Supplementary Material for the paper:

"A New Statistic for Identifying Batch Effects in High-Throughput Genomic Data that uses Guided Principal Components Analysis"

Sarah E. Reese, Kellie J. Archer, Ph.D., Terry M. Therneau, Ph.D.,
Elizabeth J. Atkinson, M.S., Celine M. Vachon, Ph.D., Mariza de Andrade, Ph.D.,
Jean-Pierre A. Kocher, Ph.D., and Jeanette E. Eckel-Passow, Ph.D.

Contents

List of Figures

List of Tables

1 Filtering Sensitivity Analyses

The sensitivity of guided principal components analysis (gPCA) to different levels of variance filtering were investigated. Filtering using an ANOVA approach was also investigated. The main goal of filtering in our analyses is to remove non-informative features and to reduce the time required for the analysis.

1.1 Sensitivity of gPCA Results to Filtering

The GENEMAM and GENOA case study data were filtered using a variance filter to retain the 1000 most variable features. The sensitivity of the results of gPCA to this filtering was investigated using the GENEMAM data. Table S1 shows the resulting p-values from retaining between 10 (0.002% of GENEMAM features and 0.043% of GENOA features) and all features from the full (a) GENEMAM data set or (b) GENOA data set. For the GENEMAM data, as long as 500 (0.076%) or more features are retained, significant batch effects are found. Since filtering to retain 500 features takes approximately 1 minute to run, there is no need to retain fewer features. For the GENOA data, as long as 50 (0.216%) or more features are retained, significant batch effects are found, which takes approximately 8 seconds to run.

Data were also simulated as in our main simulation study, but with $p = 20000$ features and $n = 90$ samples. For each of the three phenotype scenarios (no phenotype, high variance phenotype, and low variance phenotype) data were simulated with batch and phenotype means $\mu_{b_1} = \mu_{p_1} = 0$ and $\mu_{b_2} = \mu_{p_2} = 1$. The batch variance was $\sigma_b = 0.5$, the phenotype variance for the high phenotypic variance scenario was $\sigma_p = 2$, and the phenotype variance for the low phenotypic variance scenario was $\sigma_p = 0.2$. The proportion of features affected by batch (`bprop`) was held constant at 0.004, 0.017, and 0.03 for the no, high variance, and low variance phenotype scenarios, respectively, each of which had good power in the previous simulation study. The proportion of features affected by phenotype in the high variance and low variance scenarios was 0.01. Table S2 shows the filtering sensitivity results. The columns represent the same information as for the case studies.

1.2 ANOVA Filtering

An analysis of variance (ANOVA) filter was applied to the GENEMAM data to assess it as an alternative to variance filtering. The `limma` package was used to fit an ANOVA (`lmFit()`) model with batch represented in the design matrix. The `eBayes()` function was subsequently used to compute the moderated F statistics and create an indicator of features with a significant batch effect to be used to filter the centered, mean-value imputed data. The methods of Benjamini and Hochberg (1995) and Bonferroni were used to adjust for multiple comparisons at significance levels of $\alpha = 0.05$ and 0.01. Table S3 shows the number of features retained from each adjustment method. In all cases the number of features is very large owing to the large batch effect present in this dataset. As shown in section 1.1, gPCA is not sensitive to filtering, so filtering can be used to reduce the data dimension and facilitate implementing gPCA by reducing the analysis time without worry.

2 gPCA Run Time Analysis

An analysis of the time it takes to run gPCA on varying sizes of data was performed. Table S1(a) gives the time it takes to run gPCA on the GENEMAM data with between 10 and the

Table S1: Case Study Variance Filtering Sensitivity Results: δ and corresponding p-values resulting from retaining between 10 and all features from the full data set. System Time gives the system time in minutes required to run gPCA as discussed in Section 2.

	Number Features	(%)	δ	p-value	System Time (min)
1	10	(0.002)	0.6878	0.511	0.812
2	20	(0.003)	0.5617	0.706	0.779
3	50	(0.008)	0.6129	0.119	0.841
4	100	(0.015)	0.4603	0.264	0.866
5	200	(0.03)	0.4194	0.268	0.892
6	500	(0.076)	0.5428	0.012	0.965
7	1000	(0.152)	0.5987	<0.001	1.144
8	2000	(0.304)	0.6914	<0.001	1.453
9	5000	(0.761)	0.7244	<0.001	2.479
10	10000	(1.521)	0.8344	<0.001	3.895
11	20000	(3.042)	0.9814	<0.001	7.620
12	50000	(7.606)	0.9807	<0.001	15.348
13	100000	(15.212)	0.9819	<0.001	33.395
14	200000	(30.424)	0.9835	<0.001	60.809
15	500000	(76.061)	0.9839	<0.001	162.075
16	657366	(100)	0.9839	<0.001	206.657

(a) GENEMAM

	Number Features	(%)	δ	p-value	System Time (min)
1	10	(0.043)	0.8664	0.087	0.118
2	20	(0.087)	0.8117	0.063	0.118
3	50	(0.216)	0.7693	0.025	0.129
4	100	(0.433)	0.7421	0.008	0.146
5	200	(0.865)	0.8315	<0.001	0.183
6	500	(2.163)	0.9220	<0.001	0.302
7	1000	(4.326)	0.9219	<0.001	0.520
8	2000	(8.652)	0.9006	<0.001	0.977
9	5000	(21.631)	0.8811	<0.001	2.394
10	10000	(43.262)	0.8620	0.006	4.052
11	20000	(86.524)	0.8338	0.012	8.051
12	23115	(100)	0.8282	0.013	9.388

(b) GENOA

Table S2: Simulation Variance Filtering Sensitivity Results: δ and corresponding p-values resulting from retaining between 10 and all features from the full data set. System Time gives the system time in minutes required to run gPCA as discussed in Section 2.

	Number Features	(%)	δ	p-value	System Time (min)
1	10	(0.05)	0.9991	<0.001	0.027
2	100	(0.5)	0.9976	<0.001	0.030
3	1000	(5)	0.9918	<0.001	0.048
4	2000	(10)	0.9896	<0.001	0.069
5	5000	(25)	0.9856	<0.001	0.133
6	10000	(50)	0.9839	<0.001	0.241
7	15000	(75)	0.9787	<0.001	0.350
8	20000	(100)	0.9795	<0.001	0.445

(a) No Phenotype

	Number Features	(%)	δ	p-value	System Time (min)
1	10	(0.05)	0.5571	0.673	0.029
2	100	(0.5)	0.4314	0.086	0.029
3	1000	(5)	0.4015	0.037	0.049
4	2000	(10)	0.4429	0.021	0.071
5	5000	(25)	0.4977	0.009	0.138
6	10000	(50)	0.5495	0.004	0.252
7	15000	(75)	0.5846	0.003	0.366
8	20000	(100)	0.6211	0.002	0.469

(b) High Variance Phenotype

	Number Features	(%)	δ	p-value	System Time (min)
1	10	(0.05)	0.4627	0.422	0.030
2	100	(0.5)	0.2484	0.425	0.034
3	1000	(5)	0.3164	0.203	0.060
4	2000	(10)	0.3304	0.198	0.090
5	5000	(25)	0.3615	0.091	0.182
6	10000	(50)	0.4008	0.027	0.341
7	15000	(75)	0.4364	0.010	0.503
8	20000	(100)	0.4632	0.003	0.625

(c) Low Variance Phenotype

Table S3: Number of features retained using an ANOVA filtering method with different multiple comparison adjustment methods and stringencies.

	Adj. Method	α	Feat. Retained
1	BH	0.05	636141
2	BH	0.01	624797
3	Bonferroni	0.05	546012
4	Bonferroni	0.01	535708

full set of features. There were $n = 614$ samples and the data were centered and mean-value imputation was performed prior to performing gPCA. Table S1(b) gives the time it takes to run gPCA on the GENOA data allowing the number of features retained after filtering to vary from 10 to the full dataset. There were $n = 703$ samples and the data were not centered prior to performing gPCA. There were no missing values so mean-value imputation was not necessary.

3 Batch Correction Sensitivity Analysis

A simulated dataset was chosen where batch and phenotypic effects are dependent. In this scenario each feature j for $j = 1, \ldots, p$ was assigned to have no phenotypic effect, a phenotypic effect only, a batch effect only, or both batch and phenotypic effects. For feature j, we let

$$f_j = \beta_p p_j \texttt{pheno} + \beta_b b_j \texttt{batch} + e$$

where p and b are length p vectors indicating whether each feature had a phenotypic or batch effect, respectively, `pheno` and `batch` are length n vectors giving the phenotype and batch effect for each sample, and $e \sim N(0, \sigma_b)$ is a random error term. The β_p and β_b parameters determine the magnitude of the phenotypic and batch effects, respectively.

The proportion of features effected by phenotype was `pprop` $= 0.1$, the variance was $\sigma_b = 0.5$. The batch and phenotype magnitude parameters were $\beta_b = 2$ and $\beta_p = 0.5$, respectively. The number of features with a phenotypic and batch effects was set at 50 for each effect, and the number of features with both a phenotypic and batch effect was set at 100. The method of Benjamini and Hochberg (1995) for adjusting for multiple testing was used at a significance level of $\alpha = 0.1$.

After fitting a linear model using the `lmFit()` function with phenotype as the predictor, the number of significant features in simulated data was assessed using the `eBayes()` function in the `limma` package both prior to batch correction and after batch correction using the batch mean-centering method of Sims *et al.* (2008). For batch correction, the `pamr.batchadjust()` function in the `pamr` package was used.

Contingency tables (Table S4) show features found to have a significant phenotypic effect pre- and post-batch correction using batch mean-centering via the `pamr` package in `R` versus features with known true phenotypic and batch effects. There were 50 features with a phenotypic effect, 50 features with a batch effect, and 100 features with both a phenotypic and batch effect. Optimally the inferential procedure should detect 150 features having a true phenotypic effect and should fail to reject 850 features with no phenotypic effect. This shows that batch correction allows features with a true phenotypic effect that is masked by

batch to be found significant after batch correction.

Table S4: Contingency tables from simulated data with dependent batch and phenotypic effects that show the number of features truly significant versus those found to be significant using `lmFit()` and `eBayes()` on (a) raw data and (b) batch corrected data using batch mean-centering (BMC). The rows of the tables indicate truth and the columns indicate the test results.

	Fail to reject	Reject
No Phenotype Effect	850	0
True Phenotype Effect	102	48

(a) Raw

	Fail to reject	Reject
No Phenotype Effect	849	1
True Phenotype Effect	2	148

(b) BMC Corrected

4 Additional Simulations

4.1 High Proportion of Features Affected by Batch

It is of interest to investigate the performance of gPCA when the proportion of features affected by batch is high. Simulations were assessed with batch proportion between 50 and 90% of features. Table S5 shows the estimated power is 100% for all scenarios so good results can be expected even when a large proportion of features are affected by batch.

4.2 Varied Batch Variance and Phenotypic Means Greater than Batch Means

The sensitivity of gPCA results to the level of batch variance was assessed through additional simulation analyses. For the no phenotype, high variance phenotype, and low variance phenotype scenarios, estimated power was calculated while varying the variance associated with batch between $\sigma_b = 0.5$ and 2. The proportion of features affected by batch (`bprop`) and the batch means were held constant at a level found to have good power when varying the batch proportion. For the true phenotype scenarios, the phenotype means were also varied as an assessment of gPCA when the phenotypic means ($\mu_{p_1} = 0$ and $\mu_{p_2} = 1.5$ or 2) are higher than the batch means ($\mu_b = 0$ and 1). Figure S1 shows the power plots for the three scenarios. We found that as batch variance increased, so did the estimated power. The smaller the difference in the phenotypic means, the higher the power. In the no phenotype scenario, we found that power decreased as the batch variance increased. This is attributable to the first principal component from unguided PCA and gPCA being similar when no phenotype is affecting the feature data, which is unlikely in application datasets.

5 Heatmaps of the GENEMAM data

The following heatmaps (Figure S2) show how expression levels vary by sample well location on the plates. Plate 3 stands out because of poor quality issues.

Table S5: Power for detecting batch effect as a function of the proportion of features that are affected by batch at 50 to 90% when no phenotypic, high variance phenotypic, or low variance phenotypic data were included in gPCA.

	σ_b	bprop	power
1	0.5	0.500	1
2	0.5	0.633	1
3	0.5	0.767	1
4	0.5	0.900	1
5	1.0	0.500	1
6	1.0	0.633	1
7	1.0	0.767	1
8	1.0	0.900	1

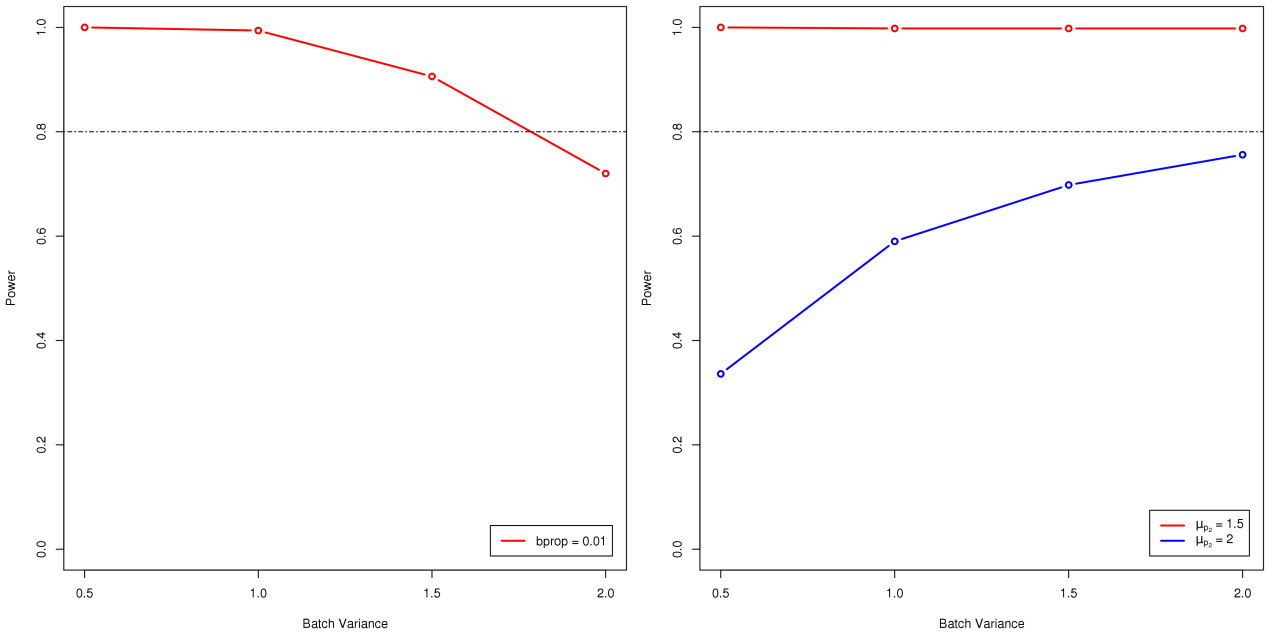
(a) No Phenotype

	σ_b	σ_p	pprop	bprop	power
1	0.5	2.0	0.01	0.500	1
2	0.5	2.0	0.01	0.633	1
3	0.5	2.0	0.01	0.767	1
4	0.5	2.0	0.01	0.900	1
5	1.0	2.0	0.01	0.500	1
6	1.0	2.0	0.01	0.633	1
7	1.0	2.0	0.01	0.767	1
8	1.0	2.0	0.01	0.900	1

(b) High Variance Phenotype

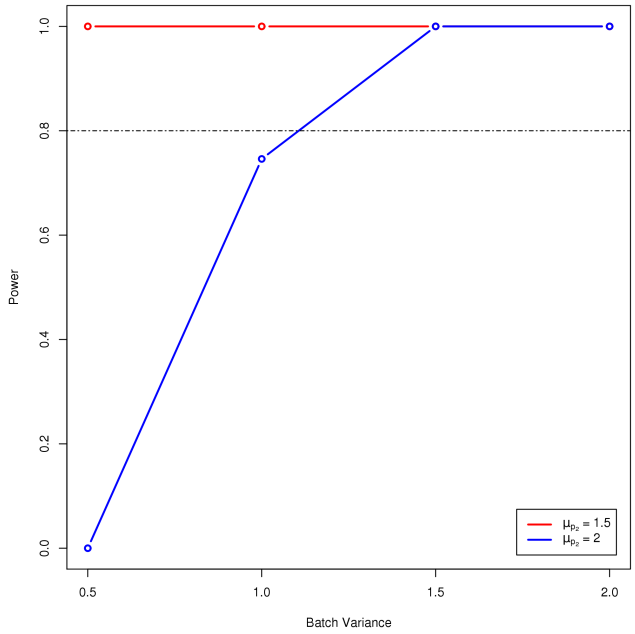
	σ_b	σ_p	pprop	bprop	power
1	0.5	0.2	0.05	0.500	1
2	0.5	0.2	0.05	0.633	1
3	0.5	0.2	0.05	0.767	1
4	0.5	0.2	0.05	0.900	1
5	0.5	0.2	0.10	0.500	1
6	0.5	0.2	0.10	0.633	1
7	0.5	0.2	0.10	0.767	1
8	0.5	0.2	0.10	0.900	1
9	1.0	0.2	0.05	0.500	1
10	1.0	0.2	0.05	0.633	1
11	1.0	0.2	0.05	0.767	1
12	1.0	0.2	0.05	0.900	1
13	1.0	0.2	0.10	0.500	1
14	1.0	0.2	0.10	0.633	1
15	1.0	0.2	0.10	0.767	1
16	1.0	0.2	0.10	0.900	1

(c) Low Variance Phenotype



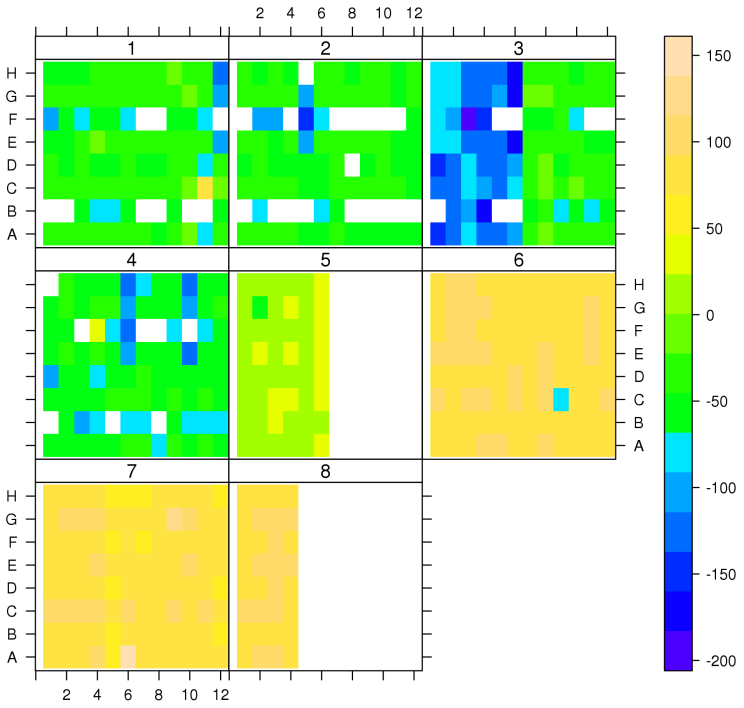
(a) No Phenotype

(b) High Variance Phenotype

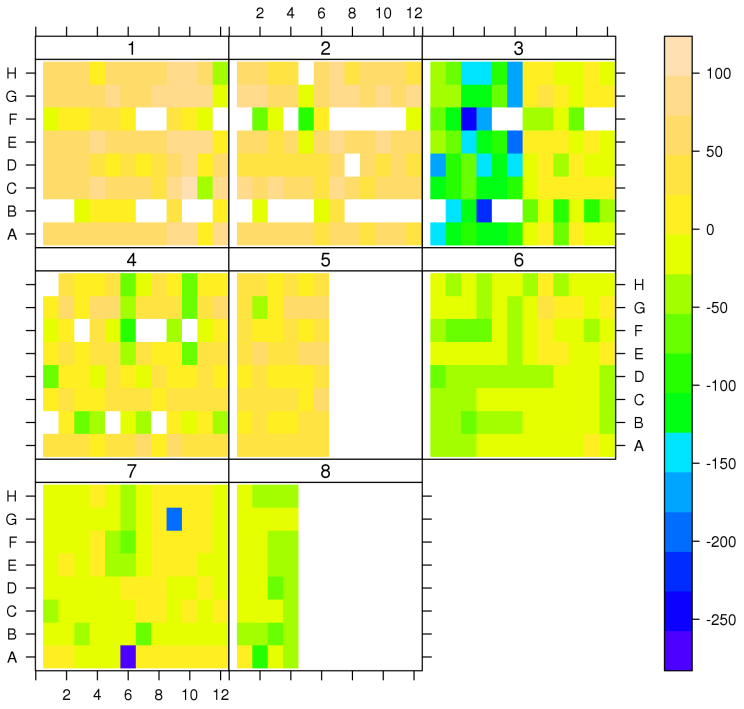


(c) Low Variance Phenotype

Figure S1: Power plots while varying the variance associated with batch and the phenotype means.



(a) PC_1



(b) PC_2

Figure S2: GENEMAM - Standardized Heatmaps showing the (a) PC_1 and (b) PC_2 values at each sample well location. White spaces indicate missing samples for the plate. Plates 5 and 8 were incomplete plates.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), pp. 289–300.

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., Miller, C. J., and Clarke, R. B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Medical Genomics*, **1**(42).