Detect differentially expressed genes with ANOVA

Analysis of variance (ANOVA) is a very powerful technique for identifying differentially expressed genes in a multi-factor experiment. In this data set, ANOVA will be used to generate a list of genes that are significantly differentially regulated by each treatment.

Adding factors and interactions

When setting up the ANOVA, the primary factors of interest, *Treatment* and *Time,* should be included. We will also include the interaction between *Treatment* and *Time, Treatment * Time,* because we are interested in whether different treatments behave differently over time. From our exploratory analysis using PCA, we also know that *Batch* is a major source of variation and needs to be included. Including *Batch* as a random factor will allow us to account for the batch effect.

- Select **Detect differentially expressed genes** from the *Analysis* section of the *Gene Expression* workflow
- Select **Treatment**, **Time**, and **Batch** in the *Experimental Factor(s)* panel
- Select **Add Factor >** to move the selections to the *ANOVA Factor(s)* panel
- Select both **Treatment** and **Time** in the *Experimental Factor(s)* panel by holding **Ctrl** on the keyboard while selecting each
- Select **Add Interaction >** to add the **Treatment * Time** interaction to the *ANOVA Factor(s)* panel (Figure 1)
- Do not select *OK* or *Apply.* We still need to add linear contrasts to the ANOVA model

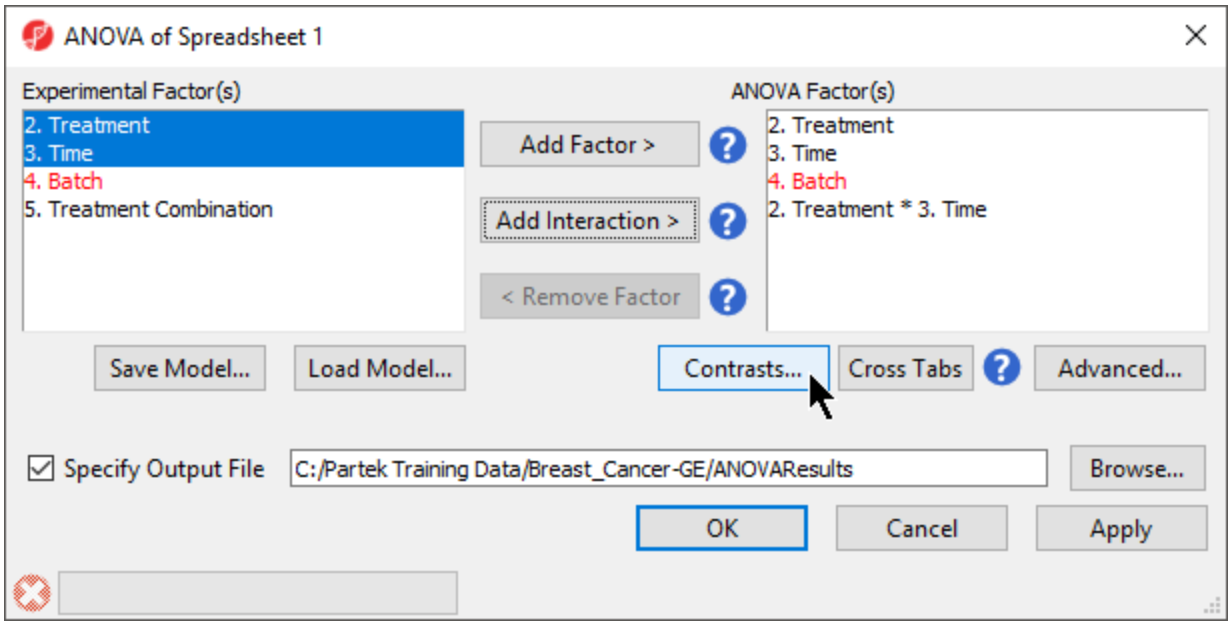


Figure 7. Adding factors and interactions to the ANOVA

Adding linear contrasts

ANOVA will output a p-value and F ratio for each factor or interaction; to get the fold-change and ratio between the different levels of a factor or interaction, linear contrasts, or comparisons, must be added.

- Select **Contrasts...** in the ANOVA dialog (Figure 1)
- Select **Yes** for *Data is already log transformed?*
- Select **Treatment * Time** from the *Select Factor/Interaction* drop-down menu

We will add contrasts comparing each of the three treatment groups to the control group.

- Select **E2 * 8** and **E2 * 48** from the *Candidate Level(s)* panel
- Select **Add Contrast Level >** to move them to the top panel (*Group 1*) on the right-hand side

The *Group 1* panel will be renamed after the contents of the panel. We can specify a name for the group.

- Set *Label* of the top panel to **E2**
- Select **Control * 0** from the *Candidate Level(s)* panel
- Select **Add Contrast Level >** to move it to the bottom panel (*Group 2*) on the right-hand side

- Set *Label* of the bottom panel to **Control**

The lower panel (*Group 2)* is considered the reference level. Because the data is $\log_2$ transformed, the geometric mean will be used to calculate the fold change and mean ratio to place both on a linear scale instead of a log scale.

- Select **Add Contrast** **(**Figure 2)

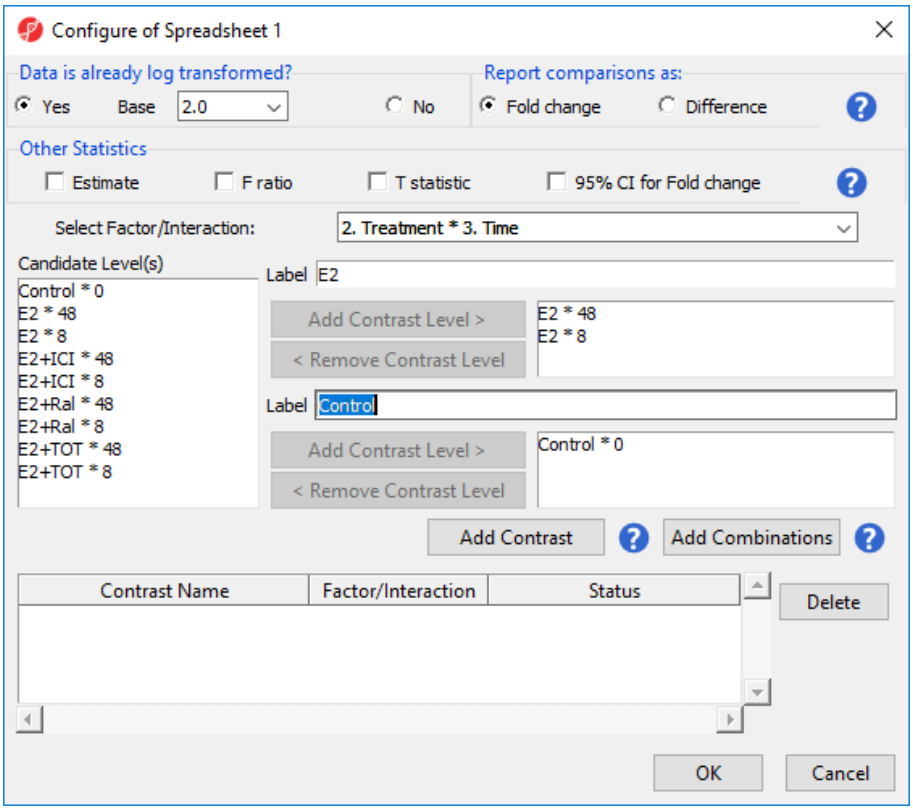


Figure 8. Adding a contrast between E2 vs. Control at all time points.

To examine the time points of each treatment condition separately, we can select *Add Combinations* instead of *Add Contrast*. This adds every possible contrast for the levels in the *Group 1* and *Group 2* panels.

- Select **E2 * 8** and **E2 * 48** from the *Candidate Level(s)* panel
- Select **Add Contrast Level >** to move them to the top panel (*Group 1*) on the right-hand side
- Select **Control * 0** from the *Candidate Level(s)* panel
- Select **Add Contrast Level >** to move it to the bottom panel (*Group 2*) on the right-hand side
- Select **Add Combinations** to add contrasts for *E2 * 8 vs. Control * 0* and *E2 * 48 vs. Control * 0* (Figure 3)

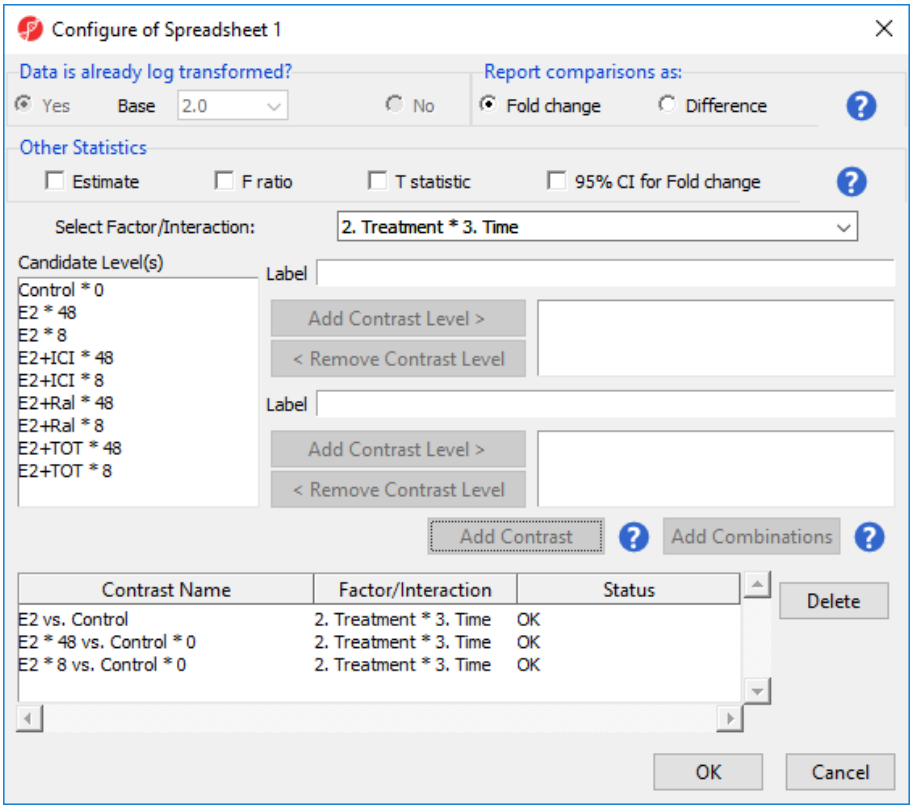


Figure 9. Add Combinations creates contrasts for every combination of levels from the two group panels.

For this tutorial, we will not be considering the time points of each treatment condition individually. We can remove the *E2 * 8 vs. Control * 0* and *E2 * 48 vs. Control * 0* contrasts.

- Select **E2 * 8 vs. Control * 0** and **E2 * 48 vs. Control * 0** from the contrasts list
- Select **Delete**

We will now add contrasts for the other treatment conditions.

- Add contrasts for *E2+ICI* vs. *Control*, *E2+Ral vs. Control,* and *E2+TOT vs. Control* following the steps outlined for *E2 vs. Control*

There should now be four contrasts added to the contrasts panel (Figure 4).

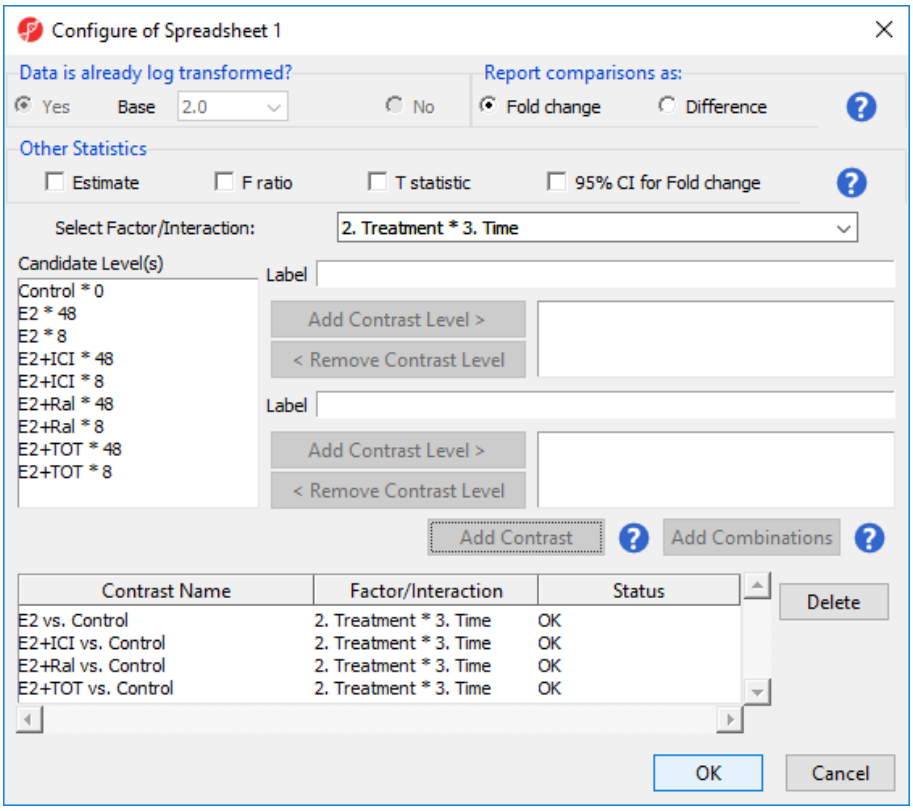


Figure 10. Fully configured contrasts for the tutorial

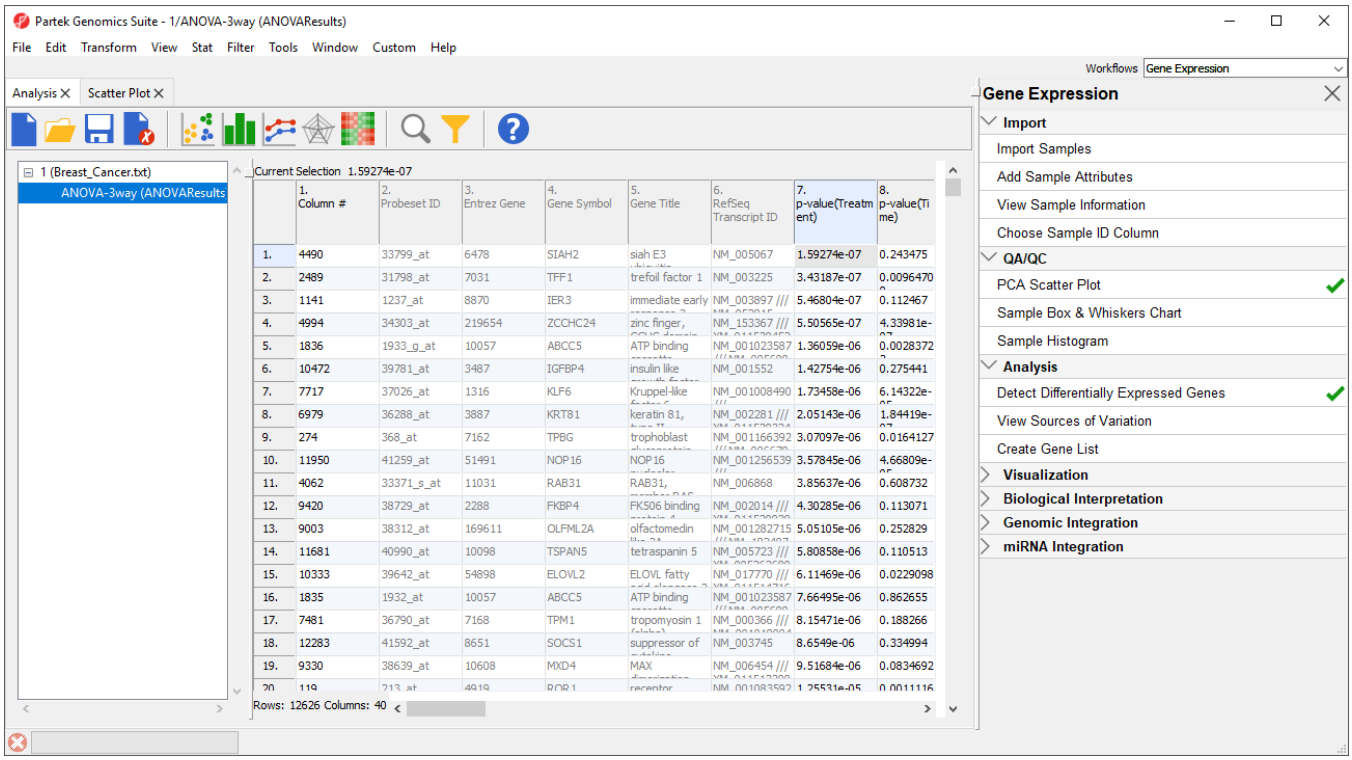
- Select **OK** to add the contrasts to the ANVOA model

The *Contrasts...* button should now read *Contrasts Included* in the ANOVA dialog.

- Select **OK** to perform the ANOVA

ANOVA results spreadsheet

The result of the 3-way mixed model ANOVA is displayed in a new spreadsheet, *ANOVA-3way (ANOVAResults)* that is a child of the *Breast_Cancer.txt* spr eadsheet. In *ANOVAResults,* each row represents a probe(set)/gene with the columns containing the results of the ANOVA (Figure 5).



	1. Column #	2. Probeset ID	3. Entrez Gene	4. Gene Symbol	5. Gene Title	6. RefSeq Transcript ID	7. p-value(Treatment)	8. p-value(Time)
1.	4490	33799_at	6478	SIAH2	siah E3	NM_005067	1.59274e-07	0.243475
2.	2489	31798_at	7031	TFF1	trefoil factor 1	NM_003225	3.43187e-07	0.0096470
3.	1141	1237_at	8870	IER3	immediate early	NM_003897 ///	5.46804e-07	0.112467
4.	4994	34303_at	219654	ZCCHC24	zinc finger,	NM_153367 ///	5.50565e-07	4.33981e-
5.	1836	1933_g_at	10057	ABCC5	ATP binding	NM_001023587	1.36059e-06	0.0028372
6.	10472	39781_at	3487	IGFBP4	insulin like	NM_001552	1.42754e-06	0.275441
7.	7717	37026_at	1316	KLF6	Kruppel-like	NM_001008490	1.73458e-06	6.14322e-
8.	6979	36288_at	3887	KRT81	keratin 81,	NM_002281 ///	2.05143e-06	1.84419e-
9.	274	368_at	7162	TPBG	trophoblast	NM_001166392	3.07097e-06	0.0164127
10.	11950	41259_at	51491	NOP16	NOP16	NM_001256539	3.57845e-06	4.66809e-
11.	4062	33371_s_at	11031	RAB31	RAB31,	NM_006868	3.85637e-06	0.608732
12.	9420	38729_at	2288	FKBP4	FK506 binding	NM_002014 ///	4.30285e-06	0.113071
13.	9003	38312_at	169611	OLFML2A	olfactomedin	NM_001282715	5.05105e-06	0.252829
14.	11681	40990_at	10098	TSPAN5	tetraspanin 5	NM_005723 ///	5.80858e-06	0.110513
15.	10333	39642_at	54898	ELOVL2	ELOVL fatty	NM_017770 ///	6.11469e-06	0.0229098
16.	1835	1932_at	10057	ABCC5	ATP binding	NM_001023587	7.66495e-06	0.862655
17.	7481	36790_at	7168	TPM1	tropomyosin 1	NM_000366 ///	8.15471e-06	0.188266
18.	12283	41592_at	8651	SOCS1	suppressor of	NM_003745	8.6549e-06	0.334994
19.	9330	38639_at	10608	MXD4	MAX	NM_006454 ///	9.51684e-06	0.0834692
20.	119	213_at	4919	ROR1	receptor	NM_001083592	1.25531e-05	0.0011116

Figure 11. Viewing the ANOVA Results spreadsheet. Probe(sets)/genes are on rows and the ANOVA results are on columns.

By default, the rows are sorted in acending order by the p-value of the first factor, which places the most significantly differentially expressed gene between different treatments at the top of the spreadsheet.

Each factor in the ANOVA adds p-value, F value, and SS value columns. F value is a ratio of signal to noise; high values indicate that the probe(set)/gene explains variation in the data set due to the factor. SS value is the sum of squares.

Each contrast in the ANOVA adds p-value, ratio, and fold-change columns. The p-value is calculated using log space. The ratio and fold change are calculated using linear space.

Viewing the sources of variation

Sources of variation captured in the ANOVA can be viewed for the entire data set or for individual probe(sets)/genes.

- Select **View Sources of Variation** from the *Analysis* section of the *Gene Expression* workflow

The Sources of Variation plot will open in a new tab (Figure 6).

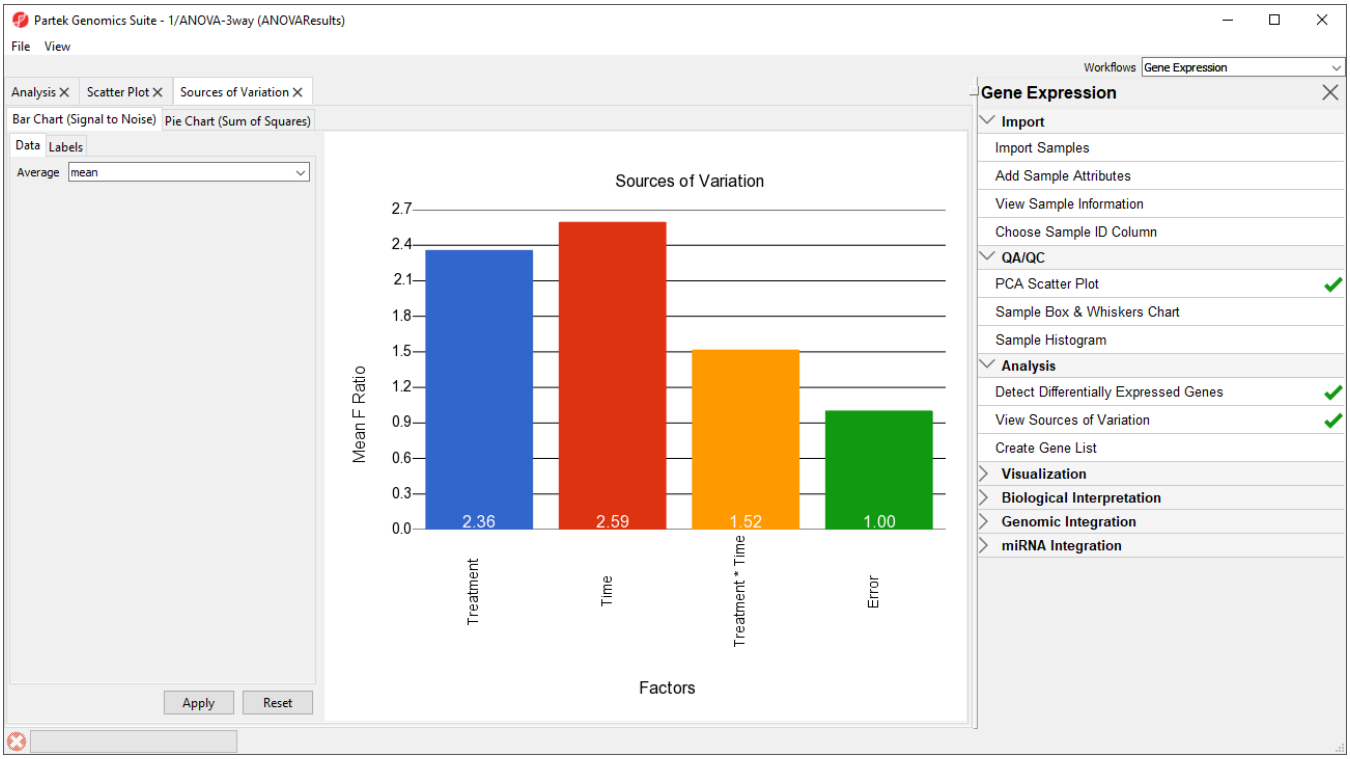


Figure 12. Viewing the sources of variation plot. Non-random factors are included when ANOVA is run using the default REML modle.

This plot presents the signal to noise ratio accross all probe(sets)/genes for each of the non-random factors and interactions in the ANOVA model. The y-axis represents the average mean square or F ratio, the ANOVA measure of variance, for all the probesets. Each bar is a factor and random error is also included. If the factor has a greater mean F ratio than Error, the factor contrinbuted significant variation to the data set.

Note that *Batch* is not included as a factor. This is beacuse *Batch* is a random factor and accounted for by the ANOVA model.

The sources of variation for each probe(set)/gene can be viewed individually.

- Right-click on a row header in the *ANOVAResults* spreadsheet
- Select **Sources of Variation** from the pop-up menu

The plot will open in a new tab. For additional plots that can be invoked from the ANOVA results spreadsheet, see the Visualizations user guide.

« Exploring the data set with PCA Removing batch effects »

Additional Assistance

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



Your Rating: Results: 41 rates