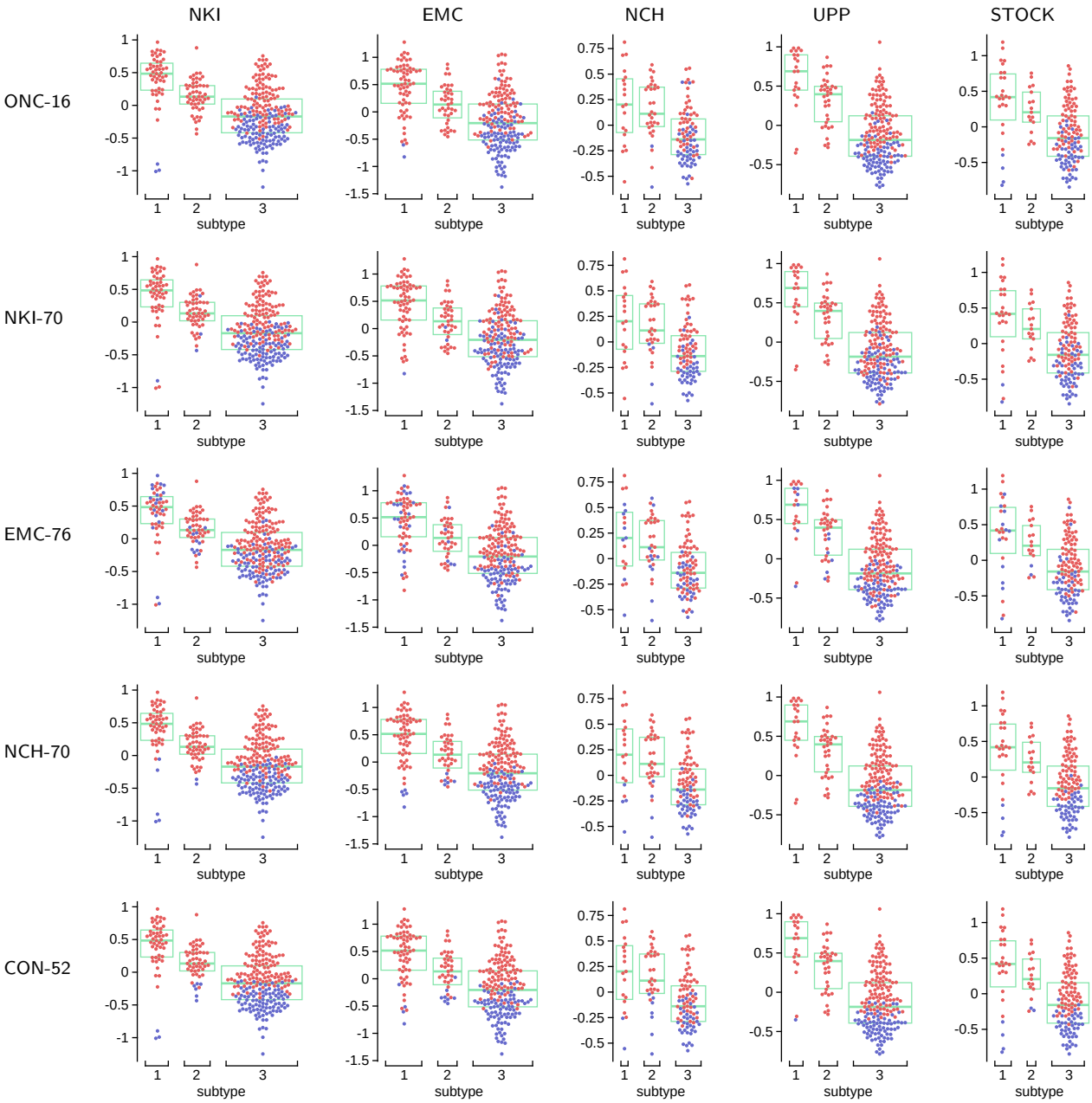
1. Their continuous prediction scores were used as explanatory variables in Cox regression (stratified by dataset).
2. The fitted coefficients were used as weights in linear combination of the two prediction scores to produce the combined prediction score
3. The combined scores were used to rank patients. 33% in each dataset were classified as low-risk.

The results for the 45 pairs are shown below as the black bars to the right of the individual signatures. The identity of each combination is not shown (the ordering from left to right is [ONC-16 + NKI-70], [ONC-16 + EMC-76p], and so on), because we only want to demonstrate that their hazard ratios and survivals are similar and within the confidence interval of the some individual signatures. Note that the combined performance may be slightly biased upward because the weights are estimated from the same data. Even so, the improvement is not clinically substantial.

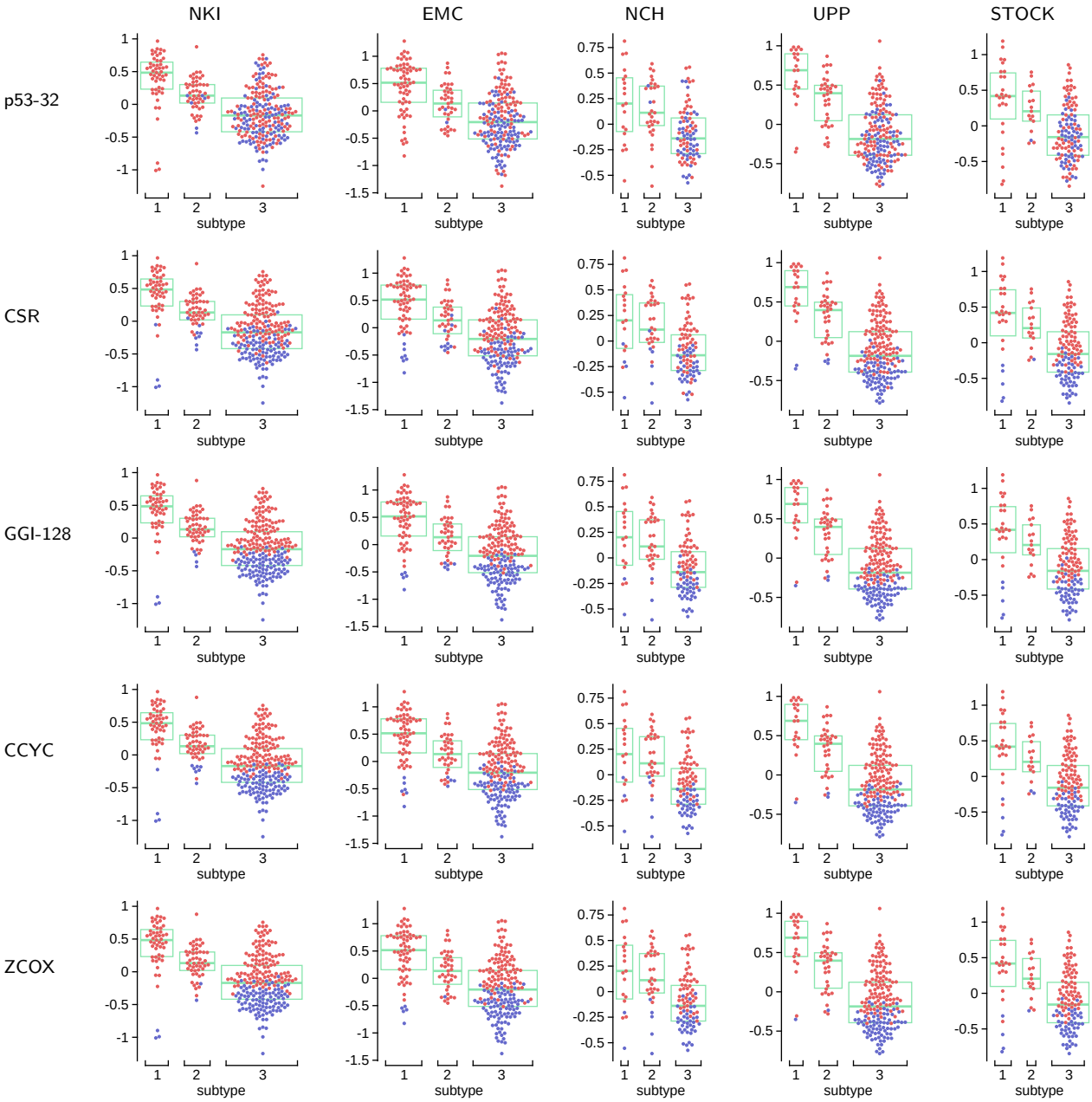


9.3 Patient classifications on proliferation-vs-subtype plots

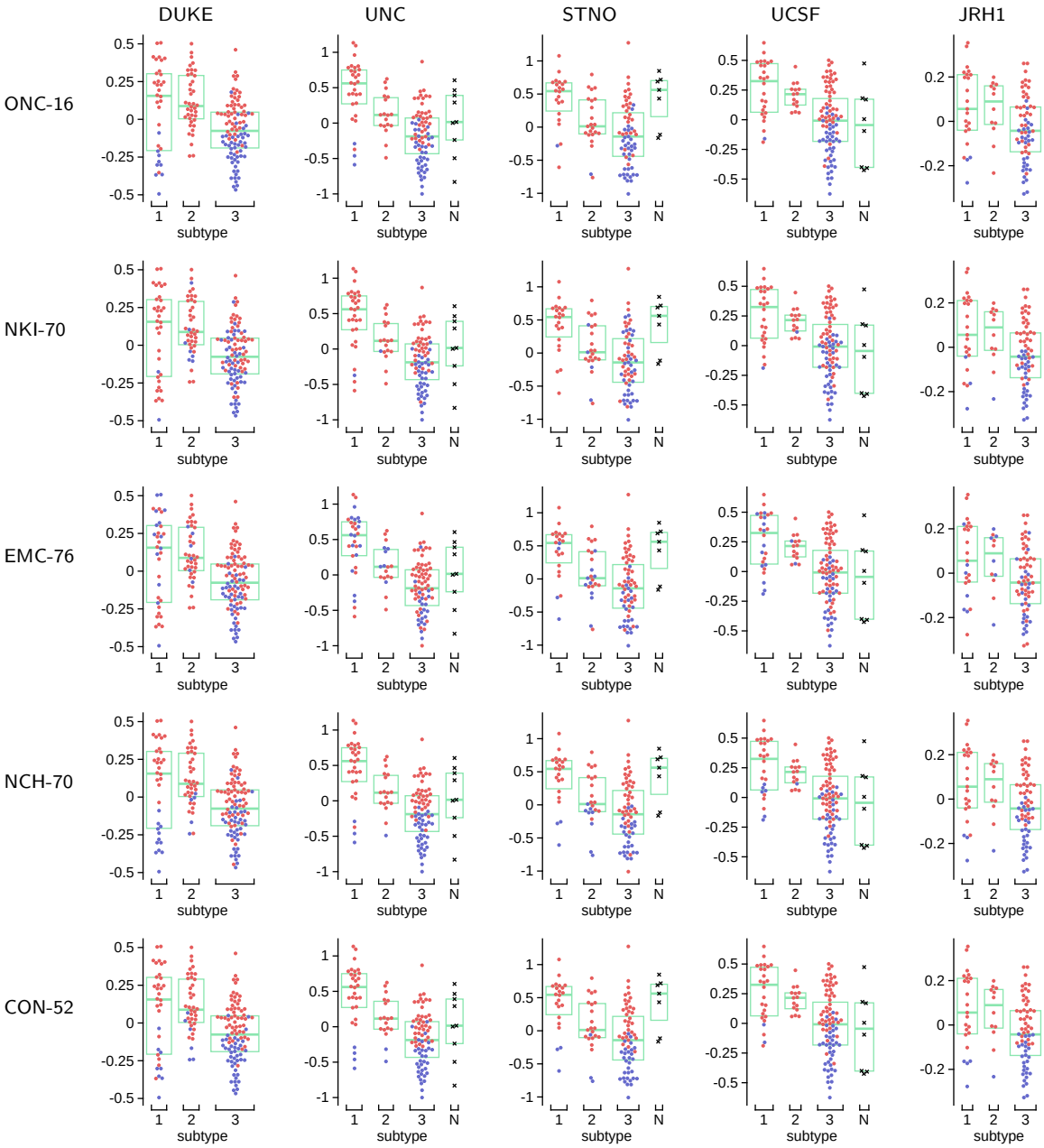
All vertical axes correspond to proliferation score.



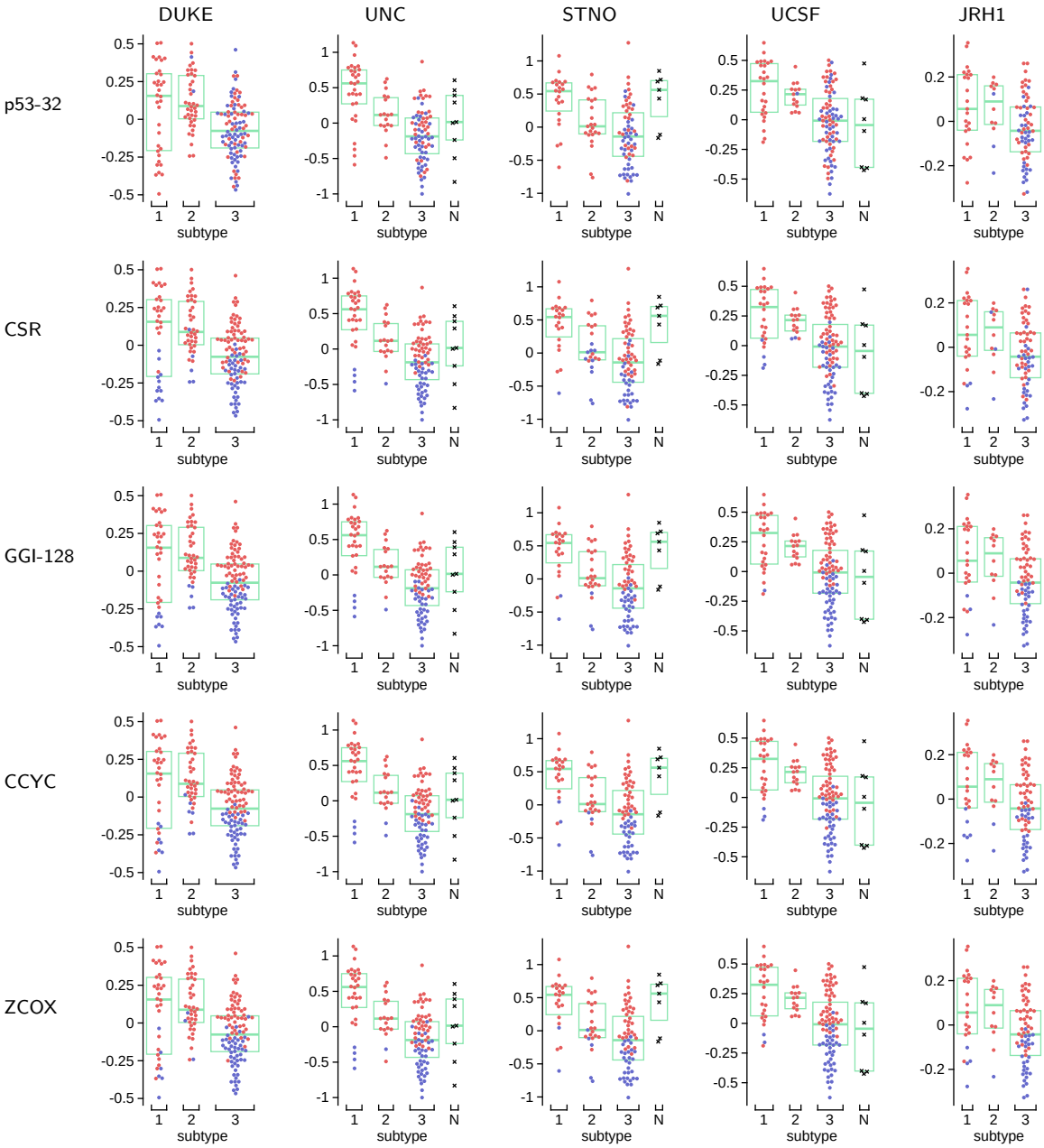
All vertical axes correspond to proliferation score.



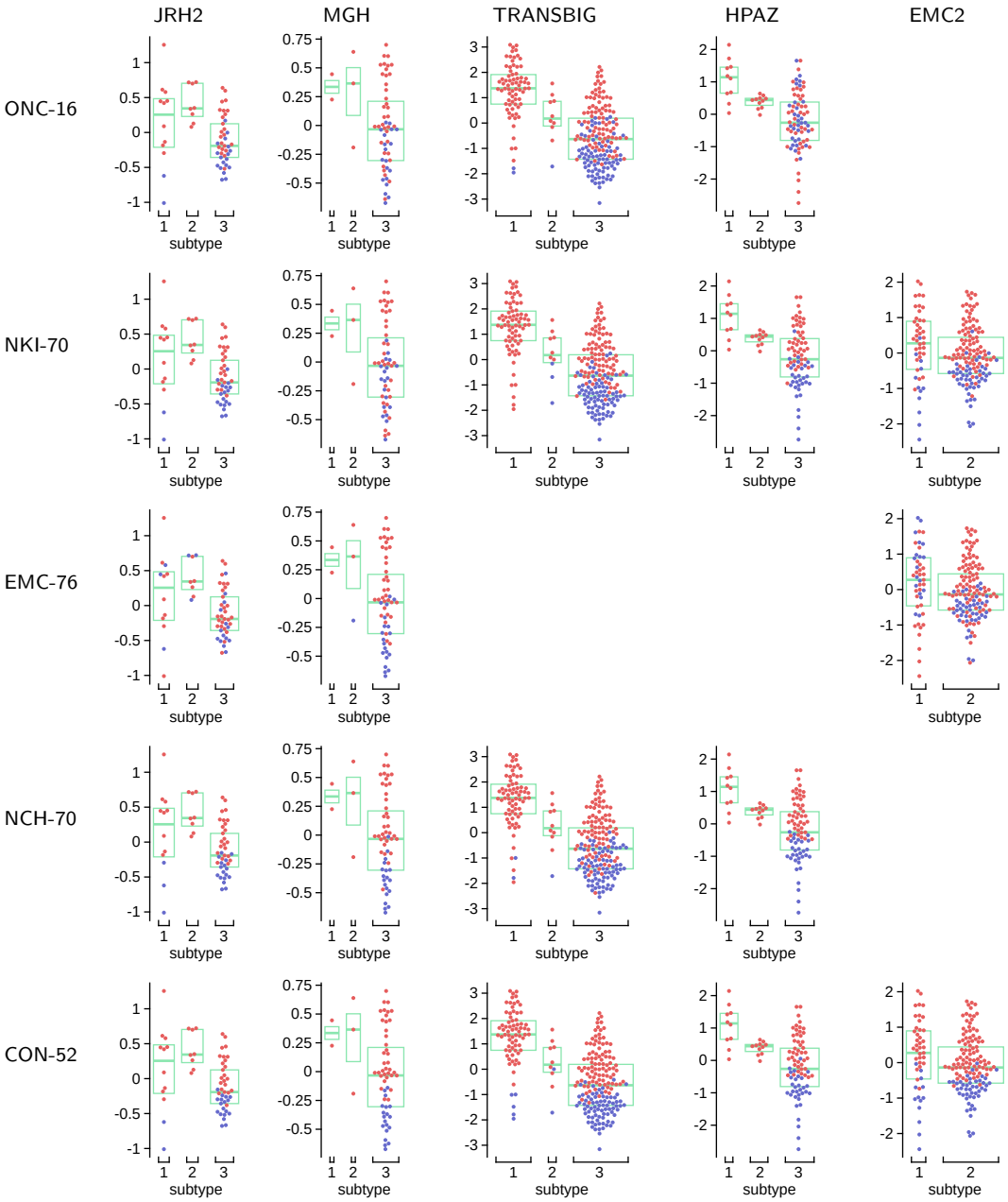
All vertical axes correspond to proliferation score.



All vertical axes correspond to proliferation score.



All vertical axes correspond to proliferation score.



All vertical axes correspond to proliferation score.