Finding Large Average Submatrices in High Dimensional Data
Supplementary Materials

Shabalin A., Weigman V.J., Perou C.M., Nobel A.B.

July 8, 2008

1 Bar Plots for the Hu data

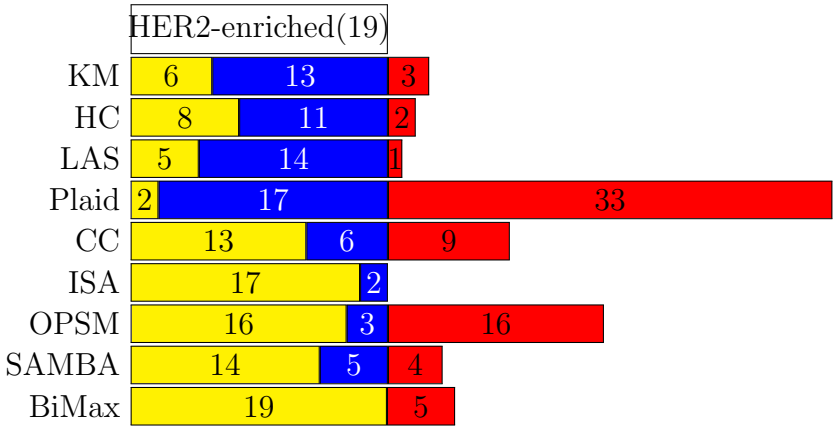


Figure 1: Bar-plot of missed, true and false discoveries for different biclustering methods and the HER2 subtype.

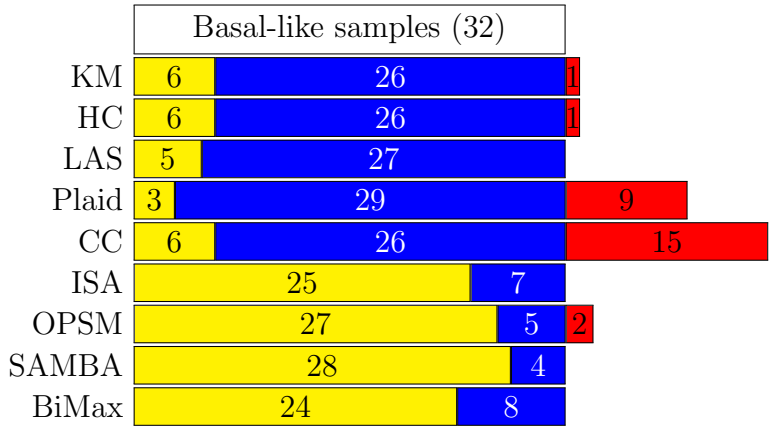


Figure 2: Bar-plot of missed, true and false discoveries for different biclustering methods and the Basal subtype.

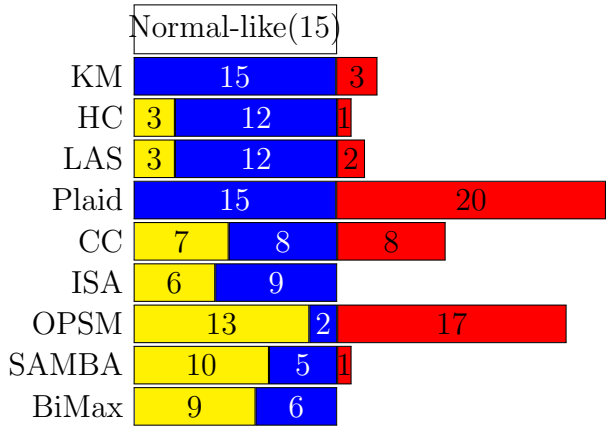


Figure 3: Bar-plot of missed, true and false discoveries for different biclustering methods and the Normal subtype

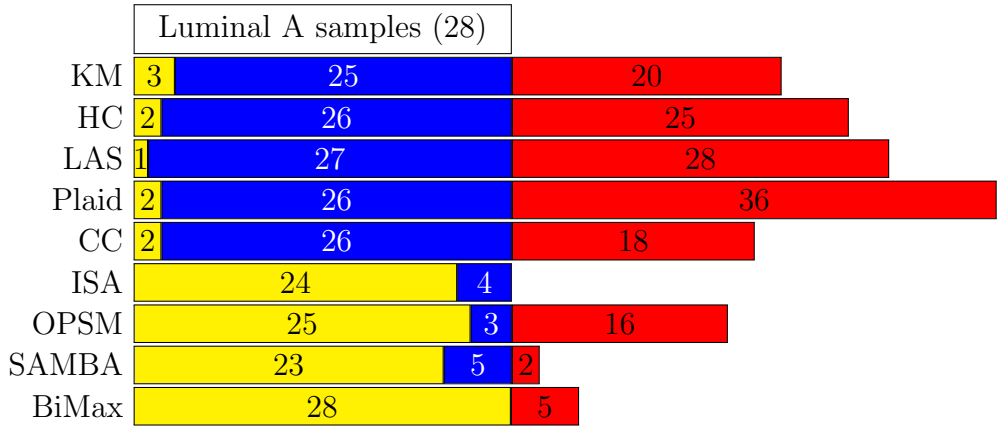


Figure 4: Bar-plot of missed, true and false discoveries for different biclustering methods and the Luminal A subtype.

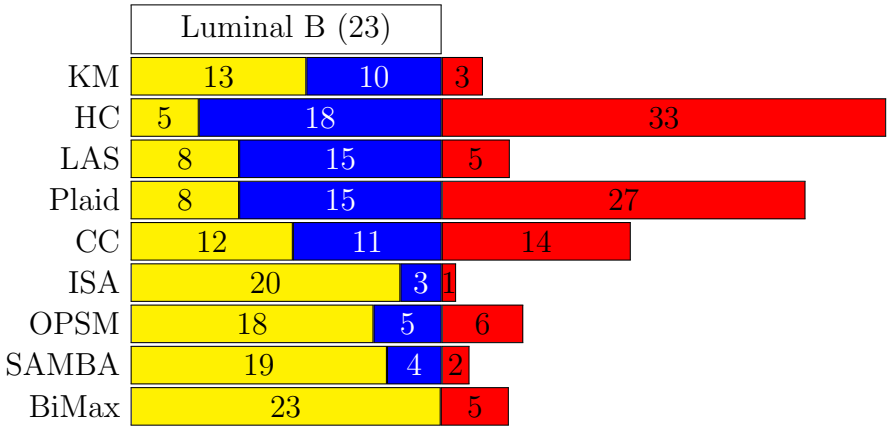
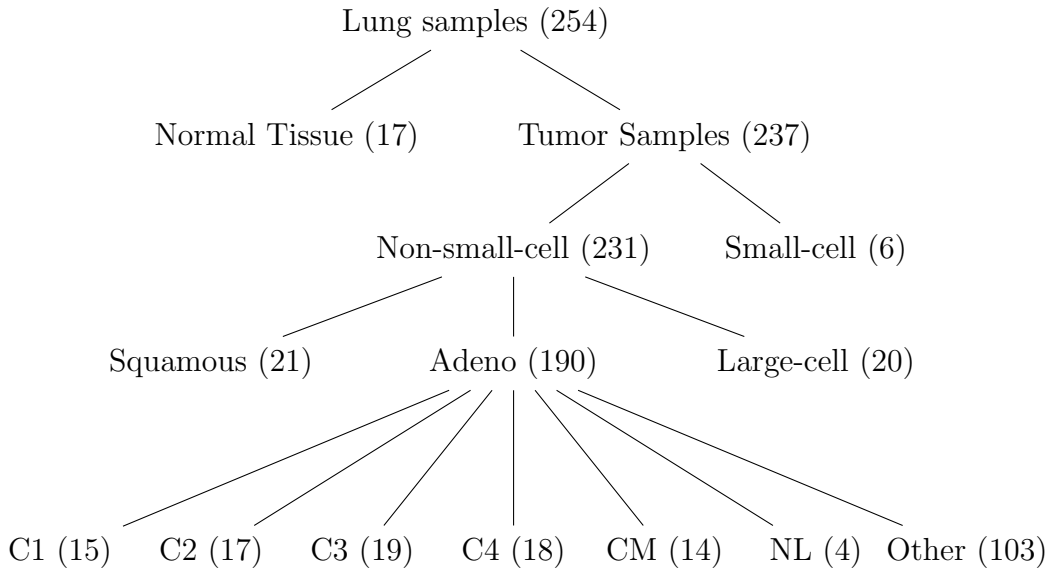


Figure 5: Bar-plot of missed, true and false discoveries for different biclustering methods and the Luminal B subtype

2 Validation Results for Lung cancer data

For the analysis of lung expression data, we used samples from the data set of Bhattacharjee *et al* (2001) which were produced using the Affymetrix human U95A oligonucleotide probe arrays. The Bhattacharjee data contains 237 lung tumor samples and 17 normal samples, measured on 12,635 genes. The tumor samples were first split based on histopathology into groups of small-cell lung carcinomas (SCLC) and non-small-cell lung carcinomas (NSCLC). The NSCLC samples were subcategorized as adenocarcinomas (ADENO), squamous cell carcinomas (SQ), and large-cell carcinomas (LC), of which adenocarcinomas are the most common. Bhattacharjee et al used hierarchical clustering to further divide the ADENO samples into six subclasses: C1-C4, colon metastasis (CM), and normal lung (NL).

The tree below illustrates the established classification of the samples.



Because of extreme computational complexity and large output we excluded BiMax from the comparison.

2.1 Validation Results for Lung cancer data (all samples)

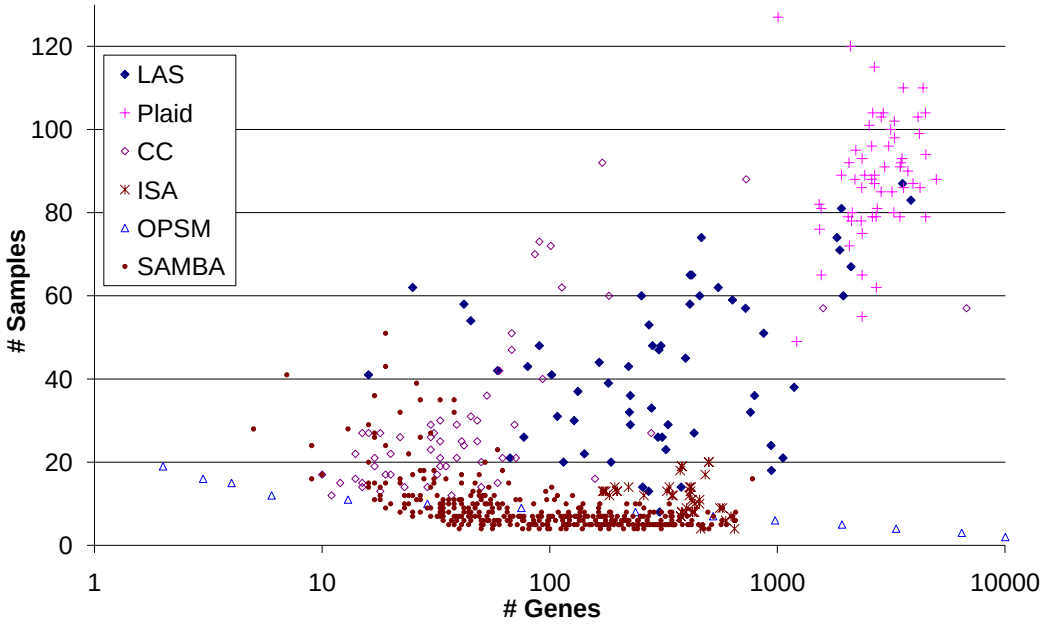


Figure 6: Bicluster sizes for different methods.

Method	# of Clusters	Eff. # of Clusters	Ratio
LAS	60	42.7	0.711
Plaid	60	11.1	0.184
CC	60	59.9	0.999
ISA	43	19.7	0.457
OPSM	14	8.5	0.608
SAMBA	389	226.1	0.581
KM	12,168	77.4	0.006
HC	206,225	985.0	0.005

Table 1: Output summary for different biclustering methods. From left to right: total number of biclusters produces; effective number of biclusters; the ratio of the effective number over the total number of biclusters.

	Correlation		Std. Dev.
	Gene	Sample	Gene
Matrix	0.01	0.03	0.83
KM	0.21	0.27	0.88
HC	0.44	0.40	0.89
LAS	0.25	0.20	1.53
Plaid	0.06	0.08	1.10
CC	0.13	0.12	1.60
ISA	0.24	0.40	1.09
OPSM	0.39	0.13	0.60
SAMBA	0.23	0.04	1.71
Subtypes		0.30	

Table 2: Average standard deviation of genes, and average pairwise correlation of genes and samples, for biclusters, IRCC clusters, and the whole data matrix.

	Adeno	Large-cell	Normal	Small-cell	Squamous	Adeno, C1	Adeno, C2	Adeno, C3	Adeno, C4	Adeno, CM	Adeno, NL
KM	11.7	24.9	21.1	6.6	21.1	14.3	10.6	14.9	13.4	15.5	17.1
HC	6.5	28.1	21.8	8.8	22.8	12.7	7.5	11.5	12.7	15.5	14.0
LAS	7.0	26.8	23.6	4.5	17.0	8.0	13.6	11.7	12.7	15.5	14.0
Plaid	8.1	14.7	9.6	2.8	8.3	8.4	4.8	3.7	7.5	5.2	11.5
CC	11.2	22.2	11.9	2.6	8.3	7.4	4.5	2.0	2.2	2.4	10.6
ISA	0.6	28.1	21.2	8.4	22.8	2.2	12.7	11.5	9.6	15.5	14.0
OPSM	0.3	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
SAMBA	0.9	3.1	1.4	1.1	0.1	1.6	0.6	0.4	0.5	0.8	1.1

Table 3: The minus $\log_{10}$ p-values of best subtype capture for different biclustering and sample clustering methods.

	# of BC's	Survival 5% level	KEGG/Cytoband Gene	KEGG/Cytoband Sample	2 out of 3	All 3
LAS	60	9	25	38	24	0
Plaid	60	7	19	28	17	0
CC	60	5	1	24	3	0
ISA	43	2	15	24	14	0
OPSM	14	0	0	1	0	0
SAMBA	389	6	50	43	9	1

Table 4: The number of biclusters passing tests for survival, and gene-set enrichment and sample-set differential expression of KEGG categories and cytobands. A detailed description of the tests is given in the text.

	Adeno	Large-cell	Normal	Small-cell	Squamous	Adeno, C1	Adeno, C2	Adeno, C3	Adeno, C4	Adeno, CM	Adeno, NL
SVM	3.9	0.0	1.3	1.0	2.4	1.8	1.9	1.7	1.5	0.0	0.4
LAS	3.6	0.4	0.4	2.1	2.0	4.4	4.1	6.3	4.9	0.0	4.2
Plaid	6.8	0.4	1.1	1.3	2.6	3.2	5.3	6.4	3.2	1.1	2.4
CC	13.4	0.7	1.8	2.4	8.9	8.3	23.5	14.0	16.9	11.7	5.7
ISA	2.1	0.0	0.8	1.2	1.6	13.5	2.9	4.9	5.9	0.0	2.9
OPSM	25.3	9.2	6.7	2.4	8.3	14.7	16.8	18.6	17.6	13.7	18.7
SAMBA	37.6	7.7	7.5	2.4	9.7	14.7	16.8	18.9	17.9	13.7	21.1

Table 5: Classification error rates (in %) for SVM on the original data and the 5-nearest neighbor with weighted Euclidean distance applied to the "pattern" matrix.

2.2 Validation Results for Lung cancer data (samples with clinical information only)

Out of total 254 samples only 125 ADENO samples have clinical information. In this subsection we repeat the analysis limiting the dataset to the samples with the clinical information.

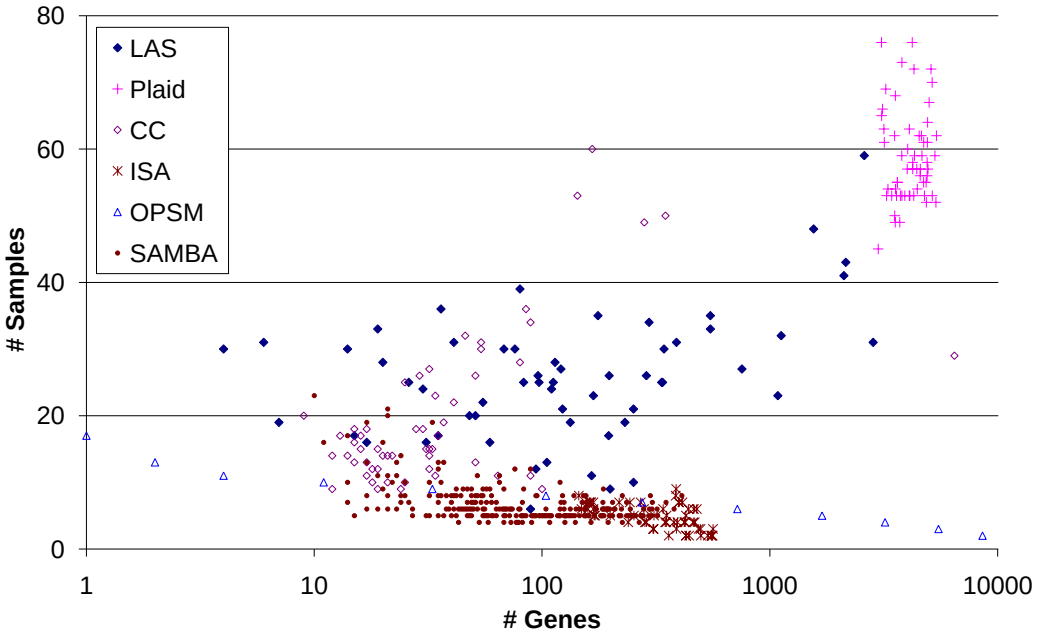


Figure 7: Bicluster sizes for different methods.

Method	# of Clusters	Eff. # of Clusters	Ratio
LAS	60	42.9	0.716
Plaid	60	6.4	0.107
CC	60	59.8	0.996
ISA	57	29.2	0.512
OPSM	12	7.3	0.609
SAMBA	281	164.8	0.586
KM	9,243	71.4	0.008
HC	80,631	692.9	0.009

Table 6: Output summary for different biclustering methods. From left to right: total number of biclusters produces; effective number of biclusters; the ratio of the effective number over the total number of biclusters.

	Correlation		Std. Dev.
	Gene	Sample	Gene
Matrix	0.01	0.04	0.83
KM	0.20	0.24	0.89
HC	0.43	0.26	0.88
LAS	0.34	0.13	1.83
Plaid	0.01	0.04	0.82
CC	0.13	0.07	1.73
ISA	0.24	0.35	0.99
OPSM	0.24	0.06	0.57
SAMBA	0.20	0.05	1.87
Subtypes		0.16	

Table 7: Average standard deviation of genes, and average pairwise correlation of genes and samples, for biclusters, IRCC clusters, and the whole data matrix.

	Adeno, C1	Adeno, C2	Adeno, C3	Adeno, C4	Adeno, CM	Adeno, NL
KM	5.9	7.3	9.5	10.4	9.8	0.4
HC	5.9	7.4	9.5	8.4	9.8	1.4
LAS	6.4	9.1	8.4	9.0	9.8	0.9
Plaid	0.0	0.8	0.0	0.2	0.7	0.0
CC	1.4	1.9	2.2	1.6	2.7	0.5
ISA	1.5	4.7	3.7	3.1	8.1	0.1
OPSM	0.0	0.0	0.0	0.0	0.8	0.0
SAMBA	0.4	0.6	0.9	1.0	0.5	0.5

Table 8: The minus $\log_{10}$ p-values of best subtype capture for different biclustering and sample clustering methods.

	# of BC's	Survival 5% level	KEGG/Cytoband Gene	KEGG/Cytoband Sample	2 out of 3	All 3
LAS	60	10	17	29	12	1
Plaid	60	1	0	5	0	0
CC	60	6	2	18	2	0
ISA	57	1	11	16	6	0
OPSM	12	0	3	1	0	0
SAMBA	281	12	32	58	13	0

Table 9: The number of biclusters passing tests for survival, and gene-set enrichment and sample-set differential expression of KEGG categories and cytobands. A detailed description of the tests is given in the text.

	Adeno, C1	Adeno, C2	Adeno, C3	Adeno, C4	Adeno, CM	Adeno, NL
SVM	5	5	4	3	1	3
LAS	5	0	3	5	2	3
Plaid	19	19	37	28	23	3
CC	16	14	24	22	14	3
ISA	15	7	27	16	2	3
OPSM	15	15	27	28	17	3
SAMBA	15	15	30	25	17	3

Table 10: Classification error rates (in %) for SVM on the original data and the 5-nearest neighbor with weighted Euclidean distance applied to the "pattern" matrix.

3 Simulations

3.1 Stability

				Average # of	
	N clusters	Eff num	ratio	Samples	Genes
LAS01	60	49.0	0.816	26.2	360.7
LAS02	60	48.6	0.811	26.0	358.5
LAS03	60	48.5	0.809	26.4	357.6
LAS04	60	48.3	0.805	26.7	357.4
LAS05	60	49.0	0.817	26.2	361.8
LAS06	60	48.6	0.810	26.0	360.8
LAS07	60	48.5	0.808	26.2	360.5
LAS08	60	48.2	0.804	27.1	358.7
LAS09	60	48.5	0.809	26.7	355.6
LAS10	60	48.8	0.814	25.9	363.0

Table 11: Summary table for 10 runs of LAS on the Hu data with different random seeds.

	ER	HER2-enriched	Basal-like	Normal-like	Luminal A	Luminal B
LAS01	9.1	10.9	18.5	8.8	8.5	4.2
LAS02	5.9	10.9	18.5	10.9	9.7	5.0
LAS03	7.4	11.3	19.9	10.1	8.5	6.3
LAS04	7.4	12.2	18.5	10.1	9.1	7.2
LAS05	7.4	10.9	16.4	10.1	10.0	4.7
LAS06	7.4	10.9	19.9	10.1	9.4	5.9
LAS07	7.4	10.9	19.9	10.1	9.4	6.9
LAS08	7.4	10.9	18.5	10.1	9.4	8.5
LAS09	8.3	11.3	18.5	8.3	9.1	4.2
LAS10	7.4	10.9	19.9	10.1	8.9	6.1

Table 12: Minus $\log_{10}$ p-values of best subtype capture for 10 runs of LAS with different random seeds.

	# of BC's	Survival 5% level	KEGG/Cytoband Gene	2 out of 2
LAS01	60	9	16	4
LAS02	60	11	16	4
LAS03	60	12	14	3
LAS04	60	13	14	4
LAS05	60	9	15	3
LAS06	60	10	13	3
LAS07	60	8	13	2
LAS08	60	13	13	3
LAS09	60	12	14	4
LAS10	60	9	14	2

Table 13: The number of LAS biclusters (for 10 runs of LAS with different random seeds) passing tests for survival, and gene-set enrichment of KEGG categories and cytobands. A detailed description of the tests is given in the paper.

3.2 Noise Sensitivity

				Average # of	
	N clusters	Eff num	ratio	Samples	Genes
$\sigma = 0.0$	60	48.3	0.806	26.2	357.3
$\sigma = 0.1$	60	49.2	0.819	26.6	359.1
$\sigma = 0.2$	60	49.4	0.823	26.9	342.8
$\sigma = 0.3$	60	47.7	0.795	26.0	338.0
$\sigma = 0.4$	60	49.5	0.825	26.9	319.8
$\sigma = 0.5$	60	50.3	0.838	26.7	297.2
$\sigma = 0.6$	60	50.3	0.839	26.8	274.5
$\sigma = 0.7$	60	50.9	0.849	26.5	252.9
$\sigma = 0.8$	60	51.8	0.863	25.5	227.3
$\sigma = 0.9$	60	51.7	0.862	25.1	208.5
$\sigma = 1.0$	59	51.6	0.875	25.8	183.0

Table 14: Summary statistics of the LAS biclusters for the data with added noise.

	ER	HER2-enriched	Basal-like	Normal-like	Luminal A	Luminal B
$\sigma = 0.0$	6.3	10.9	18.5	10.9	8.4	5.8
$\sigma = 0.1$	7.4	10.9	18.5	10.9	10.0	6.7
$\sigma = 0.2$	8.3	11.3	15.3	9.6	8.0	6.3
$\sigma = 0.3$	6.9	13.9	18.0	10.1	10.0	8.6
$\sigma = 0.4$	5.4	11.3	17.6	10.6	8.8	6.9
$\sigma = 0.5$	6.3	12.2	18.5	8.8	8.4	6.6
$\sigma = 0.6$	6.7	10.9	15.3	11.6	7.7	4.8
$\sigma = 0.7$	8.3	12.2	15.3	10.1	8.1	5.8
$\sigma = 0.8$	7.6	12.2	17.3	8.8	8.1	7.2
$\sigma = 0.9$	10.0	10.9	17.3	7.6	8.4	5.0
$\sigma = 1.0$	9.6	12.2	16.2	8.0	11.3	10.7

Table 15: Minus $\log_{10}$ p-values of best subtype capture for LAS ran on data with added noise.

	# of BC's	Survival 5% level	KEGG/Cytoband Gene	2 out of 2
$\sigma = 0.0$	60	13	16	3
$\sigma = 0.1$	60	9	15	2
$\sigma = 0.2$	60	12	14	2
$\sigma = 0.3$	60	10	14	3
$\sigma = 0.4$	60	9	13	2
$\sigma = 0.5$	60	9	14	3
$\sigma = 0.6$	60	8	13	3
$\sigma = 0.7$	60	7	13	1
$\sigma = 0.8$	60	8	12	1
$\sigma = 0.9$	60	11	9	3
$\sigma = 1.0$	59	5	9	1

Table 16: The number of LAS biclusters (on data with added noise) passing tests for survival, and gene-set enrichment of KEGG categories and cytobands. A detailed description of the tests is given in the paper.