Merging Two Gene Expression Studies via Cross Platform Normalization, Supplementary Materials

Andrey A Shabalin [1*], Håkon Tjelmeland[2],
Cheng Fan[3], Charles M Perou[3,4,5], and Andrew B Nobel[1]

February 6, 2008

[1]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology
[3]Lineberger Comprehensive Cancer Center, UNC-CH
[4]Department of Pathology and Laboratory Medicine, UNC-CH
[5]Department of Genetics, UNC-CH

Supplementary Materials

The supplementary materials of the paper consist of three sections. First section provides the heatmap illustration of the XPN model idea based on the real data. Next section presents validation results for both complete data set and on the "intrinsic" set of genes. The numbering of tables and figures here follow the numbering in the paper. Last section provides validation measures values for the analysis of stability of XPN with respect to number of clusters.

*to whom correspondence should be addressed

1 Illustration of the model idea on the real data.

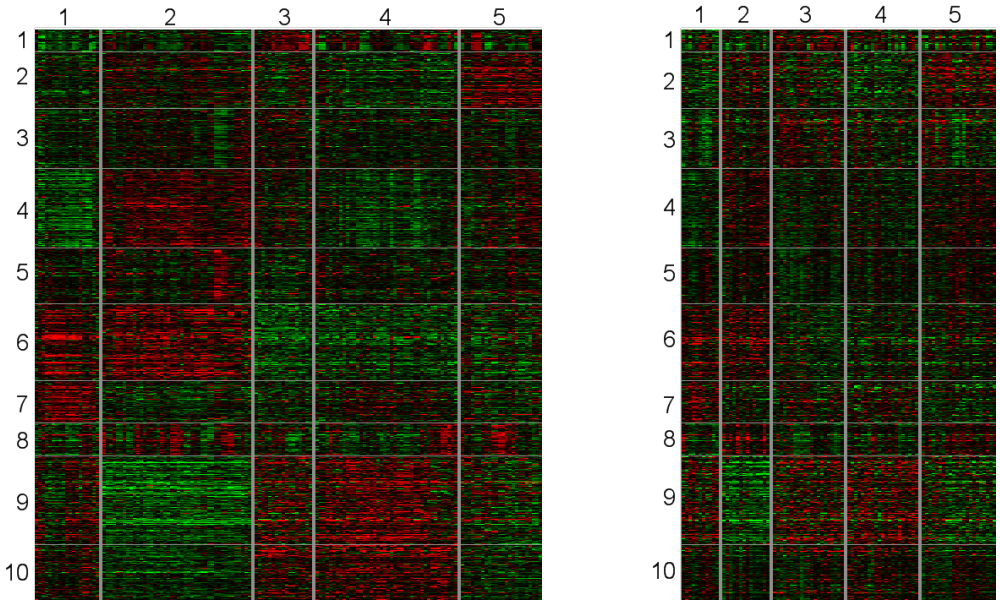

Figure 1: Heatmaps of NKI (left) and UNC (right) datasets after merging and clustering. The $K = 10$ row clusters and $L = 5$ column clusters were produced by k-means. Each column cluster (1 - 5) contains samples from both platforms.

2 Validation results for intrinsic gene list

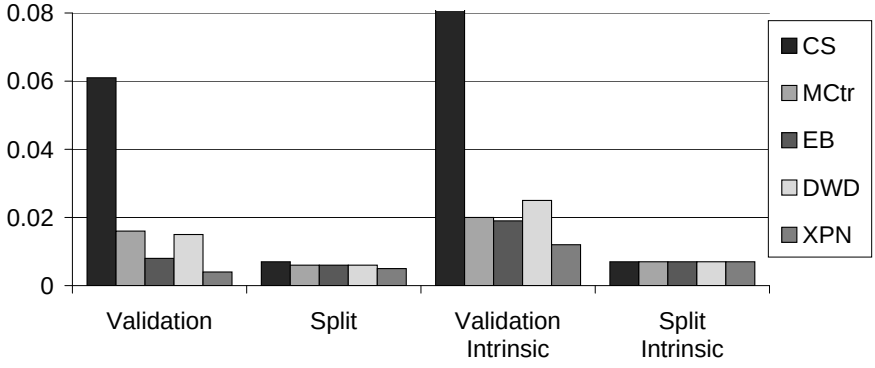

Fig. 2a. Area between the CDFs of array mean minus array median across platforms. Lower values indicate greater similarity of datasets after normalization.

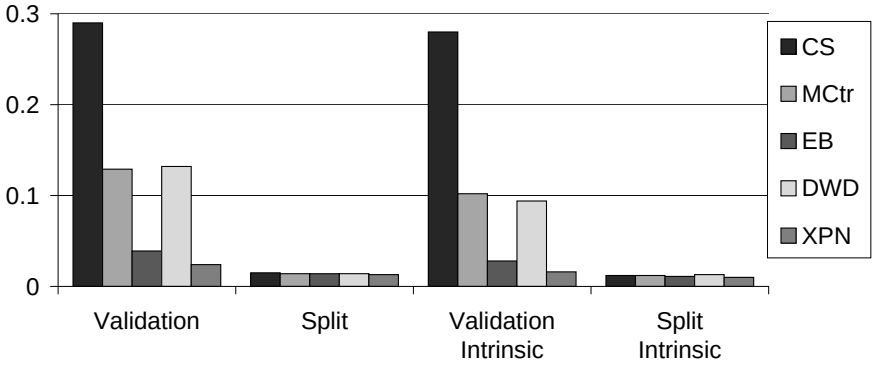

Fig. 3a. Area between the CDFs of $\sigma - MAD/\Phi(0.75)$ for arrays of different platforms. Lower values indicate greater similarity of datasets after normalization.

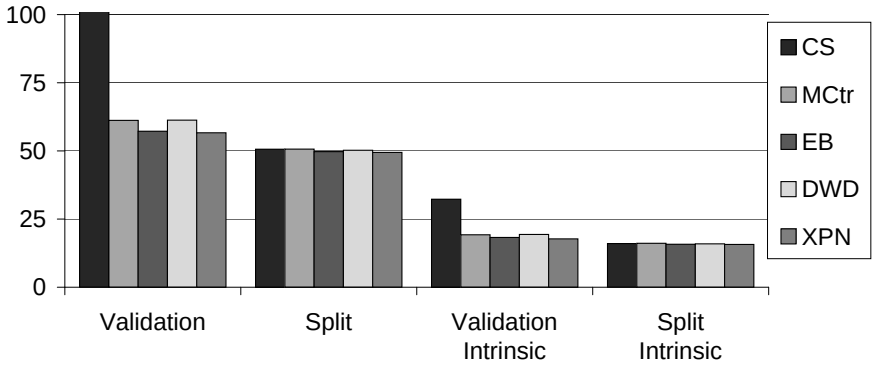

Fig. 4a. Average L_2 distance from the samples of one study to the nearest sample from the other study. Lower values indicate greater similarity of the study point "clouds" after normalization.

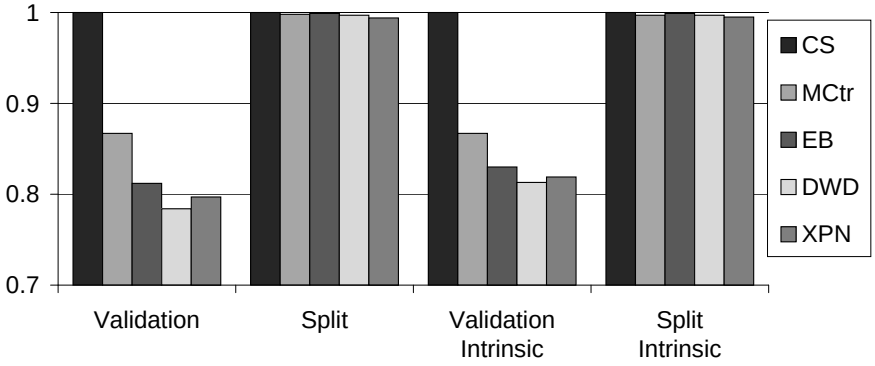

Fig. 5a. Average correlation of arrays with their values before normalization (CS). Larger values indicate less modification of the data by the normalization procedure.

		CS, MCtr, EB, DWD	XPN	Change	Change (%)
Avg gene corr w/ CS	Validn	1	0.990	-0.010	-1.0%
	Split	1	0.996	-0.004	-0.4%
GIC	Validn	0.255	0.338	0.083	32.5%
	Split	0.556	0.597	0.041	7.4%
ER t-stat correlation	Validn	0.312	0.451	0.139	44.6%
	Split	0.446	0.543	0.096	21.6%

Table 1. Gene correlation based validation measures on the **complete** set of genes.

		CS, MCtr, EB, DWD	XPN	Change	Change (%)
Avg gene corr w/ CS	Validn	1	0.993	-0.007	-0.7%
	Split	1	0.996	-0.004	-0.4%
GIC	Validn	0.471	0.579	0.108	22.9%
	Split	0.613	0.667	0.054	8.8%
ER t-stat correlation	Validn	0.478	0.669	0.191	40.0%
	Split	0.610	0.722	0.112	18.3%

Table 1b. Gene correlation based validation measures on the set of **intrinsic** genes. The first row shows the average correlation of genes with their value before normalization (CS). The second row shows global integrative correlation (GIC) between platform pairs after normalization, with larger values indicating better concordance between platforms. The third row shows the average correlation of ER t-statistics across platforms, with larger values indicating better concordance.

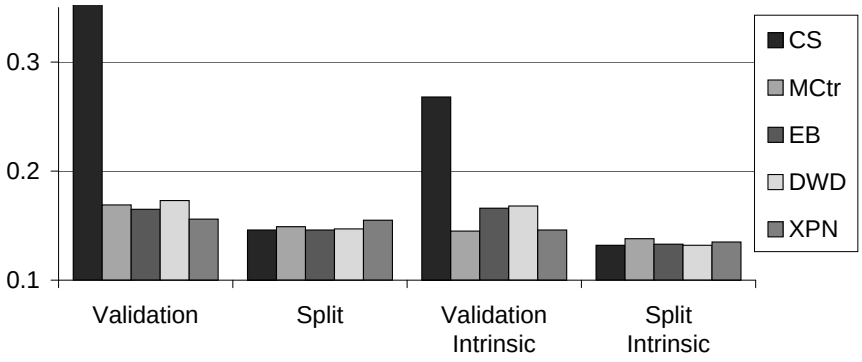

Fig. 6a. Cross platform prediction error of the PAM (nearest shrunken centroids) classifier. Smaller values indicate better concordance between platforms.

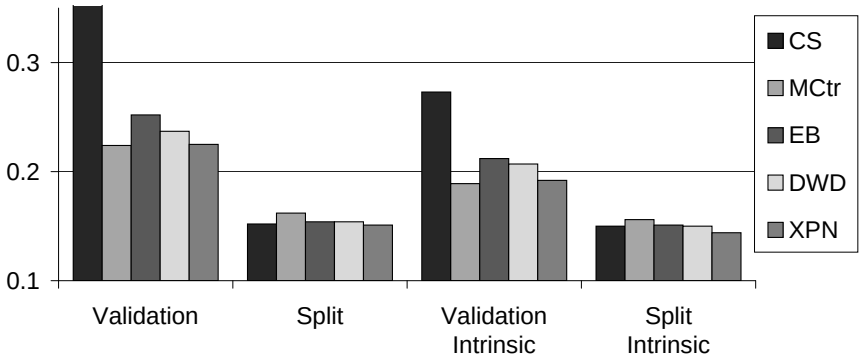

Fig. 7a. Cross platform prediction error of the SVM (Support Vector Machine) classifier. Smaller values indicate better concordance between platforms.

		CS	MCtr	EB	DWD	XPN
V_1	Validn	0.826	1	1	1	1
	Split	0.999	0.999	0.999	0.999	0.999
V_2	Validn	0.646	0.895	0.774	0.887	0.759
	Split	0.876	0.867	0.875	0.878	0.870

Table 2. Measures of preservation of gene lists on the **complete** set of genes.

		CS	MCtr	EB	DWD	XPN
V_1	Validn	0.876	1	1	1	1
	Split	1	1	1	1	1
V_2	Validn	0.689	0.938	0.886	0.937	0.879
	Split	0.946	0.941	0.945	0.947	0.940

Table 2a. Measures of preservation of gene lists on the set of **intrinsic** genes. V_1 (V_2) is the fraction of genes from the intersection (union) of platform-specific gene lists present in the list produced from the combined data $\tilde{X}$ at 0.1% level.

3 Stability of XPN with respect to number of clusters K and L

To test stability of XPN with respect to the numbers K and L of row and column clusters, we applied XPN with a range of parameters. For $L = 5$ we tried $K = 2, 10, 20, 25, 30, 50, 100, 500$, and for $K = 25$ we tried $L = 2, 4, 5, 6, 7, 8, 10$. The results indicate that XPN is generally insensitive to the choice of the K and L. However, we do see (expected) degradation of performance in situations where K or L is below 4, in which case the clustering is too coarse to adequately capture homogenous blocks of samples or genes.

		MC	EB	DWD	XPN	XPN	XPN	XPN	XPN	XPN	XPN	XPN	XPN	XPN
	K (gene clusters)				25	2	2	10	20	25	30	50	100	500
	L (sample clusters)				5	2	5	5	5	5	5	5	5	5
Fig 2	CDFs of $\mu-$ median	0.016	0.008	0.015	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
Fig 3	CDFs of $\sigma - MAD/\Phi(0.75)$	0.129	0.039	0.132	0.024	0.024	0.021	0.023	0.024	0.024	0.023	0.023	0.024	0.024
Fig 4	Avg min dist to other array	61.2	57.2	61.3	56.6	57.4	57.0	56.7	56.6	56.6	56.6	56.5	56.5	56.2
Fig 5	avg array corr w/ CS	0.867	0.812	0.784	0.797	0.798	0.796	0.797	0.798	0.797	0.798	0.798	0.798	0.798
Table 1(1)	Avg gene corr w/ CS	1	1	1	0.991	0.992	0.986	0.99	0.991	0.991	0.991	0.991	0.991	0.991
Table 1(2)	GIC	0.255	0.255	0.255	0.337	0.31	0.293	0.333	0.338	0.337	0.341	0.344	0.344	0.349
Table 1(3)	ER t-stat corr	0.312	0.312	0.312	0.452	0.459	0.422	0.457	0.455	0.452	0.454	0.449	0.452	0.478
Fig 6	PAM ER class error	0.169	0.165	0.173	0.161	0.193	0.195	0.16	0.154	0.161	0.154	0.158	0.158	0.163
Fig 7	SVM ER class error	0.224	0.252	0.237	0.223	0.242	0.241	0.228	0.225	0.223	0.219	0.221	0.223	0.224
Table 2(1)	V_1	1	1	1	1	1	1	1	1	1	1	1	1	1
Table 2(2)	V_2	0.895	0.774	0.887	0.753	0.764	0.651	0.745	0.747	0.753	0.753	0.755	0.75	0.75

Table 1: Validation measures for different values of gene clusters K. Presented in the same order as in the paper.

		MC	EB	DWD	XPN	XPN	XPN	XPN	XPN	XPN	XPN	XPN
	K (gene clusters)				25	25	25	25	25	25	25	25
	L (sample clusters)				5	2	4	5	6	7	8	10
Fig 2	CDFs of $\mu-$ median	0.016	0.008	0.015	0.004	0.005	0.004	0.004	0.004	0.004	0.004	0.004
Fig 3	CDFs of $\sigma - MAD/\Phi(0.75)$	0.129	0.039	0.132	0.024	0.024	0.022	0.024	0.023	0.024	0.023	0.022
Fig 4	Avg min dist to other array	61.2	57.2	61.3	56.6	57.4	56.7	56.6	56.5	56.5	56.4	56.3
Fig 5	avg array corr w/ CS	0.867	0.812	0.784	0.797	0.798	0.798	0.797	0.798	0.798	0.798	0.797
Table 1(1)	Avg gene corr w/ CS	1	1	1	0.991	0.991	0.991	0.991	0.991	0.99	0.99	0.989
Table 1(2)	GIC	0.255	0.255	0.255	0.337	0.294	0.333	0.337	0.345	0.351	0.356	0.36
Table 1(3)	ER t-stat corr	0.312	0.312	0.312	0.452	0.474	0.467	0.452	0.44	0.431	0.423	0.413
Fig 6	PAM ER classif error	0.169	0.165	0.173	0.161	0.195	0.154	0.161	0.149	0.149	0.144	0.148
Fig 7	SVM ER classif error	0.224	0.252	0.237	0.223	0.24	0.232	0.223	0.223	0.223	0.22	0.22
Table 2(1)	V_1	1	1	1	1	0.999	1	1	1	1	1	1
Table 2(2)	V_2	0.895	0.774	0.887	0.753	0.719	0.734	0.753	0.762	0.761	0.755	0.755

Table 2: Validation measures for different values of sample clusters L. Validation measures are presented in the same order as in the paper.