

GSA

GSA stands for gene specific analysis, the goal of which is to identify the statistical model that is the best for a specific gene among all the selected models, and then use that best model to calculate p-value and fold change.

- [GSA dialog](#)
- [GSA advanced options](#)
- [GSA report](#)
- [References](#)

GSA dialog

The first step of GSA is to choose which attributes to include in the test (Figure 1). All sample attributes including numeric and categorical attributes are displayed in the dialog, so use the check button to select between them. An experiment with two attributes *Cell type* (with groups A and B) and *Time* (time points 0, 5, 10) is used as an example in this section.

Choose which attributes to include in the statistical test

Categorical attributes

Cell type

☒

Time

☒

Back

Next

Figure 14. Choosing attributes to include in the statistical test by selecting the corresponding check button

Click **Next** to display the levels of each attribute to be selected for sub-group comparisons (contrasts).

To compare *A* vs. *B*, select **A** for *Cell type* on the top, **B** for *Cell type* on the bottom and click **Add comparison**. The specified comparison is added to the table below (Figure 2).

Cell type

☒ A

☐ B

Time

☐ 0

☐ 5

☐ 10

vs.

Cell type

☐ A

☒ B

Time

☐ 0

☐ 5

☐ 10

Add comparison

Figure 15. Specifying attribute levels for sub-group comparisons (contrast): Select A for Cell type on the top, B for Cell type on the bottom, and click Add comparison to compare A vs B

To compare *Time* point 5 vs. 0, select **5** for *Time* on the top, **0** for *Time* on the bottom, and click **Add comparison** (Figure 3).

<div>Cell type</div> <div> <input type="checkbox"/> A <input type="checkbox"/> B </div>	<div>Time</div> <div> <input type="checkbox"/> 0 <input checked="" type="checkbox"/> 5 <input type="checkbox"/> 10 </div>	5
vs.		
<div>Cell type</div> <div> <input type="checkbox"/> A <input type="checkbox"/> B </div>	<div>Time</div> <div> <input checked="" type="checkbox"/> 0 <input type="checkbox"/> 5 <input type="checkbox"/> 10 </div>	0
<div>Add comparison</div>		

Figure 16. Specifying attribute levels for sub-group comparisons (contrast): Select 5 for Time on the top, 0 for Time on the bottom, click Add comparison to compare 5 vs 0

To compare cell types at a certain time point, e.g. time point 5, select **A** and **5** on the top, and **B** and **5** on the bottom. Thereafter click **Add comparison** (Figure 4).

<div>Cell type</div> <div> <input checked="" type="checkbox"/> A <input type="checkbox"/> B </div>	<div>Time</div> <div> <input type="checkbox"/> 0 <input checked="" type="checkbox"/> 5 <input type="checkbox"/> 10 </div>	A and 5
vs.		
<div>Cell type</div> <div> <input type="checkbox"/> A <input checked="" type="checkbox"/> B </div>	<div>Time</div> <div> <input type="checkbox"/> 0 <input checked="" type="checkbox"/> 5 <input type="checkbox"/> 10 </div>	B and 5
<div>Add comparison</div>		

Figure 17. Specifying attribute levels for subgroup comparisons (contrast): Select A and 5 on the top, B and 5 on the bottom, click Add comparison to compare A*5 vs B*5

Multiple comparisons can be computed in one GSA run; Figure 5 shows the above three comparisons are added in the computation.

	Cell type	Time	vs.	Cell type	Time	
1	A		vs.	B		✗
2		5	vs.		0	✗
3	A	5	vs.	B	5	✗

Figure 18. Three comparisons included in GSA computation: A vs B; 5 vs 0; and A*5 vs B*5

In terms of design pool, i.e. choices of model designs to select from, two **2** factors in this example data will lead to seven possibilities in the design pool:

- *Cell type*
- *Time*
- *Cell type, Time*
- *Cell type, Cell type * Time*
- *Time, Cell type * Time*
- *Cell type * Time*
- *Cell type, Time, Cell type * Time*

In GSA, if a 2nd order interaction term is present in the design, then all first order terms must be present, which means, if *Cell type * Time* interaction is present, the two factors must be included in the model. In the other words, the following designs are not considered:

- *Cell type, Cell type * Time*
- *Time, Cell type * Time*
- *Cell type * Time*

If a comparison is added, some models that don't have the comparison factors will also be eliminated. E.g. if a comparison on *Cell type A vs. B* is added, only designs that have *Cell type* factor included will be in the computation. These are:

- *Cell type*
- *Cell type, Time*
- *Cell type, Time, Cell type * Time*

The more comparisons on different terms are added, the fewer models will be included in the computation. If the following comparisons are added in one GSA run:

- A vs B (*Cell type*)
- 5 vs 0 (*Time*)

only the following two models will be computed:

- *Cell type, Time*
- *Cell type, Time, Cell type * Time*

If comparisons on all the three terms are added in one GSA run:

- A vs B (*Cell type*)
- 5 vs 0 (*Time*)
- A*5 vs B*5 (*Cell type * Time*)

then only one model will be computed:

- *Cell type, Time, Cell type * Time*

If GSA is invoked from a quantification output data node directly, you will have the option to use the default normalization methods before performing differential expression detection (Figure 6).

- If invoked from a Partek E/M method output, the data node contains raw read counts and the default normalization is:
 - Normalize to total count (RPM)
 - Add 0.0001 (offset)
- If invoked from a Cufflinks method output, the data node contains FPKM and the default normalization is:
 - Add 0.0001 (offset)

Read count normalization

To customize normalization, run the normalization task prior to gene analysis

Apply normalization ⓘ
☒ Default
☐ None

Figure 19. Applying default normalization if differential gene detection dialog is invoked from a quantification output data node (see text for details)

If advanced normalization needs to be applied, perform the *Normalize counts* task on a quantification data node before doing differential expression detection (GSA or ANOVA).

GSA advanced options

Click on **Configure** to customize *Advanced options* (Figure 7).

Advanced options

Low value

Low-value filter ⓘ

☒ Lowest average coverage ⓘ 1.0
☐ Lowest maximum coverage ⓘ 1.0
☐ Lowest total coverage ⓘ 10.0
☐ None

Multiple test correction

Multiple test correction ⓘ
☒ FDR step-up
☐ Storey q-value
☐ Bonferroni
☐ None

Report option

Model selection criterion ⓘ

☒ AICc
☐ AIC

Enable multimodel approach ⓘ

☒ Yes
☐ No

Use only reliable estimation results ⓘ

☒ Yes
☐ No

Model types configuration ⓘ

☒ Default
☐ Custom

Apply
Save as new
Cancel

Figure 20. Configuring advanced GSA options

Low-expression feature

Low-expression feature section allows you to specify criteria to exclude features that do not meet requirements for the calculation. If there is filter feature task performed in the upstream analysis, the default of this filter is set to "None", otherwise, the default is Lowest average coverage is set to 1.

- *Lowest average coverage*: the computation will exclude a feature if its geometric mean across all samples is below than the specified value
- *Lowest maximum coverage*: the computation will exclude a feature if its maximum across all samples is below the specified value
- *Minimum coverage*: the computation will exclude a feature if its sum across all samples is below than the specified value

- *None*: include all features in the computation

Multiple test correction

Multiple test correction can be performed on the p-values of each comparison, with FDR step-up being the default (1). Other options like *Storey q-value* (2), and Bonferroni are provided, select one method at a time; None means no multiple test correct will be performed.

FDR step-up:

Suppose there are n p-values (n is the number of features). The p-values are sorted by ascending order and m represents the rank of a p-value. The calculation compares p-value*(n/m) with the specified alpha level, and the cut-off p-value is the one that generates the last product that is less than the alpha level. The goal of step up method is to find:

$$k^* = \max \left\{ m: P_m \leq \frac{m}{n} \cdot \alpha \right\} = \max \left\{ m: P_m \cdot \frac{n}{m} \leq \alpha \right\}$$

Define the step-up value as:

$$S_m = P_m \frac{n}{m}$$

Then an equivalent definition for K^* is :

$$k^* = \min \{ j: S_m > \alpha \text{ for all } m \text{ between } (j + 1) \text{ and } n \}$$

So when

$$S_m > \alpha \text{ and } S_{m-1} > S_m, \text{ then } S_{m-1} > \alpha$$

the step up value is

$$S_n^* = P_n$$

$$S_{m-1}^* = \min \{ S_{m-1}, S_m^* \}$$

In order to find K^* , start with S_n^* and then go up the list until you find the first step up value that is less or equal to alpha.

Storey q-value:

q-value is the minimum "positive false discovery rate" (pFDR) that can occur when rejecting a statistic.

For an observed statistic $T=t$ and nested set of rejection area $\{C\}$,

$$q - value(t) = \min_{\{C: t \in C\}} pFDR(C)$$

$$pFDR(C) = \frac{\pi_0 \cdot Prob(T \in C | H = 0)}{Prob(T \in C)} = Prob(H = 0 | T \in C)$$

Bonferroni:

Suppose there are n p-values (n is the number of features), the expected number of Type I errors would be given by $n^* \alpha$, thus the significance level of each individual test should be adjusted to α/n . Alternatively the p-values should be adjusted as $p_B = p^* n$, p_B is Bonferroni corrected p-value. If p_B is greater than 1, it is set to 1

Report option

This section configures how to select the best model for a feature. There are two options for *Model selection criterion*.

AICc (Akaike Information Criterion corrected) and AIC (Akaike Information Criterion). AICc is recommended for small sample size, while AIC is recommended for medium and large sample size What about large samples?(3). Note that when sample size grows from small to medium, AICc converges to AIC. Taking the AICc/AIC value into account, GSA considers the model with the lowest information criterion as the best choice.

In the results, the best model's Akaike weight is also generated. The model's weight is interpreted as the probability that the model would be picked as the best if the study were reproduced. The range of Akaike weight is between 0 to 1, where 1 means the best model is very superior to the other candidates from the model pool; if the best model's Akaike weight is close to 0.5 on the other hand, it means the best model is likely to be replaced by other candidates if the study were reproduced. One still uses the best shot model, however, the accuracy of the best shot is fairly low.

The default value for *Enable multimodel approach* is *Yes*. It means that the estimation will utilize all models in the pool by assigning weights to them based on AIC or AICc. If **No** is selected instead, the estimation is based on only one best model which has the smallest AIC or AICc. The output p-value will be different depending on the selected option for multimodel, but the fold change is the same. Multimodel approach is recommended when the best model's Akaike weight is not close to 1, meaning that the best model is not compelling.

There are situations when a model estimation procedure does not outright fail, but still encounters some difficulties. In this case, it can even generate p-value and fold change for the comparisons, but those values are not reliable, and can be misleading. It is recommended to use only reliable estimation results, so the default option for *Use only reliable estimation results* is set **Yes**.

Model types configuration

Partek® Flow® provides five response distribution types for each design model in the pool, namely:

- Normal
- Lognormal (the same as ANOVA task)
- Lognormal with shrinkage (the same as limma-trend method 4)
- Negative binomial
- Poisson

We recommend to use lognormal with shrinkage distribution (the default), and an experienced user may want to click on **Custom** to configure the model type and p-value type (Figure 8).

Figure 21. Five response distribution types for each design model

If multiple distribution types are selected, then the number of total models that is evaluated for each feature is the product of the number of design models and the number of distribution types. In the above example, suppose we have only compared A vs B in *Cell type* as in Figure 2, then the design model pool will have the following three models:

- *Cell type*
- *Cell type, Time*
- *Cell type, Time, Cell type * Time*

If we select *Lognormal with shrinkage* and *Negative binomial*, i.e. two distribution types, the best model fit for each feature will be selected from $3 * 2 = 6$ models using AIC or AICc.

The design pool can also be restricted by *Min error degrees of freedom*. When "Model types configuration" is set to *Default*, this is automated as follows: it is desirable to keep the error degrees of freedom at or above six. Therefore, we automatically set to the largest k , $0 \leq k \leq 6$ for which admissible models exist. Admissible model is one that can be estimated given the specified contrasts. In the above example, when we compare A vs B in *Cell type*, there are three possible design models. The error degree of freedom of model *Cell type* is largest and the error degree of freedom of model *Cell type, Time, Cell type * Time* is the smallest:

$$k(\text{Cell type}) > k(\text{Cell type, Time}) > k(\text{Cell type, Time, Cell type*Time})$$

If the sample size is big, $k \geq 6$ in all three models, all the models will be evaluated and the best model will be selected for each feature. However, if the sample size is too small, none of the models will have $k \geq 6$, then only the model with maximal k will be used in the calculation. If the maximal k happens to be zero, we are forced to use Poisson response distribution only.

There are two types of p-value, *F* and *Wald*, Poisson, negative binomial and normal models can generate p-value using either Wald or F statistics. Lognormal models always employ the F statistics; the more replicates in the study, the less the difference between the two options. When there are no replicates, only Poisson can be used to generate p-value using Wald.

Note: Partek Flow keeps tracking the log status of the data, and no matter whether GSA is performed on logged data or not, the LSMeans, ratio and fold change calculation are always in linear scale. Ratio is the ratio of the two LSMeans from the two groups in the comparison (left is the numerator, right is the denominator); Fold change is converted from ratio: when ratio is greater than 1, fold change is same as ratio; when ratio is less than one, fold change is -1/ratio. In other words - fold change value is always ≥ 1 or ≤ -1 , there is no fold change value between -1 and 1. When the LSmean of numerator group is greater than that of denominator group, fold change is greater than 1; when LSmean of numerator group is less than denominator group, fold change is less than 1; when the group groups are the same, fold change is 1. Logratio is ratio is log2 transformed, which is equivalent to logfoldchange is some other software.

GSA report

If there are multiple design models and multiple distribution types included in the calculation, the fraction of genes using each model and type will be displayed as pie charts in the task result (Figure 9).



Figure 22. Pie charts of proportion of genes using each model and distribution in gene-specific analysis calculation

Feature list with p-value and fold change generated from the best model selected is displayed in a table with other statistical information (Figure 10). By default, the gene list table is sorted by the first p-value column.

Optional columns

						A vs B						
	View	Gene ID	Transcript ID	Total reads	Average coverage	P-value	FDR step up	Ratio	Fold change	LSMean(A)	LSMean(B)	
1		PLXDC2	NM_001282736	9.62	1.07	3.37E-6	0.08	0.88	-1.13	1.01	1.14	
2		BATF2	NM_001300807	11.01	1.21	4.43E-6	0.08	0.72	-1.39	1.04	1.45	
3		WDR20	NM_001242417	12.24	1.31	5.42E-6	0.08	0.60	-1.67	1.05	1.75	
4		GGT6	NM_001122890	12.66	1.35	7E-6	0.08	0.59	-1.71	1.07	1.83	
5		AGAP2	NM_001122772	9.73	1.08	2.45E-5	0.22	0.85	-1.18	1.00	1.18	
6		THTPA	NM_001256322	10.06	1.11	3.8E-5	0.29	1.18	1.18	1.20	1.02	
7		ICA1L	NM_001288624	9.17	1.02	4.87E-5	0.31	0.96	-1.04	1.00	1.04	
8		SLC35B3	NM_015948	15.20	1.59	5.86E-5	0.31	0.51	-1.96	1.18	2.32	
9		RTFDC1	NM_001283036	10.69	1.18	6.23E-5	0.31	0.77	-1.30	1.05	1.36	
10		TOR2A	NM_001252023	9.17	1.02	6.73E-5	0.31	0.97	-1.03	1.00	1.04	
11		INSRR	NM_014215	10.75	1.19	1.2E-4	0.45	0.86	-1.17	1.11	1.30	
12		HIGD1A	NM_001099668	15.49	1.62	1.23E-4	0.45	0.53	-1.90	1.23	2.34	
13		MRPL22	NM_001014990	10.05	1.11	1.27E-4	0.45	0.83	-1.21	1.02	1.23	
14		MFAP3L	NM_001009554	9.80	1.08	1.55E-4	0.49	0.85	-1.18	1.01	1.19	
15		AGER	NM_001206929	9.36	1.04	1.67E-4	0.49	0.92	-1.09	1.00	1.09	
16		MFAP3L	NM_001301647	9.43	1.05	1.89E-4	0.49	0.91	-1.10	1.00	1.10	
17		PCDH20	NM_022843	9.22	1.02	1.96E-4	0.49	1.04	1.04	1.04	1.00	
18		RRP12	NM_015179	345.58	37.83	1.97E-4	0.49	0.72	-1.39	32.79	45.41	
19		CHRM3-AS2	NR_103776	9.61	1.07	2.04E-4	0.49	1.09	1.09	1.11	1.02	
20		GALNT18	NM_198516	9.33	1.04	2.76E-4	0.63	0.93	-1.07	1.00	1.08	
21		MECP2	NM_001316337	10.73	1.17	3.8E-4	0.83	0.70	-1.43	1.00	1.43	
22		MRPL18	NR_134864	13.05	1.42	4.21E-4	0.84	0.69	-1.44	1.21	1.75	
23		BAG6	NM_001199697	34.20	3.49	4.87E-4	0.84	0.48	-2.09	2.56	5.35	
24		BAG6	NM_004639.2	104.27	11.14	5.05E-4	0.84	0.60	-1.66	8.97	14.86	
25		ACTA2	NM_001613	9.65	1.07	5.52E-4	0.84	0.89	-1.12	1.02	1.14	

Rows per page 25
(1 of 1830)
Download


Figure 23. Feature list on the gene-specific analysis result. Clicking on the column header sorts the table. Panel on the left filters the table


The following information is included in the table by default:

- Feature ID information: if transcript level analysis was performed, and the annotation file has both transcript and gene level information, both gene ID and transcript ID are displayed. Otherwise, the table shows only the available information.
- Total counts: total number of read counts across all the observations from the input data.
- Each contrast outputs p-value, FDR step up p-value, ratio and fold change in linear scale, LSmean of each group comparison in linear scale

When you click on the Optional columns link on the top-left corner of the table, extra information will be displayed in the table when select:


- Maximum count: maximum number of reads counts across all the observations from the input data.
- Geometric mean: geometric mean value of the input counts across all observations.
- Arithmetic mean: arithmetic mean value of input counts across all observations.

Click on **View extra details report** () icon under View section to get more statistical information about the feature. In a case that the task doesn't fail, but certain statistical information is not generated, e.g. p-value and/or fold change of a certain comparison are not generated for some or all feature, click on this icon to get more information by mousing over the read exclamation icon


Lognormal with shrinkage AICc ?

Model estimation failed. Estimation of shrinkage hyperparameters failed. Please try increasing the threshold in Low expression filter.

By clicking on **Optional columns**, you can retrieve more statistics result information, e.g. Average coverage which is the geometric mean of normalized reads in linear scale across all the samples; fold change lower/upper limits generated from 95% confidence interval; feature annotation information if there are any more annotation fields in the annotation model you specified for quantification, like genomic location, strand information etc.

On the right of each contrast header, there is volcano plot icon (). Select it to display the volcano plot on the chosen contrast (Figure 11).

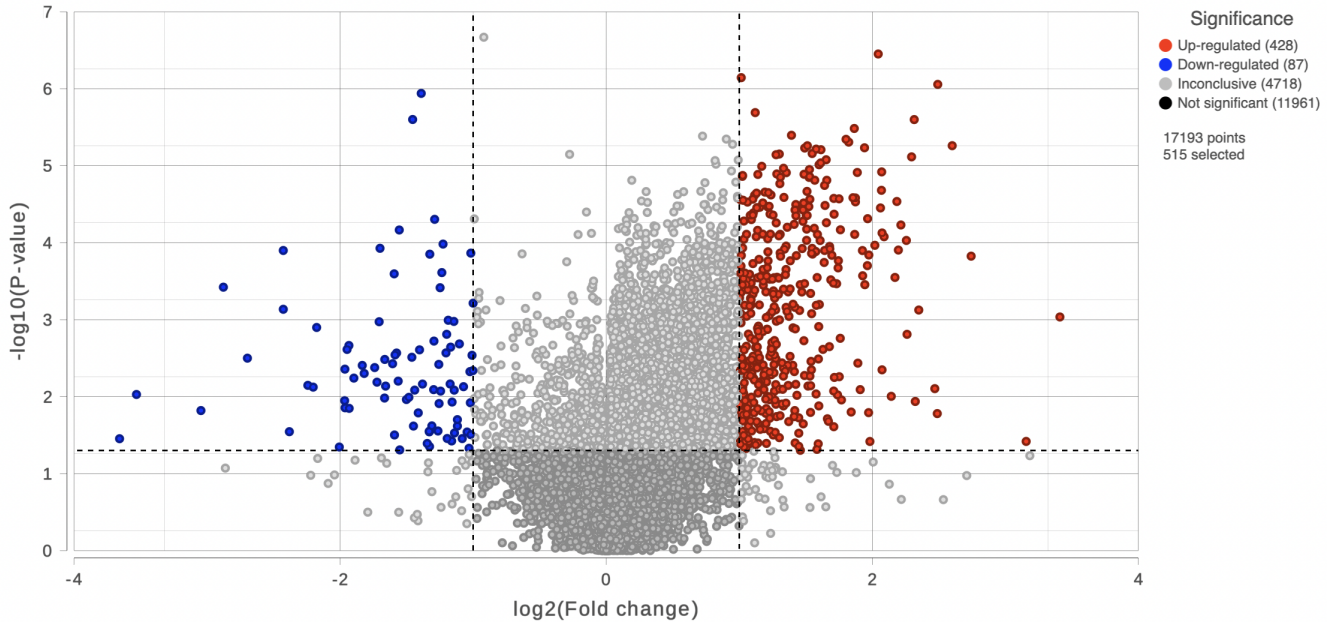



Figure 24. Volcano plot in comparison A vs B. X-axis represents fold change (linear scale), Y-axis represents negative logged p-value (unadjusted), each dc change of -2 and 2. Lower left corner displays number of features passing the fold-change and p-value criteria

Feature list filter panel is on the left of the table (Figure 12). Click on the **black triangle** () to collapse and expand the panel.

Select the **check box** of the field and specify the cutoff by typing directly or using the slider. Press **Enter** to apply. After the filter has been applied, the total number of included features will be updated on the top of the panel (*Result*).

Note that for the LSMean, there are two columns corresponding to the two factors in the contrast. The cutoffs can be specified for the left and right columns separately. For example, in Figure 6, the LSMean (left) corresponds to A while the LSMean(right) is for the B.

Results: 45741

Filter

☐ Gene ID

☐ Transcript ID

☐ Total reads

☐ P-value

☒ FDR step up

Less than or equ ▾

1

01

☐ Ratio

☒ Fold change

From

-1

to

1

☒ Exclude range

☐ LSMean

☐ Low expressed

Save filter

Clear filter

Saved filters

⚙️ ▾

(No saved filters available)

🔍

Generate filtered node

Figure 25. Feature list filter panel

The filtered result can be saved into a filtered data node by selecting the **Generate list** button at the lower-left corner of the table (

? Unknown Attachment

). Selecting the **Download** button at the lower-right corner of the table downloads the table as a text file to the local computer.

If lognormal with shrinkage method was selected for GSA, a shrinkage plot is generated in the report (Figure 13). X-axis shows the log2 value of average coverage. The plot helps to determine the threshold of low expression features. If there is an increase before a monotone decrease trend on the left side of the plot, you need to set a higher threshold on the low expression filter. Detailed information on how to set the threshold can be found in the [GSA white paper](#).

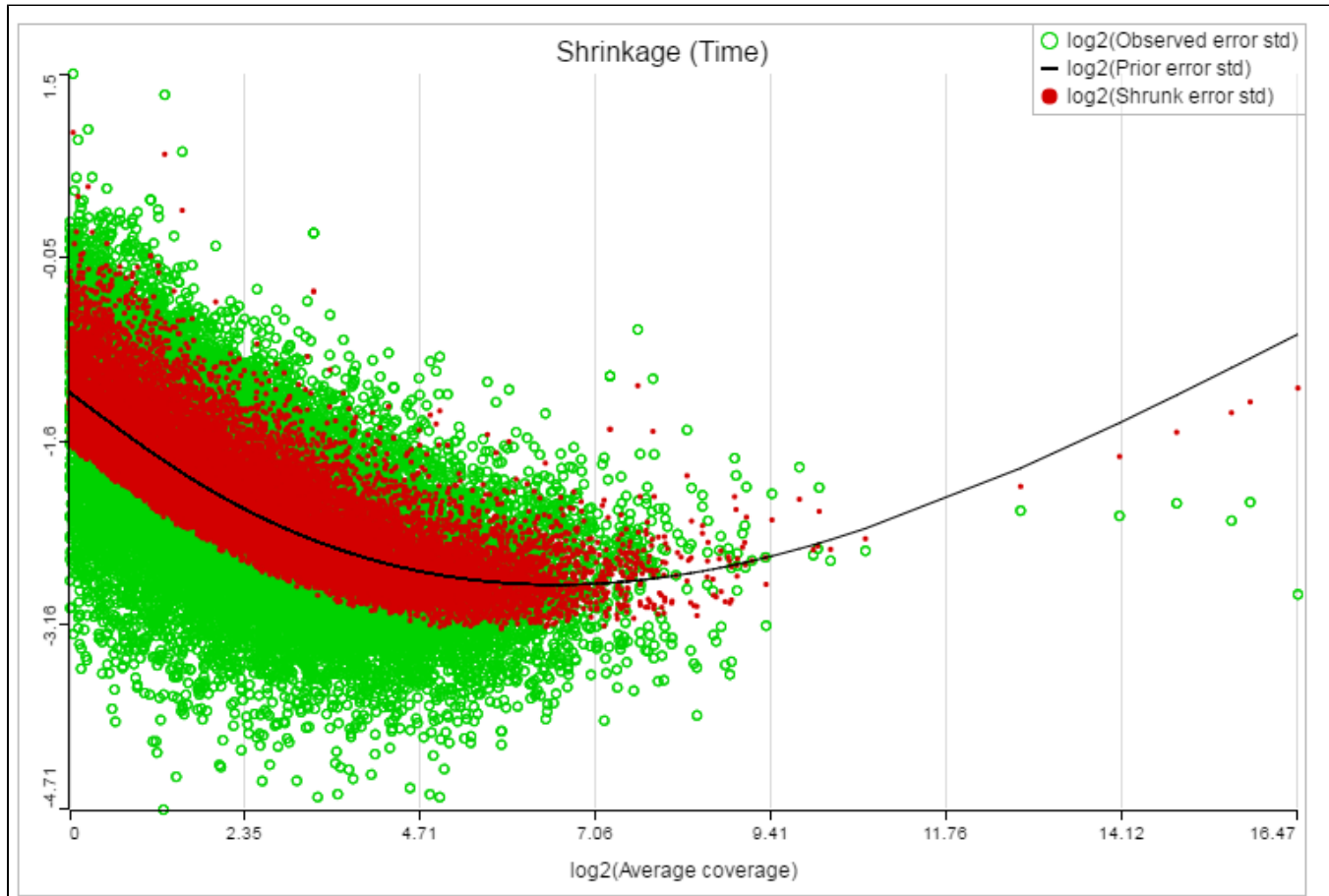


Figure 26. Shrinkage plot generated on lognormal with shrinkage model. X-axis is represents average coverage in log2 scale; Y-axis represents log2 standard deviation of error term. Green dot represents standard deviation of residual error obtained from lognormal linear model on a gene; black line represents the trend how the errors change depending on the average gene expression; red dot represents adjusted (shrunk) standard deviation of error on a gene

References

1. Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. JRSS, B, 57, 289-300.
2. Storey JD. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Annals of Statistics, 31: 2013-2035.
3. Auer, 2011, A two-stage Poisson model for testing RNA-Seq
4. Burnham, Anderson, 2010, Model selection and multimodel inference
5. Law C, Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology, 2014 15:R29.
6. <http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html#cuffdiff-output-files>
7. Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biology, 2010

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:  43 rates