Subtype-Aware Batch Correction Retains Biological Signal of Integrated Breast Cancer Datasets

Gil Tomás
gil.tomas@igmm.ed.ac.uk

February 9, 2018

Contents

1 Dataset Acquisition

Raw CEL files from ten breast cancer gene expression datasets where downloaded from GEO and normalized with fRMA, RMA and MAS5. Demographics of each dataset are shown in Table 1.

2 Micorarray Dataset Integration

Microarray dataset integration needs to account for technical, non-biological, variation across multiple batches of independently acquired datasets. The goal of conventional batch correction (BC) is to remove batch effects while retaining biological variation conveyed by each dataset. Several methods exist to address this task. One popular approach is ComBat, which proposes "parametric and nonparametric empirical Bayes frameworks for adjusting data for batch effects that is robust to outliers in small sample sizes and performs comparable to existing methods for large samples". However, most BC integration strategies do not account for imbalanced subtype composition within datasets.

This manuscript introduces a novel procedure for integrating expression profile datasets of known dissimilar composition, called *subtype-aware batch correction* (SABC). Molecular subtypes are initially assigned to each sample in each dataset with a publicly available single sample predictor (SSP). Then, batch effects are resolved between datasets on a per-subtype basis. This two-layered approach allows

Table 1: Demographics of datasets in this study.

GSE	platform	nSamples	fracER+	fracHER2+
GSE5327	HG-U133A	58	0.00	—
GSE25065	HG-U133A	198	0.62	0.01
GSE2034	HG-U133A	286	0.73	—
GSE17705	HG-U133A	298	1.00	—
GSE16446	HG-U133Plus2	120	0.00	0.33
GSE17907	HG-U133Plus2	33	0.47	1.00
GSE21653	HG-U133Plus2	266	0.57	0.12
GSE5460	HG-U133Plus2	127	0.58	0.24
GSE2109	HG-U133Plus2	353	0.65	0.27
GSE23177	HG-U133Plus2	116	1.00	0.00

for the biological specifics captured by the single sample predictor of choice to be accounted for during the batch correction step, and thus carried over into the integrated dataset. By ignoring distinct subtype compositions between datasets, conventional BC incorrectly apprehends biological variation as technical batch effect, and consequently distorts true biological signal in the integrated dataset.

SABC still requires a methodology to address batch effects; we chose ComBat for this task. Due to the split of samples by subtype prior to integration, some subtype-scpecific batches may not have enough samples to draw summary statistics from; or all samples of a lower frequency subtype may only be present in one given batch. Samples that fall in these categories are not considered (removed from the analysis) by SABC.

Breast cancer is widely understood to be subdivided into five intrinsic or molecular subtypes, which can be assigned by gene expression profile single sample predictors. Breast cancer hence provides an ideal case study for the evaluation of SABC. We used conventional BC and SABC to integrate four breast cancer datasets hybridized onto the Affymetrix HG-U133a chip and six breast cancer datasets hybridized onto the HG-U133Plus2 chip (Table 1).

To evaluate the performance of each method, we compared the distributions of expression values for the 205225_at and the 216836_s_at probesets in each chip, respectively targeting for the *ESR1* and *ERBB2* gene transcripts. Estrogen receptor (ESR1) and Erb-B2 Receptor Tyrosine Kinase 2 receptor (ERBB2) status are strong predictors of breast cancer prognosis and are traditionally assessed by immunohistochemestry (IHC). Post dataset integration, the expression values of both genes should remain in line with the biological signal conveyed by the independent assessment given by IHC for both proteins. In addition, we compared the agreement between single sample predictor class assignments for individual samples prior and post integration. The batch correction procedure should not interfere with the molecular subtype identity of each sample, and for that reason higher agreement rates should be indicative of higher transcriptional fidelity of the integrated dataset.

To guide the integration process with SABC, we used two SSPs implemented in the Genefu Bioconductor package. The first, sorlie2003, is based on 534 diagnostic genes and is a five-subtype classifier; the second, desmedt2008, is based on three genes (*ESR1*, *ERBB2* and *AURKA*), and is a three-subtype classifier. To assess classifier agreement prior and post integration, and in order to circumvent the redundancy caused by using the same single sample predictor to integrate and to validate the integration process, we used the Genefu implementation of the PAM50 single sample predictor, based on 50 genes and yielding a five-subtype classifier.

Table 2: Partition of desmedt2008 subtypes in the ten datasets in this study (fRMA normalization).

	ER-/HER2-	ER+/HER2-	HER2+
HG-U133a			
GSE5327	37	13	8
GSE25065	75	93	30
GSE2034	61	189	36
GSE17705	54	218	26
HG-U133Plus2			
GSE16446	77	19	24
GSE17907	6	15	12
GSE21653	75	149	42
GSE5460	35	74	18
GSE2109	97	210	46
GSE23177	15	78	23

2.1 Prior SSP predictions

Prior to datset integration with SABC, we compute SSP classes for each dataset in our analysis. Tables 2 and 3 respectively show the partitions of each of the 3- and 5-class SSPs across the four HG-U133a datasets in this study. We then computed SSP predictions for both the desmedt2008 and sorlie2003 classifier in each integrated expression matrix.

2.2 Integration

Integrated matrices of gene expression were computed for each normalization method (fRMA, RMA and MAS5) and each integration method (BC and SABC, driven by 3- and 5-subtype classifier).

3 Comparison of Integration Methods

3.1 Distortion of Molecular *ESR1* measurements

We compared the distributions of expression values prior and post dataset integration for the 205225_at probeset in 840 breast tumours, hybridized onto the HG-U133a chip, from four datasets with distinct fractions of ER+ samples (Table 1 and Figure 1). Regardless of the normalization method, BC integration significantly distorts *ESR1* expression measurements in samples from datasets with extreme fractions of ER+ samples (GSE5327 and GSE17705), to the point where the two distributions no longer can tell the difference between ER− and ER+ samples (Figure 2, second column of panels). In both cases, SABC integration, whether driven by a 3-subtype SSP (SABC3) or a 5-subtype SSP (SABC5), succeeds in retaining the biological signal conveyed by the IHC status in datasets with extreme compositions (Figure 2, third and fourth columns of panels). Further comparison of ESR1 transcript abundance in these two datasets prior and post integration reveals that expression values depart significantly from original measurements when datasets are integrated with BC, yet are preserved from extreme distortion by SABC integration (Figure 3).

Table 3: Partition of sorlie2003 subtypes in the ten datasets in this study (fRMA normalization).

	Basal	Her2	LumA	LumB	Normal
HG-U133a					
GSE5327	32	—	1	25	—
GSE25065	48	—	39	108	3
GSE2034	48	1	84	152	1
GSE17705	6	—	144	140	8
HG-U133Plus2					
GSE16446	68	1	1	50	—
GSE17907	2	4	3	23	1
GSE21653	68	7	89	95	7
GSE5460	32	—	47	48	—
GSE2109	65	5	128	144	11
GSE23177	2	—	33	81	—

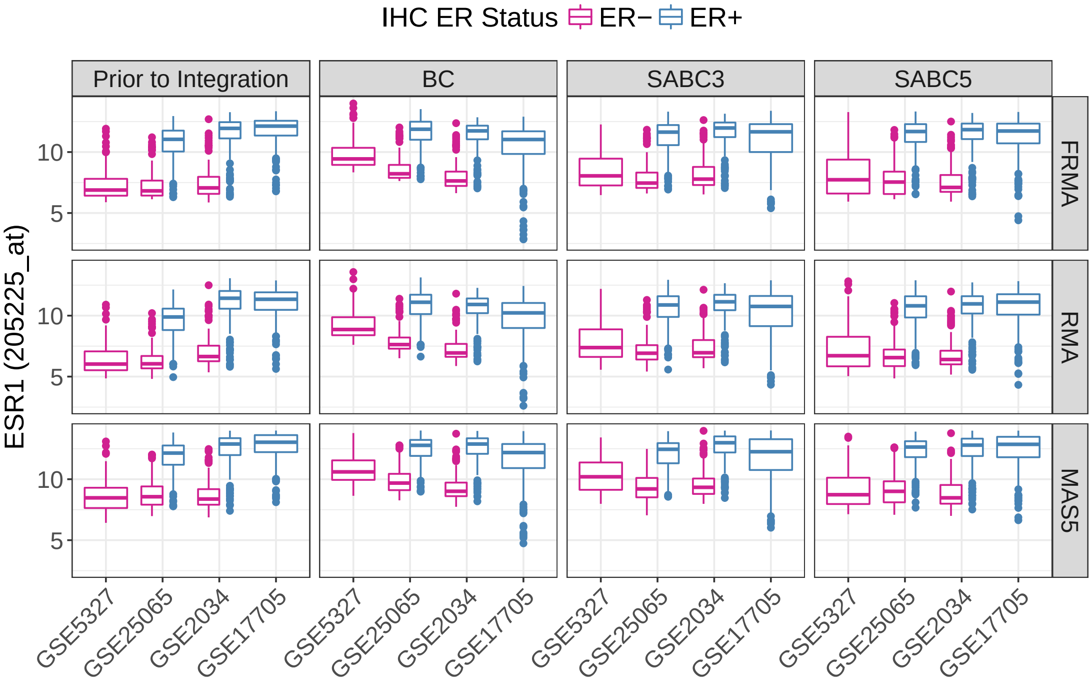

Figure 1: 20225_at probeset measurements hybridized onto the HG-U133a chip broken by dataset prior and post integration. Integration was done using standard batch correction (BC, with ComBat) and subtype-aware batch correction (SABC3, driven by a three-subtype SSP—desmedt2008; and SABC5, driven by a five-subtype SSP—sorlie2003). The distributions are further split by ER status, independently assessed by IHC on fresh frozen specimens. Raw data was normalised with FRMA, RMA and MAS5.

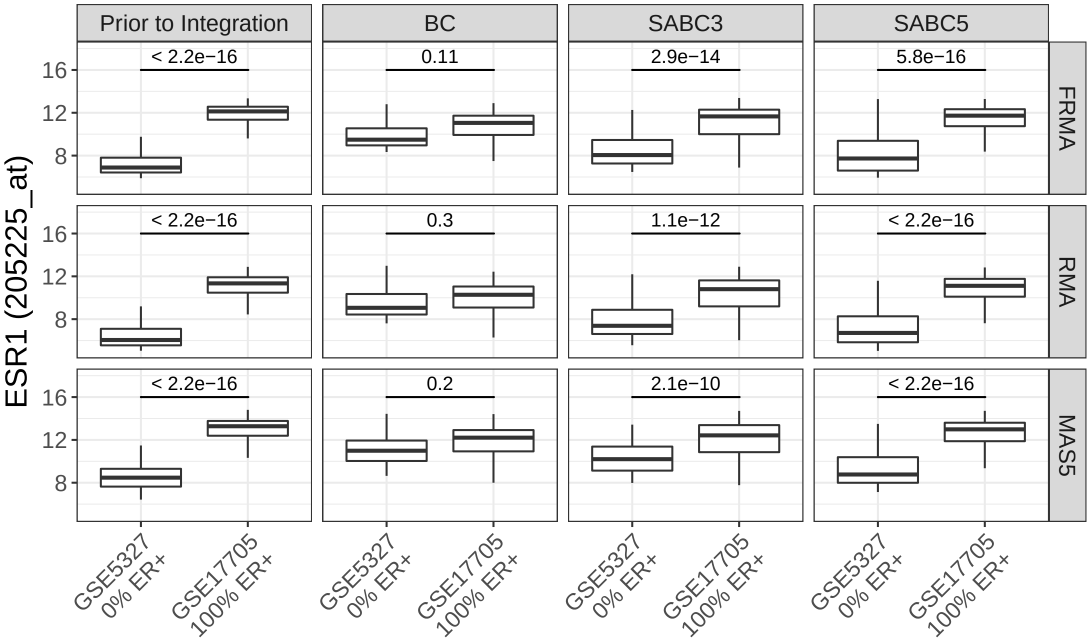

Figure 2: Distributions of probeset 205225_at measurements from datasets GSE5327 (n=58, all ER-) and GSE17705 (n=298, all ER+), taken from Figure 1, are compared side by side. See Figure 1 for details.

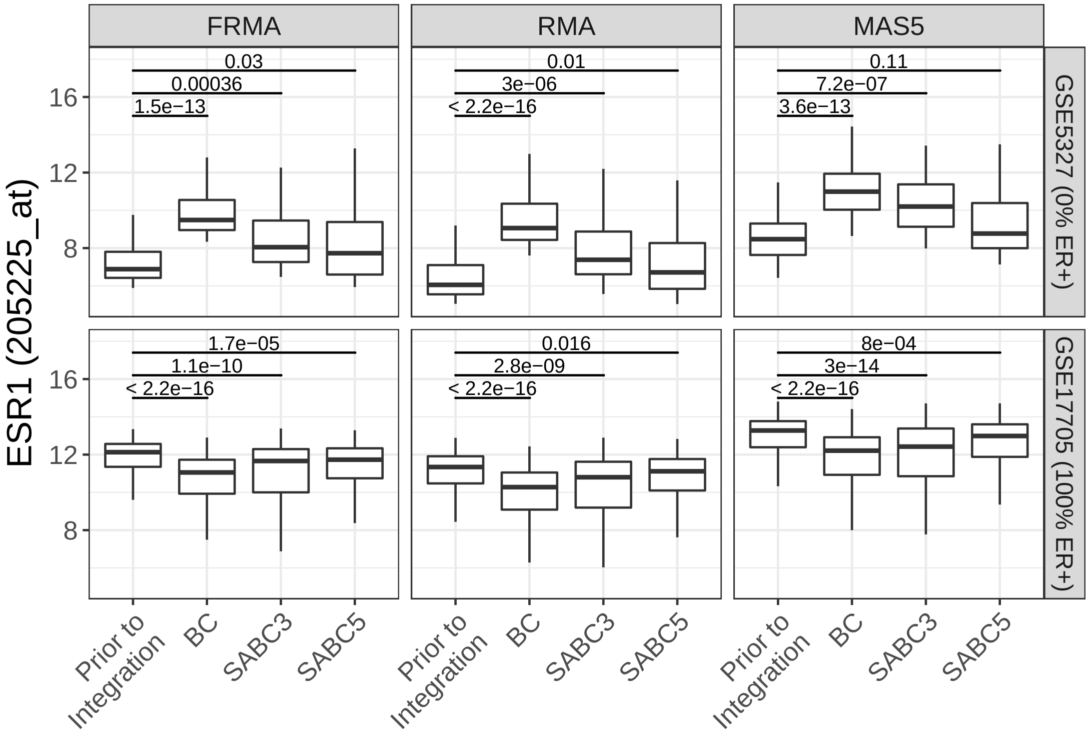

Figure 3: Distributions of probeset 205225_at measurements from datasets GSE5327 (n=58, all ER-) and GSE17705 (n=298, all ER+), taken from Figure 1, are each compared prior and post dataset integration. See Figure 1 for details.

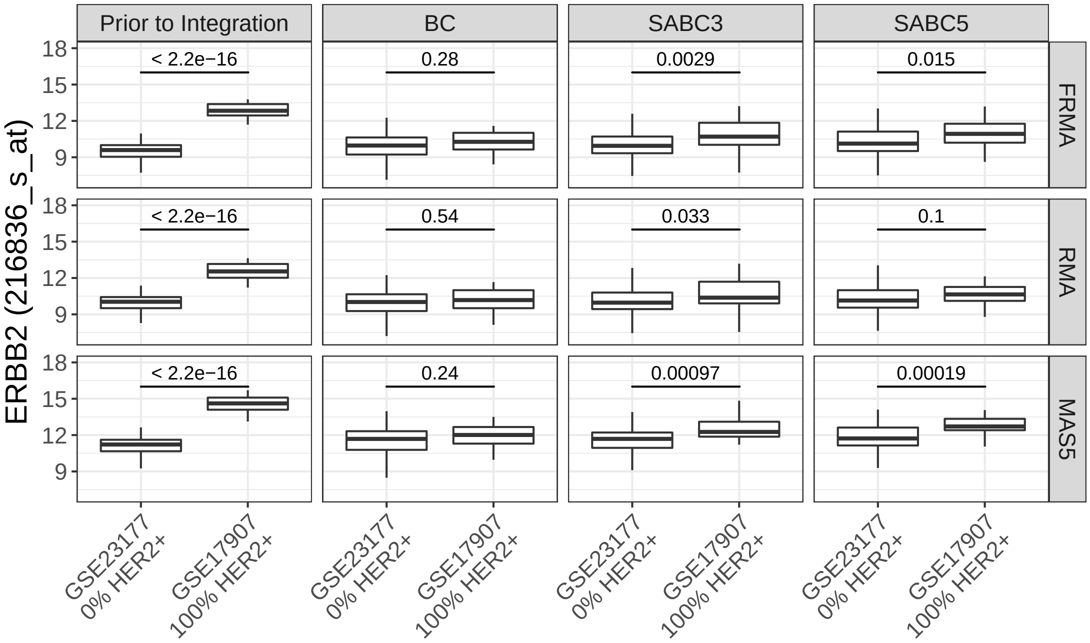

Figure 4: 216836_s_at probeset measurements hybridized onto the HG-U133Plus2 chip are shown for datasets GSE23177 (n=116, all HER2-) and GSE17907 (n=37, all HER2+), in the leftmost column. These two datasets were integrated with GSE16446, GSE21653, GSE5460 and GSE2109 (Table 1), with BC, SABC3 and SABC5. ERBB2 expression values for the samples in the two datasets with extreme HER2+ compositions are then shown post-integration with each of these methods, broken by normalisation procedure (cf. Figure 1 for more details).

3.2 Distortion of Molecular *ERBB2* measurements

Although less pronounced, a similar trend is observed when comparing the expression values prior and post- dataset integration for the 216836_s_at probeset, in 1015 breast tumours hybridized onto the HG-U133Plus2 chip, from six datasets with distinct proportions of HER2+ samples (Table 1 and Figure 4). Regardless of the normalisation protocol, The ERBB2 probeset distributions clearly reflect the IHC HER2 receptor status in the two datasets in our analysis with extreme HER2 compositions (GSE23177, all HER2–; and GSE17907, all HER2+). When the five datasets are integrated with BC, this biological signal is erased, yet preserved (albeit to a lesser extent than in the original datasets), when integrated with SABC3 and SABC5 (Figure 4).

The different degree to which subtype-aware batch correction successfully retains biological signal from these two molecular correlates of breast cancer biology, in datasets with extreme compositions, could be explained by how well the classifiers used to drive integration capture the underlying biology of the ER and HER2 receptors. Because the clinical HER2 breast cancer phenotype is the result of a gene amplification, it is possible that gene expression classifiers are less apt to model binary gene expression distributions (HER2) rather than continuous ones (ER). In addition, the top level split highlighted by most molecular characterisations of breast tumours is lead by ER status, consigning most SSP derived from molecular data to be particularly sensitive to biological signal conveyed by this marker.

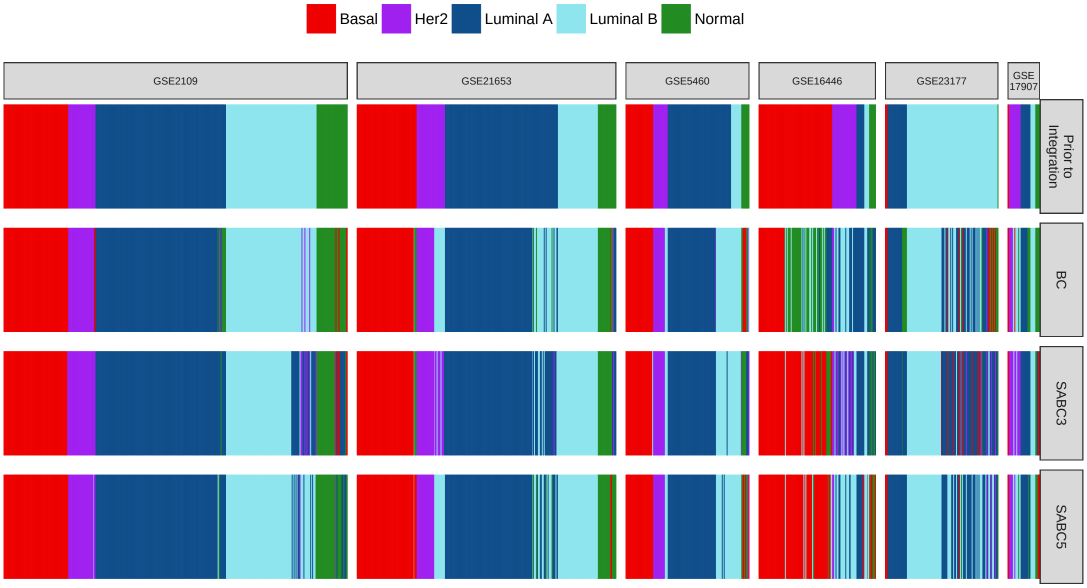

Figure 5: PAM50 subtype assignments for 1015 samples from six datasets (Table 1) hybridized on the HG-U133Plus2 chip and normalized with FRMA. Subtype assignements were computed prior to dataset integration (first row) and post integration with BC, SABC3 and SABC5 (subsequent rows).

3.3 Single Sample Predictor Agreement

Single sample predictor assignments for the Genefu implementation of the PAM50 breast cancer classifier were computed for each of the 1015 samples in GSE2109, GSE21653, GSE5460, GSE16446, GSE23177 and GSE17907, after fRMA normalization (Table 1). We then integrated the six datasets with BC, SABC3 and SABC5, and computed each sample's PAM50 subtype post-integration. Comparisons of subtype assignments prior- and post-integration can be seen in Figure 5 and in Table 4.

With the exception of the largest dataset, GSE2109, PAM50 interrater agreement was always higher with SABC integration than with conventional integration. Incidentally, the datasets that showed lesser subtype agreement post BC integration are the ones with most extreme ER and HER2 compositions (GSE16446, GSE23177 and GSE17907). For these datasets, SABC integration was able to increase subtype assignment agreement (with the exception of SABC3 for GSE23177).

3.4 Effect of Sample Size

Because sample size can further confound the effect of dataset composition towards retention of biological signal post-integration, we compared the PAM50 interrate agreement prior and post-integration for datasets GSE2109, GSE21653, GSE5460, GSE16446 and GSE23177, sampling, in each dataset, 116 samples with the same IHC ER+ proportions as the original datasets. Comparisons of subtype assignements prior and post-integration in this experiment can be seen in Figure 6 and Table 5. Notice how dataset size normalization further emphasises the need for subtype-aware batch correction.

Table 4: Interrater agreement of PAM50 subtype assignment prior- and post-dataset integration (see text for details).

	all	GSE2109	GSE21653	GSE5460	GSE16446	GSE23177	GSE17907
Clinical							
nSamples	1015	353	266	127	120	116	33
% ER+	—	65	57	58	0	100	47
% HER2+	—	27	12	24	33	0	100
Cohen's Kappa (%)							
BC	72	95	80	72	21	30	58
SABC3	78	85	90	75	63	25	79
SABC5	80	93	85	76	50	40	61

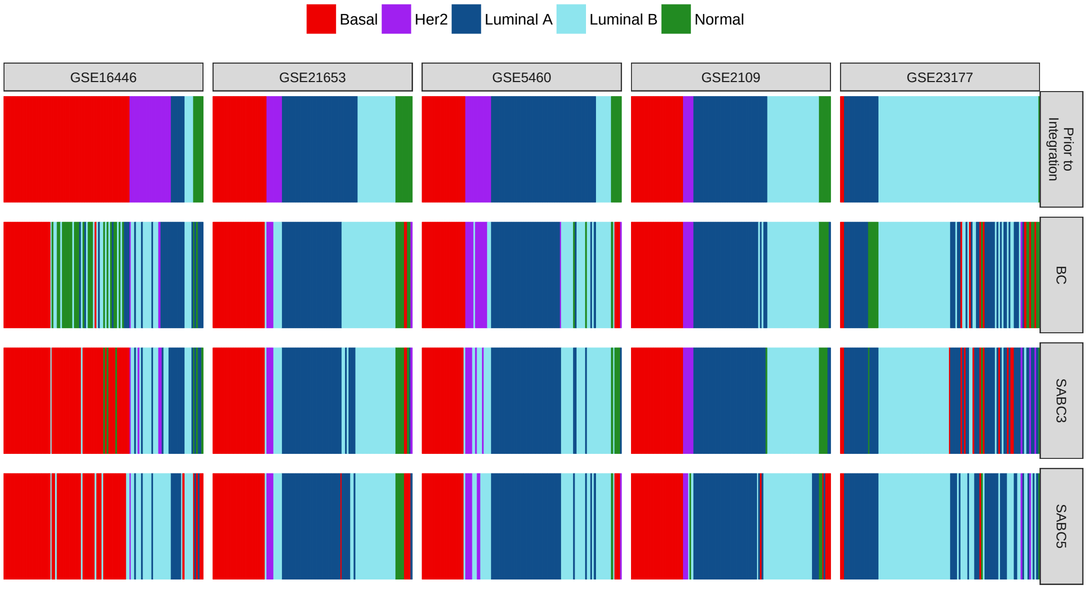

Figure 6: PAM50 subtype assignments for 580 samples from five datasets (Table 1) hybridized on the HG-U133Plus2 chip and normalised with FRMA. For each dataset, 116 samples were randomly selected in order to respect the original IHC ER+ fraction. Subtype assignments were computed prior- (top row) and post- (subsequent rows) dataset integration with BC, SABC3 and SABC5

Table 5: Interrater agreement of PAM50 subtype assignment prior- and post-dataset integration, when datasets are all brought down to the same size (n = 116, see text for details).

	all	GSE16446	GSE21653	GSE5460	GSE2109	GSE23177
Clinical						
nSamples	580	116	116	116	116	116
% ER+	—	0	57	58	65	100
% HER2+	—	33	12	24	27	0
Cohen's Kappa (%)						
BC	62	23	79	69	96	32
SABC3	71	59	84	65	96	31
SABC5	69	45	82	63	81	47

4 Comparison with Published Methods

SABC is not the first method to account for external biological variables or clinical covariates during dataset integration. ComBat offers the possibility of using clinical variables as covariates during dataset integration. Harman, a more recent integration procedure using a PCA and a constrained optimisation technique, also calls for factor coding for an "experimental grouping variable" to drive integration. We ran Harman to integrate the 840 breast tumours hybridised onto the the HG-U133a chip (normalized with fRMA), using the the output of the 5-class sorlie2003 SSP as covariate and experimental grouping variable respectively. As seen in Figure 7, dataset integration with Harman fails to retain the original difference in ESR1 measurements prior to integration for datasets GSE5327 and GSE17705.

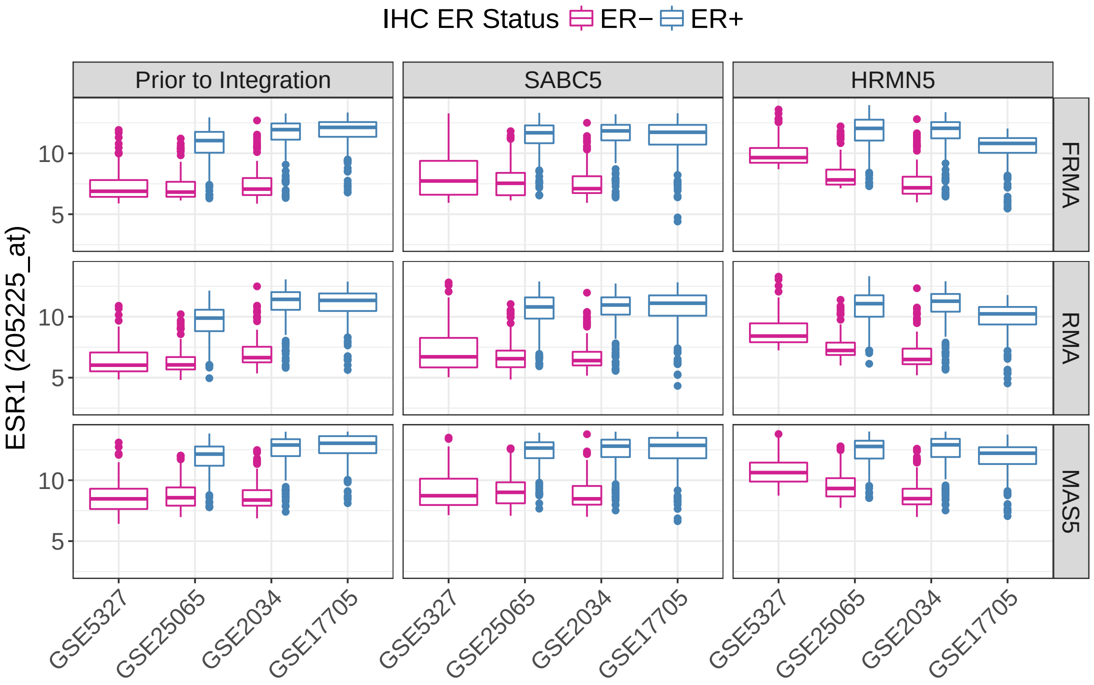

Figure 7: 205225_at probeset measurements hybridized onto the HG-U133a chip broken by dataset prior and post integration. Integration was done using subtype-aware batch correction (SABC5, driven by a five-subtype SSP) and Harman (HRMN5, using a five-subtype SSP as a grouping variable). The distributions are further split by ER status, independently assessed by IHC on fresh frozen specimens. Raw data was normalised with FRMA, RMA and MAS5.