Supplemental Methods

A. SVA (Surrogate Variable Analysis): Identification of Latent Batch Effect

The Surrogate Variable Analysis (SVA) assumes a general linear model on latent (or surrogate) variables and the primary factor of interest histological type for log counts: $\log(Y) = \mu + \alpha X + \beta W + E_1$, where

Y is raw counts for *N* samples and *G* genes, *X* is design matrix for histological type, *W* is design matrix for all surrogate variables (SV's), and E_1 is the random error matrix following normal distribution.

Step 1: Fit the model $Y = \mu + \alpha X +$ R with OLS and generate the residual matrix R.

Step 2: Construct principal components (PCs) of R by singular value decomposition and select K most significant PCs by either method: 'BE' or 'Leek', denoted as $e_{k,}\ k = 1, \dots, K$.

Step 3: Regress e_k on each gene and calculate the association p-value between e_k and each gene.

Step 4: For each e_k, pick h genes with smallest p-values from the previous step and form the $h \times m$ reduced expression matrix Y_k. h is determined by the q-value approach.

Step 5: Construct the PCs of Y_k by singular value decomposition, denoted as $e_i^r, i = 1, \dots, m$.

Step 6: Calculate the correlation between e_k and each e_i^r; the kth surrogate variable is the e_i^r with the largest correlation with e_k.

We used 'BE' method to determine the number (14) of SV's, because 'Leek' method gives only 2 SV's that are both highly associated with known batch (i.e. sequencing project). The 'BE' method not only generated SV's associated with known batch, but also identified latent variables independent of known batch. Association between each SV and known batch by either method is shown by the p-value of Fisher's exact test in **Table 1**.

Table 1: Association between SV's and Known Batch by Fisher's Exact Test

SV by Leek Method		SV by BE Method						
		SV1	SV2	SV3	SV4	SV5	SV6	SV7
SV10	SV2	<0.001	<0.001	0.062	<0.001	0.01	0.16	0.02
<0.001	<0.001	SV8	SV9	SV10	SV11	SV12	SV13	SV14
		0.47	0.002	0.48	0.04	0.32	0.11	0.044

B. ComBat: Adjustment of All Batch Effects

We applied the random effect model on histology and SV's to adjust for latent batch effects. Empirical Bayes method is used in the estimation of this model. Since SV's given by Bioconductor package *sva* are continuous, they must be converted to categorical factors (with 4 levels) by quantiles before fitting the random effect model. This model can be specified as $Y = \mu + \alpha X + \beta W^* + \delta E_2$.

Y, X are the same matrices defined in SVA, W^* is design matrix for only one SV, and $E_2$is random error matrix with gene-wise variance $\sigma^2 = (\sigma_1^2, \dots, \sigma_G^2)$. We do not use log counts in ComBat, because the SV's are all transformed to be categorical and treated as 'unknown' batch factors. This method is implemented by the *combat* function in *sva* package, which restricts the number of confounders to be 1. Hence, W^* involves only one SV and the normalization must be applied repeatedly, see **Figure 1**.

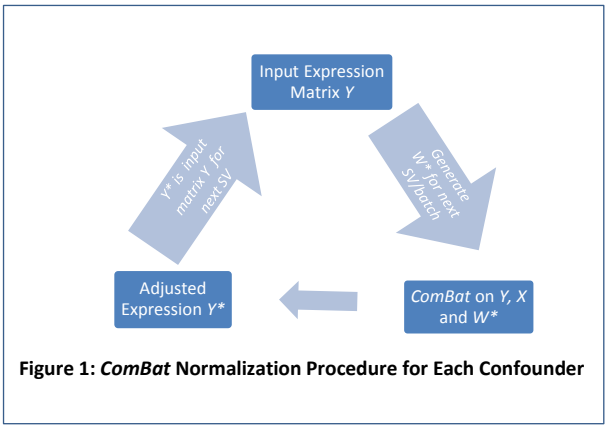

Figure 1: *ComBat* Normalization Procedure for Each Confounder

Step 1: Compute OLS estimates $\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\delta}, \hat{\sigma}$; standardize Y (i.e. adjust for histological effect) by $Z = (Y - \hat{\mu} - \hat{\alpha}X)/\hat{\sigma}$.

Step 2: Compute empirical Bayes estimate $\hat{\beta}^*, \hat{\delta}^*$ on Z.

Step 3: Adjust Z for $\hat{\beta}^*, \hat{\delta}^*$ effect with adjusted matrix Z^*; add and multiply the effect of $\hat{\mu}, \hat{\alpha}, \hat{\sigma}$ to Z^*, i.e. $Y^* = \hat{\sigma}Z^* + \hat{\mu} + \hat{\alpha}X$.

After adjusting unknown batch effects represented by SV's, we perform *ComBat* again on the adjusted data to further adjust for known batch, because SV's may not fully capture known batch effect.

We check for the remained batch effects by PCA visualization. **Figure 2** presents the noise effect of sequencing projects along with technical artifacts, which is not observed in **Figure 3**. Some samples from project 2 & 3 still cluster in **Figure 3**, because they are all clear cell samples as shown by **Figure 4**, which illustrates that histological effect remains in the adjusted data.

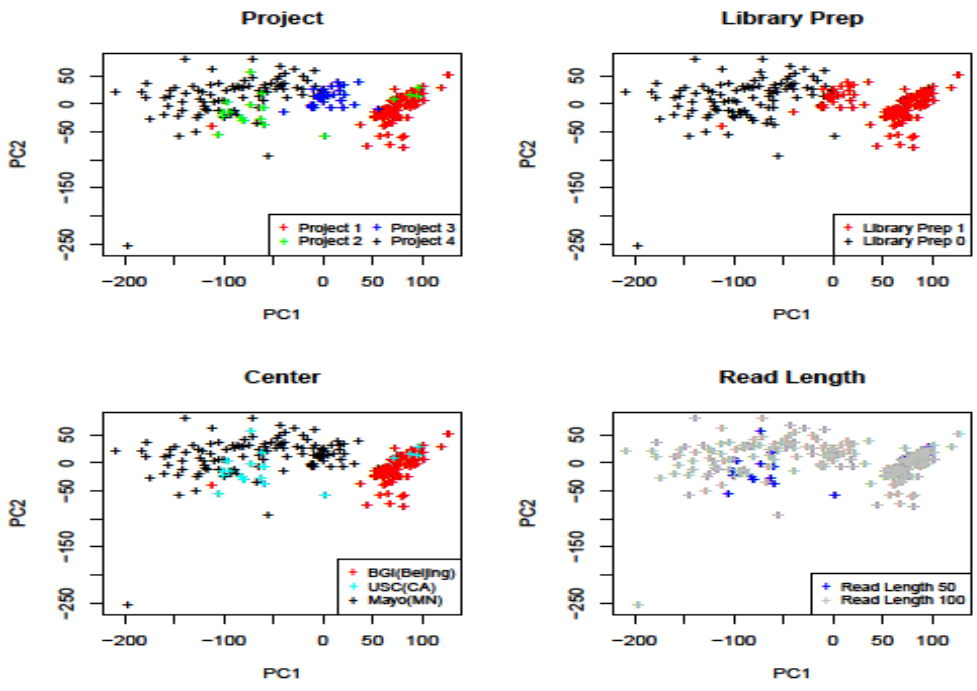

Figure 2: PCA analysis of raw (non-normalized) log counts

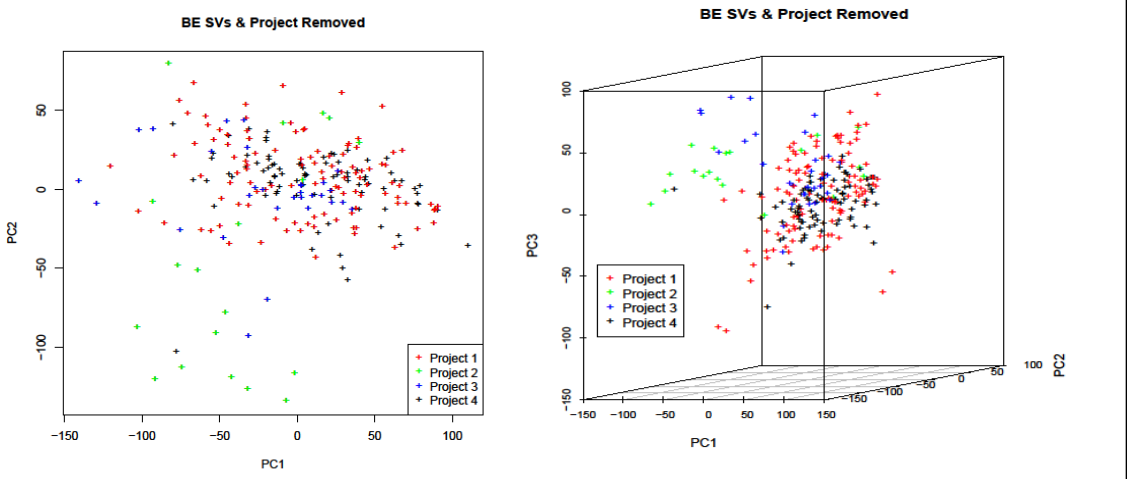

Figure 3: PCA analysis of normalized log counts (BE method) with plotting colors representing different projects or batches.

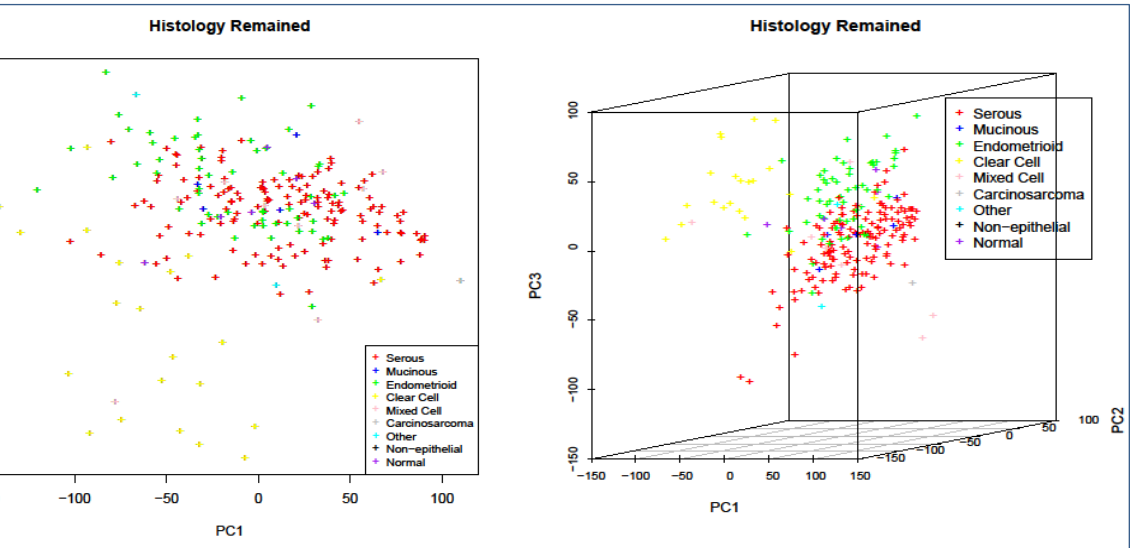

Figure 4: PCA analysis of normalized log counts (BE method) with plotting colors representing different histologies.

We also applied ComBat to the SV's identified by Leek method and visualized the remained histological effects in **Figure 5**. Contrast between **Figures 4** and **5** illustrates that BE method removes more unwanted noises that confounds with histology.

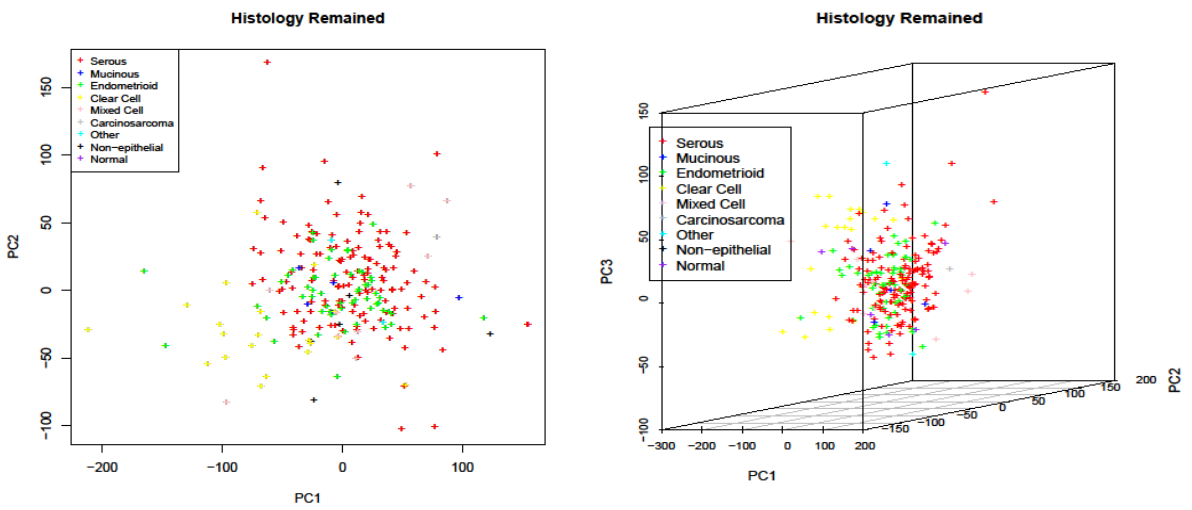

Figure 5: PCA analysis of normalized log counts (Leek method) with plotting colors representing different histologies.

C. Purity Score Estimation and Association with SVAs

It is necessary to confirm if tumor purity is confounding with histology in DE analysis. The tumor purity score for each sample can be computed based on raw and normalized data, respectively, using package ESTIMATE. The adjusted data from SVA/ComBat contains negative values, hence, must be shifted above 0 before computing purity score. We perform ANOVA test for the association between raw data purity score and histology for 186 DE samples, showing significant p-value (<0.001). The same test is performed on normalized data purity score and histology, showing no significance difference in purity scores across histologies (p = 0.36 in all tumor samples, p = 0.11 in 186 tumor samples included in this study) (**Figure 6**). We also check for association between raw data purity score and SV's by ANOVA (for categorical SV) and GLM (for continuous SV) tests. Results indicate that almost half SV's are significantly associated with tumor purity sore (**Table 2**). Therefore, we conclude that the normalization by SVA and ComBat successfully captures and removes tumor purity effect.

ANOVA on Combat Normalized Data: 186 Tumor Samples

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
histology	2	5014139.80	2507069.90	2.21	0.11
Residuals	183	207304213.73	1132809.91		

ANOVA on Raw Data: 186 Tumor Samples

- All Samples

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
histology	2	24738532.94	12369266.47	5.59	0.00
Residuals	183	404580487.58	2210822.34		

Figure 6: Association of Purity Scores Estimated Pre-normalization (raw data) and Post-normalization (Combat Normalized data).

Table 2: Tests for Association between Purity Score of Raw Data and SV's (BE method).

	SV1	SV2	SV3	SV4	SV5	SV6	SV7
Continuous SV	0.00551	0.00382	0.931	0.157	<2e-16	0.0276	0.726
Categorical SV	0.842	0.09473	0.1818	0.7682	<2.2e-16	0.0067	0.3793
	SV8	SV9	SV10	SV11	SV12	SV13	SV14
Continuous SV	0.0162	0.0429	0.0296	0.00045	0.772	0.539	0.208
Categorical SV	0.07185	0.0348	0.1326	0.00411	0.3479	0.9098	0.3276

D. Sensitivity Analysis for 10 genes: edgeR, ANOVA and Non-parametric Anova (Kruskal-Wallis Test)

We randomly pick 10 genes (non-negative expression so able to fit edgeR negative binomial model); plots of the 10 genes are below. For each gene, we fit 3 models to determine differentially expressed genes: ANOVA, Kruskal-Wallis Test, Negative-Binomial Model using edgeR. We also completed three models for testing difference in two histologies (not all three histologies) using non-parametric t-test (Wilcoxon test), parametric two sample t-test, and Negative Binomial Model using edge R. **Figure 7** shows the density for the 10 randomly selected genes. Results from the three models for the 10 genes are displayed in **Table 3**. QQ-plots showed deviation from normality assumption for the ANOVA tests, thus this model was removed as a possible candidate. Due to normalization of the data, measurements of gene expression levels for each gene no longer positive integer values, and since results similar between edgeR and non-parametric tests, we choose to use the conservative non-parametric tests for both testing

differential expression between three histologies, as well as, pairwise comparisons between histologies, as no distributional assumptions are required.

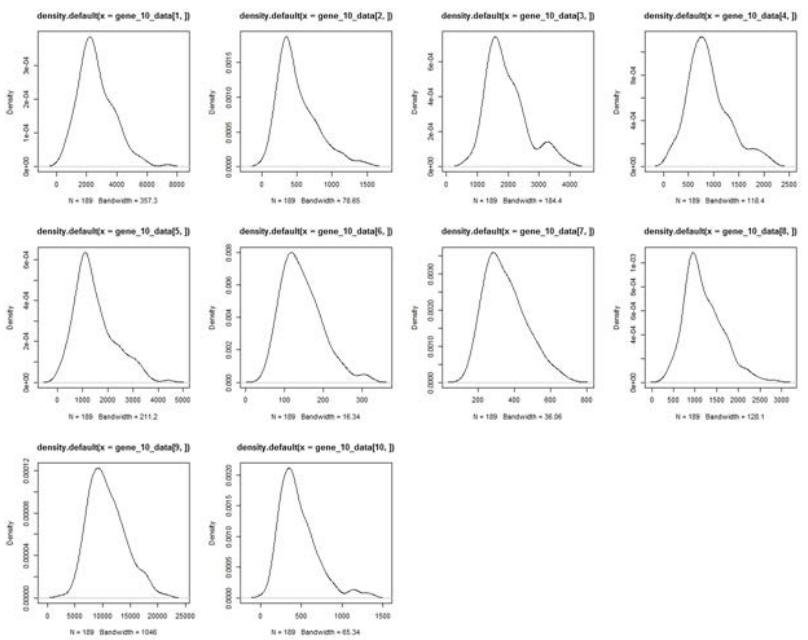

Figure 7: Density plots for 10 randomly selected genes for sensitivity analysis.

Table 3: Sensitivity Analysis Results for 10 randomly selected genes

Gene	3 Histology Comparison			Comparison of 2 histologies (EC vs CC)		
	ANOVA	Kruskal-Wallis Test	edgeR	2 sample T -test	Non-parametric T-test	edgeR
ENSG00000039650	5.13E-17	1.61E-11	3.15E-06	1.34E-05	1.80E-06	1.18E-04
ENSG00000106524	3.85E-01	2.93E-01	6.48E-06	6.31E-01	4.98E-01	5.64E-02
ENSG00000134899	3.99E-01	2.86E-01	6.29E-04	1.72E-01	1.15E-01	2.01E-01
ENSG00000137413	5.52E-01	7.30E-01	4.63E-02	4.79E-01	5.77E-01	2.78E-01
ENSG00000141858	8.26E-01	8.49E-01	1.02E-01	8.67E-01	7.06E-01	2.91E-01
ENSG00000144182	2.09E-01	3.17E-01	1.57E-01	1.66E-01	1.98E-01	5.06E-01
ENSG00000144713	7.37E-01	9.05E-01	2.37E-01	7.40E-01	9.09E-01	6.04E-01
ENSG00000145592	1.00E-03	3.30E-04	2.71E-01	4.78E-01	9.67E-01	7.01E-01
ENSG00000166886	4.87E-01	3.60E-01	7.11E-01	9.52E-01	9.57E-01	7.80E-01
ENSG00000279641	7.69E-02	7.04E-02	9.50E-01	2.53E-02	5.45E-02	9.59E-01