Removing batch effects

- Using the Remove Batch Effect tool
- Batch effects in PCA
- Batch effects in ANOVA results visualizations

By including *Batch* in the ANOVA model, the variability due to the batch effect is accounted for when calculating p-values for the non-random factors. In this sense, the batch effect has already been removed. However, visualizing biological effects can be very difficult if batch effects are present in the original intensity data used to generate visualizations. We can modify the original intensity data to remove the batch effect using the *Remove Batch Effect* tool.

Using the Remove Batch Effect tool

The *Remove Batch Effect* tool functions much like ANOVA in reverse, calculating the variation attributed to the factor being removed then adjusting the original intensity values to remove the variation. Once the variation caused by the batch effect has been removed, tools like PCA or clustering can be used to visualize what the data would look like if the batch effect was not present.

- Select the1 (Breast_Cancer.txt) spreadsheet
- Select Stat from the main tool bar
- Select Remove Batch Effect... (Figure 1)

🤣 Partek Genomics Suite - 1	(Breast_Car	ncer.txt)									-	×
File Edit Transform View	Stat Fil	ter Tools Windo	w Custom I	Help								
	Descri	ptive 🕨									Workflows Gene Expression	~
Analysis × Scatter Plot ×	Correl	ate 🔸	ces of Variation	×							Gene Expression	\times
	Param	etric Tests	Q	T	2						∽ Import	
	ANOV	A		-							Import Samples	
1 (Breast_Cancer.txt)	Remo	ve Batch Effect	GSM13097.txt							^	Add Sample Attributes	
ANOVA-3way (ANOV	/∆ Surviv Fisher	al Analysis 🔨	2. Treatment	3. Time	4. Batch	5. Treatment Combination	6. 100_g_at	7. 101_at	8. 102_at	9. 103	View Sample Information	
	Logisti	ic Regression	ct Control	0	A	Control-0	7.4733	6.52826	1.88753	2.58	Choose Sample ID Column	
	Multip	le Test Correction	ct Control	0	в	Control-0	8.18834	7.27984	3.74416	2.82		
	Power	Analysis	tt E2	8	A	E2-8	7.93369	6.87774	4.49185	5.24	DCA Seatter Blat	_
		4. GSM13138.b	tt E2	8	В	E2-8	8.09803	6.81121	3.12102	3.35	FCA Scatter Flot	~
		5. GSM13139.b	ct E2	48	A	E2-48	8.25644	6.80349	4.65535	3.01	Sample Box & Whiskers Chart	
		6. GSM13140.b	d E2	48	В	E2-48	7.61912	6.98982	3.87774	4.27	Sample Histogram	
		7. GSM15900.b	d E2+ICI	8	A	E2+ICI-8	7.80025	6.81634	3.53605	3.41	✓ Analysis	
		8. GSM15901.tx	d E2+ICI	8	В	E2+ICI-8	7.88692	7.0433	2.76553	2.53	Detect Differentially Expressed Genes	~
		9. GSM15902.tx	d E2+ICI	48	A	E2+ICI-48	8.01569	6.71699	4.34341	3.91	View Sources of Variation	
		10. GSM15903.b	d E2+ICI	48	В	E2+ICI-48	7.87897	6.77215	3.01792	3.95		—
		11. GSM15904.b	ct E2+Ral	8	A	E2+Ral-8	7.66035	6.5157	3.3505	3.20	Create Gene List	
		12. GSM15905.b	ct E2+Ral	8	В	E2+Ral-8	7.61324	6.77479	2.96347	2.45	> Visualization	
		13. GSM15906.b	ct E2+Ral	48	A	E2+Ral-48	7.79896	6.20555	4.15381	4.56	> Biological Interpretation	
		14. GSM15907.b	ct E2+Ral	48	В	E2+Ral-48	7.67596	6.84172	3.16993	2.76	> Genomic Integration	
		15. GSM15908.b	d E2+TOT	8	A	E2+TOT-8	8.12619	6.62352	4.32193	2.72	> miRNA Integration	
		16. GSM15909.tx	d E2+TOT	8	В	E2+TOT-8	8.03782	6.88874	1.53605	2.51		
		17. GSM15910.b	d E2+TOT	48	A	E2+TOT-48	8.24935	6.67525	4.03562	3.10		
		18. GSM15911.b	d E2+TOT	48	В	E2+TOT-48	7.72042	7.13031	3.85798	3.10		
	~	,										
<	>	Rows: 18 Columns	: 12631 <							> ~		
A												
w												

Figure 14. Invoking the Remove Batch Effect tool

The *Remove Batch Effects* dialog will open. The tool functions by performing an ANOVA then modifying the original intensities values to remove the effects of the specified factor(s).

- Select Treatment, Time, and Batch
- Select Add Factor > to add them to the ANOVA Factor(s) panel
- Select **Batch** in the ANOVA Factor(s) panel
- Select Add Factor > to add Batch to the Remove Effect(s) of These Factor(s) panel

By default, the results will be displayed in a new spreadsheet. Options to overwrite the current spreadsheet and specify the output file appear in the bottom of the dialog (Figure 2).

🥬 Remove Batch Effects					×
Experimental Factor(s) 2. Treatment 3. Time 4. Batch 5. Treatment Combination Save Model Load Model	AN Add Factor > Add Interaction > CREMOVE Factor	IOVA Factor(s) 2. Treatment 3. Time 4. Batch	R Add Factor > < Remove Factor	emove Effect(s) of Th 4. Batch	nese Factor(s)
Results Displayed in New Spreadsheet Current Spre Specify Output File C:/Partek The Specify Outpu	adsheet (Change Values) raining Data/Breast_Cancer-GE	E/batch-remove	ОК	Cancel	Browse Apply
Figure 15. Configuring the Remove Bai	Cn Effects tool to remove B	latch and create a new spreads	Sheet		

• Select OK

The new spreadsheet, 1-removeresult (batch-remove) will open in the Analysis tab (Figure 3).

													Workflows Gene Expression	
lysis X Scatter Plot X Sources	of V	ariati	on X	Sour	ces of Variatio	n×							Gene Expression	2
) 🧀 🔲 💽 🛛 🚅 I							2						└─ Import	
		· ·	_					, 					Import Samples	
1 (Breast_Cancer.txt)	$\sim -$	Curr	ent Se	election	GSM13097.txt			-				^	Add Sample Attributes	
ANOVA-3way (ANOVAResults			1. Filen	ame	2. Treatment	3. Time	4. Batch	5. Treatment Combination	6. 100 <u>g</u> at	7. 101_at	8. 102_at	9. 103	View Sample Information	
I-removeresult (batch-remove)		1.	GSM	13097.t	xt Control	0	A	Control-0	7.44022	6.6821	1.51407	2.36	Choose Sample ID Column	
		2.	GSM	13098.t	xt Control	0	в	Control-0	8.22142	7.126	4.11762	3.04		
		3.	GSM	13099.t	xt E2	8	Α	E2-8	7.90061	7.03158	4.11839	5.02		
		4.	GSM	13138.t	xt E2	8	в	E2-8	8.13111	6.65737	3.49448	3.57	PCA Scatter Plot	
		5.	GSM	13139.t	xt E2	48	Α	E2-48	8.22336	6.95733	4.28189	2.79	Sample Box & Whiskers Chart	
		6.	GSM	13140.t	xt E2	48	в	E2-48	7.6522	6.83598	4.2512	4.49	Sample Histogram	
		7.	GSM	15900.t	xt E2+ICI	8	A	E2+ICI-8	7.76717	6.97018	3.16259	3.19	✓ Analysis	
		8.	GSM	15901.t	et E2+ICI	8	в	E2+ICI-8	7.92	6.88946	3.13899	2.75	Detect Differentially Expressed Genes	
		9.	GSM	15902.t	xt E2+ICI	48	A	E2+ICI-48	7.98261	6.87083	3.96995	3.69	View Sources of Veriation	
		10.	GSM	15903.t	xt E2+ICI	48	в	E2+ICI-48	7.91205	6.61831	3.39138	4.17	view Sources of Variation	
		11.	GSM	15904.t	xt E2+Ral	8	Α	E2+Ral-8	7.62727	6.66954	2.97704	2.97	Create Gene List	
		12.	GSM	15905.t	xt E2+Ral	8	в	E2+Ral-8	7.64632	6.62095	3.33693	2.68	> Visualization	
		13.	GSM	15906.t	xt E2+Ral	48	A	E2+Ral-48	7.76588	6.35939	3.78035	4.34	> Biological Interpretation	
		14.	GSM	15907.t	xt E2+Ral	48	в	E2+Ral-48	7.70904	6.68788	3.54339	2.98	> Genomic Integration	
		15.	GSM	15908.t	xt E2+TOT	8	A	E2+TOT-8	8.09311	6.77736	3.94847	2.50	> miRNA Integration	
		16.	GSM	15909.t	xt E2+TOT	8	в	E2+TOT-8	8.0709	6.7349	1.90951	2.73		
		17.	GSM	15910.t	xt E2+TOT	48	Α	E2+TOT-48	8.21627	6.82909	3.66216	2.88		
		18.	GSM	15911.t	xt E2+TOT	48	в	E2+TOT-48	7.7535	6.97647	4.23144	3.32		
>		Row	s: 18	Columns	: 12631 🧹							> ~		

Figure 16. Viewing the new spreadsheet with batch effects removed

Batch effects in PCA

We can visualize the effects of removing the batch effects using PCA.

- Select 1 (Breast_Cancer.txt) from the spreadsheet tree
- Select (
- Select (
- Set Drawing Mode to Mixed
- Select the Ellipsoids tab
 Select Add Centroid

- Add Batch to the *Grouping Variable(s)* panel
 Set the colors of the two centroids as shown (Figure 4) to pink and yellow

🥵 Edit Centroid	×
Show Vectors 🔽 Label Centroids	
Categorical Variable(s) 2. Treatment 3. Time 5. Treatment Combination <-	Grouping Variable(s) 4. Batch
Ellipses/Ellipsoids to draw	Set All Clear All
	OK Cancel

Figure 17. Adding a centroid for Batch

- Select OK to close the Add Centroid ...
- Select OK to close the Configure Plot Properties dialog

The two centroids are distinct, showing the batch effect (Figure 5).



Figure 18. Viewing a batch effect using PCA. The batches are shown as the pink (A) and yellow (B) centroids. The clear separation of the centroids indicates a batch effect

• Repeat the above steps for 1-removeresult (batch-remove)

For 1-removeresult (batch-remove), the centroids of the two batches overlap, showing that the batch effect has been removed (Figure 6).



Figure 19. Overlapping centroids for batches A and B show that the batch effect has been removed.

Batch effects in ANOVA results visualizations

Visualization of ANOVA results for single probe(sets)/genes also benefits from batch removal. To illustrate this, we first need to repeat our ANOVA using the new *batch-remove* intesitiy values spreadsheet.

- Select the Analysis tab
- Select 1-removeresult (batch-remove) in the spreadsheet tree
- Select Stat from the main toolbar
- Select ANOVA...
- Add Treatment, Time, and Batch factors to the ANOVA Factor(s) panel
- Add Treatment * Time interaction to the ANOVA Factor(s) panel
- Select Contrasts...
- Select Treatment from the Select Factor Interaction drop-down menu
- Select Yes for Data is already log transformed?
- Set up contrasts of treatment vs. control for E2, E2+ICI, E2+Ral, and E2+TOT (Figure 7)

Configure of Spreadsheet	t 1-removeresult			Х
Data is already log transform	ed?	•		
Yes Base 2.0		• Fold change		•
Estimate Fr	ratio 🗌 T statistic	🗐 95% CI f	for Fold change	0
Select Factor/Interaction: Candidate Level(s) Control E2 E2+ICI E2+Ral E2+TOT	2. Treatment Label Add Contrast Level < Remove Contrast Level Label Add Contrast Level < Remove Contrast Level	> evel		~
	A	dd Contrast 🛛 ?	Add Combinat	ions ?
Contrast Name	Factor/Interaction	Status		Delete
E2 vs. Control E2+ICI vs. Control E2+Ral vs. Control E2+TOT vs. Control	2. Treatment 2. Treatment 2. Treatment 2. Treatment	Non-Estimable(p-val Non-Estimable(p-val Non-Estimable(p-val Non-Estimable(p-val	ue ratio {folc ue ratio {folc ue ratio {folc ue ratio {folc	
			OK	Cancel

Figure 20. Configuring ANOVA to comparing treatment groups to control

- Select **OK** to add contrasts
- Change output file name to ANOVAResults_batch-remove
- Select **OK** to perform the ANOVA

The ANOVAResults_batch-remove spreadsheet will open in the Analysis tab.

- Select the ANOVAResults spreadsheet
 Right-click on the row header for row 2, *TFF1*Select Dot Plot (Orig. Data) (Figure 8)

Partek Genomics Suite - 1/ANOVA-3way (ANOVAResults)								>	<
File Edit Transform View Stat Filter Tools Windo	w Custom Hel	p						w 18	_
								Workflows Gene Expression	$\tilde{}$
Analysis X Scatter Plot X Sources of Variation X Sour	ces of Variation X	Scatter Plot X	Scatter Plot	×				-Gene Expression	×
								✓ Import	
								Import Samples	
1 (Breast_Cancer.txt) Current Selection	1.59274e-07	-		-		-	· ·	Add Sample Attributes	
ANOVA-3way (ANOVAResults 1. 1. Column #	2. Probeset ID	3. Entrez Gene	4. Gene Symbol	Gene Title	6. RefSeq Transcript ID	p-value(Treatm ent)	8. p-value(Ti me)	View Sample Information	
ANOVA-3way (ANOVAResults								Choose Sample ID Column	
1. 4490	33799_at	6478	SIAH2	siah E3	NM_005067	1.59274e-07	0.243475	∨ QA/QC	
2. 2480	31708 at	7031	TFF1	trefoil factor 1	NM_003225	3.43187e-07	0.0096470	PCA Scatter Plot	~
3. Paste		8870	IER3	immediate early	/ NM_003897 ///	5.46804e-07	0.112467	Sample Box & Whiskers Chart	-
4. Eilter Inc	lude	219654	ZCCHC24	zinc finger,	NM_153367 ///	5.50565e-07	4.33981e-		
5. Filter Ex	dude	10057	ABCC5	ATP binding	NM_001023587	1.36059e-06	0.0028372	Sample Histogram	
6. Filter Inc	Filter Include (Orig. Data)	3487	IGFBP4	insulin like	NM_001552	1.42754e-06	0.275441	✓ Analysis	
7. Select (C	Select (Orig. Data)		KLF6	Kruppel-like	NM_001008490	1.73458e-06	6.14322e-	Detect Differentially Expressed Genes	~
8. Insert		3887	KRT81	keratin 81,	NM_002281 ///	2.05143e-06	1.84419e-	View Sources of Variation	~
9. Delete	Delete		TPBG	trophoblast	NM_001166392	3.07097e-06	0.0164127	Create Gene List	
10. HTML Re	port	51491	NOP 16	NOP16	NM_001256539	3.57845e-06	4.66809e-	Vieualization	
11. Dot Plot	(Orig. Data)	11031	RAB31	RAB31,	NM_006868	3.85637e-06	0.608732	Pielogical Interpretation	
12. XY Plot (Orig. Data) T	2288	FKBP4	FK506 binding	NM_002014 ///	4.30285e-06	0.113071		_
13. Bar Char	t (Orig. Data)	169611	OLFML2A	olfactomedin	NM_001282715	5.05105e-06	0.252829		
14. ANOVA	interaction Plot	10098	TSPAN5	tetraspanin 5	NM_005723 ///	5.80858e-06	0.110513	> miRNA Integration	_
15. Profile (Drig. Data)	54898	ELOVL2	ELOVL fatty	NM_017770 ///	6.11469e-06	0.0229098		
16. Probe St	t Details	10057	ABCC5	ATP binding	NM_001023587	7.66495e-06	0.862655		
17.		7168	TPM1	tropomyosin 1	NM_000366 ///	8.15471e-06	0.188266		
18. Browse	to Location	8651	SOCS1	suppressor of	NM_003745	8.6549e-06	0.334994		
19. Create L	50005 _at	10608	MXD4	MAX	NM_006454 ///	9.51684e-06	0.0834692		
v 20 119	213 at	4919	ROR 1	recentor	NM 001083592	1 25531e-05	0.0011116		
< Rows: 12626 Colu	mns: 40 <						> `	×	
]	

Figure 21. Invoking a dot plot from the ANOVAResults spreadsheet

A dot plot for trefoil factor 1 (TFF1) will open (Figure 9). The dot plot shows gene intensity values (y-axis) for each sample. Samples are grouped by *Treatm* ent.



Figure 22. Viewing the dot plot for trefoil factor 1 (TIFF1) across different treatment groups

To visualize the batch effect we will make a few changes to the plot.

- Select H/V to switch the horizontal and vertical axis
- Select (
- Set *Color* to **Batch**
- Set Size to Time
- Set Connect to Treatment Combination (Figure 10)

Plot Rendering Properties	×
Style Ellipsoids Labels Box&Whiskers Titles Color O Fixed ■ ● 4. Batch ✓ Manual	Axes Color Legend Text
Shape auto ~ Manual Outline Thin ~	Style direct V
Lighting Effects Fogging Transparency	Normal O Mixed
Save Load	OK Cancel

Figure 23. Configuring the dot plot (part 1 of 2)

- Select the Labels tab
 Select Column for *In Point Labels*Select Time from the *Column* drop-down list (Figure 11)

Plot Rendering Properties	×
<u>S</u> tyle <u>Ellipsoids</u> <u>Labels</u> Box& <u>W</u> hiskers <u>T</u> itles <u>A</u> xes <u>C</u> olor <u>L</u> egend Te <u>x</u> t	
Label Font Size 14	
Point Labels Points to Label Selected Label Offset 3 (a) Letters / Digits Imit to 1 (c) 1 (c) Label Format Imit to 1 (c) 1 (c) Label Format Imit to 1 (c) 1 (c)	
In Point Labels Letters / Digits Show all Limit to In Point Label None Row # Row Name Column In Point Label None degrees Rotate Point Labels 0 degrees	
Save Load OK Canc	:el

Figure 24. Configuring the dot plot (part 2 of 2)

• Select OK

The dot plot now clearly shows the batch effect (Figure 12). Samples within treatment groups are separated clearly between the two batches shown in blue and red.



Figure 25. Viewing a dot plot showing a batch effect. Each dot is a sample. The y-axis is treatment combinations; the x-axis is the expression value of the TFF1 gene. Dots are colored by batch, sized by time, connected by treatment combination, and labeled by time.

To view the effects of batch removal, we can view this dot plot for the ANOVAResults_batch-remove spreadsheet.

- Select the Analysis tab
- Select ANOVA-3way (ANOVAResults_batch-remove) from the spreadsheet tree
- Repeat the steps shown above to create the dot plot for trefoil factor 1

The dot plot invoked from the ANOVAResults_batch-remove) spreadsheet shows that the batch effect has been removed as all the samples no longer clearly separate by color within treatment groups (Figure 13).



Figure 26. Viewing the dot plot that shows batch effect removal. The plot configuration matches Figure 12.

« Detect differentially expressed genes with ANOVA Creating a gene list using the Venn Diagram »

Additional Assistance

If you need additional assistance, please visit our support page to submit a help ticket or find phone numbers for regional support.

Copyright © 2018 by Partek Incorporated. All Rights Reserved. Reproduction of this material without express written consent from Partek Incorporated is strictly prohibited.

