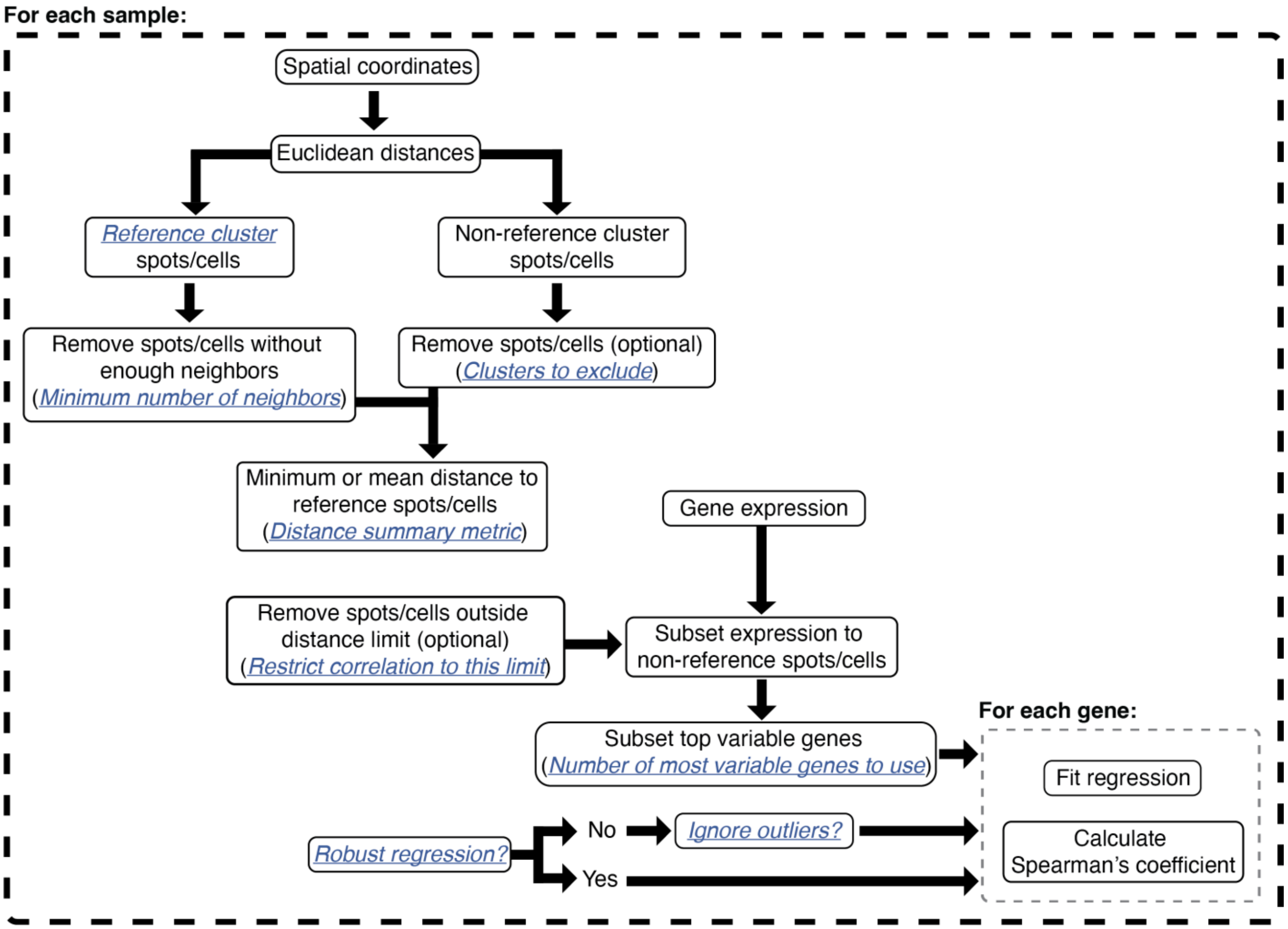
Module: Spatial gradients

spatialGE documentation

The availability of spatial transcriptomics (ST) technologies has further enabled the study of gene expression at the interface of domains. Researchers studying gene expression at domain-domain interfaces often aim to elucidate interactions between those domains. These interactions are of special interest for the identification of biomarkers of drug targets, for example, genes involved in tumor cell proliferation.

The traditional approach to study domain-domain interface gene expression involves clustering followed by differential gene expression analysis of the cluster identified as the “interface cluster”. While this approach might yield reasonable results, it involves the somewhat biased identification of the interface cluster. In addition, depending on tissue characteristics, some clustering algorithms might fail to predict that cluster. In spatialGE, the **Spatial gradients** module has been designed as an alternative to the study of gene expression at domain-domain interfaces. The module runs the STgradient function from the spatialGE R package in the background. The STgradient function still requires identification of domains, however it does not require prediction of a “interface cluster”. Instead, the STgradient algorithm tests for correlations between gene expression of the spots/cells in a domain and the distance to a *reference domain*. For example, researchers could use the **Spatial gradients** module to identify genes with high expression when closer to the white matter region of an ST sample, or genes that exhibit low expression as spots/cells are closer to the tumor domain of a cancer ST sample. The diagram below illustrates the STgradient algorithm:

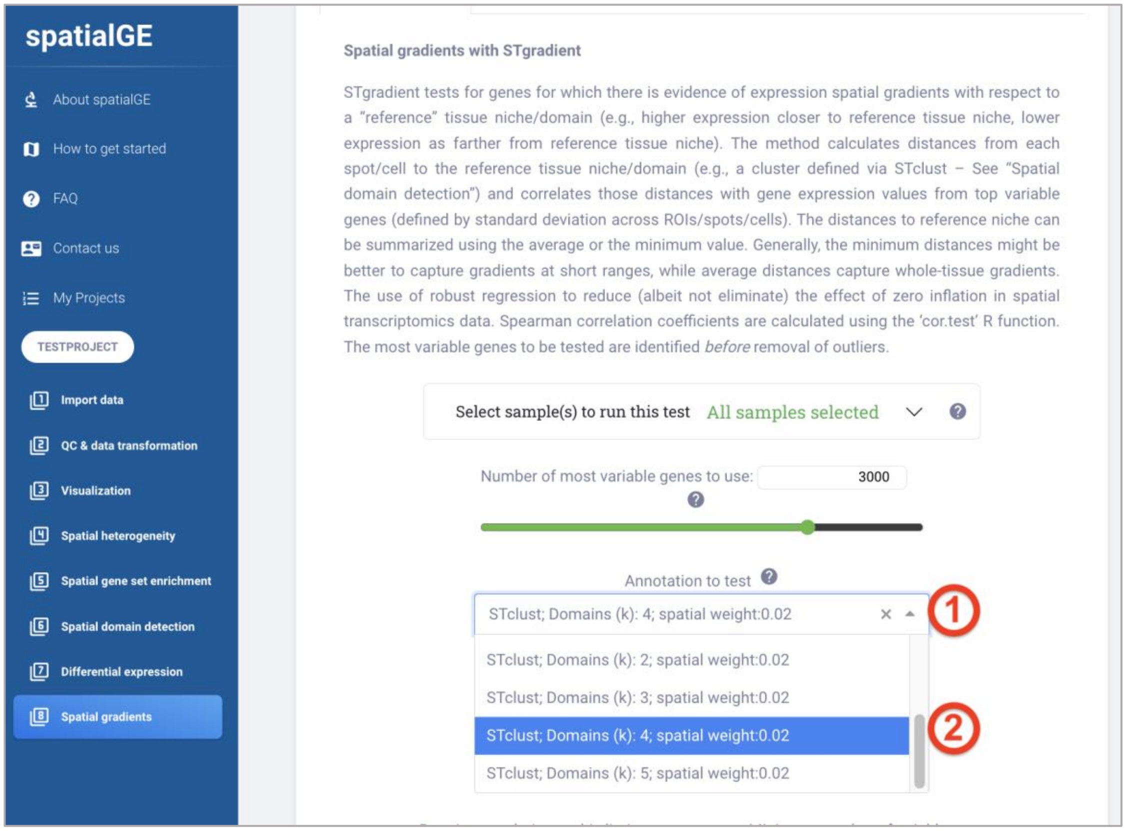

For each sample:
Spatial coordinates
Euclidean distances
Reference cluster spots/cells
Non-reference cluster spots/cells
Remove spots/cells without enough neighbors (Minimum number of neighbors)
Remove spots/cells (optional) (Clusters to exclude)
Minimum or mean distance to reference spots/cells (Distance summary metric)
Gene expression
Remove spots/cells outside distance limit (optional) (Restrict correlation to this limit)
Subset expression to non-reference spots/cells
Subset top variable genes (Number of most variable genes to use)
For each gene:
Fit regression
Robust regression?
No
Ignore outliers?
Yes
Calculate Spearman’s coefficient

In the STgradient algorithm, the spatial coordinates are used to calculate Euclidean distances of each spot/cell to each of the other spots/cells. The domain assignments obtained from the **Spatial domain detection** module are used to classify spots/cells into reference and non-reference domains. The *reference domain* is the domain from which the distances will be calculated. The *non-reference domain* is the domain on which the user wants to study gradients. Suppose that the user wants to identify genes in the stromal domain in which expression changes with respect to the tumor domain. In this case, the *reference domain* is the tumor compartment, and the *non-reference domain* is the stromal compartment.

Next, the STgradient algorithm removes isolated spots/cells in the reference domain (i.e., have a number of immediate neighbors smaller than parameter **Minimum number of neighbors**). The **Minimum number of neighbors** parameter intends to reduce the effects of small "pockets" of the reference domain on the correlation coefficient. The user also can set STgradient to remove non-reference spots/cells assigned to another domain. For example, an user could remove from the calculations those spots/cells assigned to immune-infiltrated areas embedded in the stromal domain, in case such immune-infiltrated domain was identified by the **Spatial domain detection** module. At this point, either the minimum or average distance of each non-reference spot/cell to each reference spot is calculated (**Distance summary metric**). The minimum distance is easier to interpret, however, it is more sensitive to highly fragmented reference domains. On the other hand, the interpretability of the average distance is not as easy as the minimum distance, but it might detect in a better way the gradients when the reference domain is distributed across the sample. It is also possible to exclude non-reference spots that fall outside a limit of minimum/average distance (**Restrict correlation to this limit**). While it is true that gradients can be observed within certain distances from the reference domain, this parameter should be used with caution. Interpretation of the excluded spots/cells is challenging given that the current implementation of the algorithm does not convert the Euclidean distances into microns. If the user thinks that correlations should be calculated within a specific distance range, then it is suggested to run the **Spatial gradients** module without restriction and then multiple times with different values in the **Restrict correlation to this limit** argument.

The gene expression data is subset to the non-reference spots/cells, and the top variable genes (**Number of most variable genes to use**) are selected. Results will be produced only for those genes. Depending on the user's selection, the ordinary least squares (OLS) or robust regression is used to calculate the slope of the regression line between the minimum/average distance and the expression of a gene. If OLS was selected, users might opt to remove outliers via the interquartile method. Finally, the Spearman's coefficient is calculated in addition to the regression line coefficient. The slope indicates the direction of the correlation. If positive, the gene expression tends to be higher when farther from the reference domain. If negative, gene expression is higher when closer to the reference domain. The interpretation of the sign for the Spearman's coefficient is similar, however its magnitude indicates how strong the correlation is between gene expression and distance to the reference domain.

To start an analysis with the **Spatial gradients** module, an analysis with the **Spatial domain detection** module should be completed beforehand. The **Spatial gradients** analysis can be run on all samples or a subset by de-selecting the sample names under the **Select sample(s) to run this test** drop-down. To facilitate comparative analysis, most users might want to run the analysis on all samples. For this tutorial, all samples will be analyzed. Next, the **Number of most variable genes to use** slider can be moved to specify how many genes to test. The default (3000 genes) is a good starting place and will be kept for this tutorial. In the **Annotation to test** dropdown, one of the clustering solutions generated by the **Spatial domain detection** module is selected. For consistency with the tutorial of **Differential expression** module, the K=4 solution with spatial weight=0.02 will be selected. Please click the **Annotation to test** drop-down, followed by the "*STclust; Domains (k): 4; spatial weight:0.02*" option.



Spatial gradients with STgradient

STgradient tests for genes for which there is evidence of expression spatial gradients with respect to a "reference" tissue niche/domain (e.g., higher expression closer to reference tissue niche, lower expression as farther from reference tissue niche). The method calculates distances from each spot/cell to the reference tissue niche/domain (e.g., a cluster defined via STclust – See "Spatial domain detection") and correlates those distances with gene expression values from top variable genes (defined by standard deviation across ROIs/spots/cells). The distances to reference niche can be summarized using the average or the minimum value. Generally, the minimum distances might be better to capture gradients at short ranges, while average distances capture whole-tissue gradients. The use of robust regression to reduce (albeit not eliminate) the effect of zero inflation in spatial transcriptomics data. Spearman correlation coefficients are calculated using the 'cor.test' R function. The most variable genes to be tested are identified *before* removal of outliers.

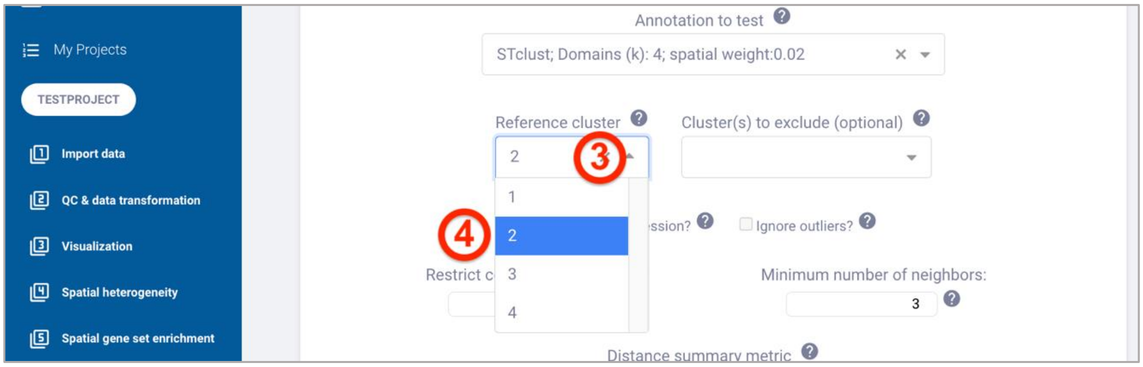
Select sample(s) to run this test: All samples selected

Number of most variable genes to use: 3000

Annotation to test: STclust; Domains (k): 4; spatial weight:0.02

spatialGE
About spatialGE
How to get started
FAQ
Contact us
My Projects
TESTPROJECT
1 Import data
2 QC & data transformation
3 Visualization
4 Spatial heterogeneity
5 Spatial gene set enrichment
6 Spatial domain detection
7 Differential expression
8 Spatial gradients
1
STclust; Domains (k): 2; spatial weight:0.02
STclust; Domains (k): 3; spatial weight:0.02
STclust; Domains (k): 4; spatial weight:0.02
STclust; Domains (k): 5; spatial weight:0.02
2

The next step is to select a **Reference cluster**. By looking the domains in the **Spatial domain detection** module, and the results from the **Differential expression** module, it seems like Cluster 2 represent the tumor domain in the "ductal_carcinoma" sample. To test for genes that show gradients with respect to the tumor domain, please select "2" from the **Reference cluster** drop-down. No domains will be excluded, however, should the user decide to remove specific domains, the user can do so by selecting the domains in the **Cluster(s) to exclude** drop-down.



Annotation to test: STclust; Domains (k): 4; spatial weight:0.02

Reference cluster: 2

Cluster(s) to exclude (optional)

Minimum number of neighbors: 3

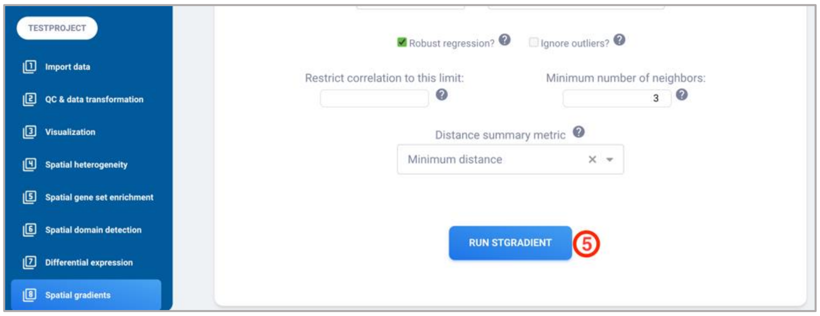
Note: The names given to each cluster by the **Spatial domain detection** module are not shared across samples. For example, if Cluster 2 in Sample_1 represents the tumor domain, in Sample_2, the tumor domain could be represented by Cluster_4. To compare results from the **Spatial gradients** module across samples, it is suggested to run the module several times using a different domain as **Reference cluster** each time.

The **Spatial gradients** module presents two alternatives to estimate the regression slope. The default option is to use robust regression via the MASS and sfsmisc packages R packages. The robust regression method reduces the effects of outliers on the estimation of regression coefficients. This method is useful in the analysis of ST data given the excess of zeroes. The robust regression method is activated by checking the **Robust regression?** checkbox. By unchecking the **Robust regression?** checkbox, the estimation method changes to OLS. If **Robust regression?** is left unchecked and the **Ignore outliers?** checkbox is checked, then outliers in each gene are removed from the calculation of regression coefficients via the interquartile method. Most ST studies would probably benefit from robust regression, and in this tutorial, **Robust regression?** will be left checked.

The **Restrict correlation to this limit** textbox enables calculation of regression and correlation coefficients based only on spots/cells within a given range. If the user thinks that expression gradients occur mostly at the immediate interface of two domains, this option can be used. Nevertheless, interpretation of this parameter may not be straightforward given that it does not have units (i.e., not measured in microns), but rather indicates the maximum Euclidean distance. The best approach to find an appropriate value to restrict the correlations is to run the analysis several times and using different values each time. For this tutorial, no value will be input in this textbox.

The **Minimum number of neighbors** textbox is an important parameter that helps reducing the effect of isolated reference spots/cells in the correlation. The parameter controls how many immediate neighbors a reference spot/cell has to have so that distances to that spot/cell are included in the correlation. Analyses on samples with uniform compact domains might benefit by using larger neighborhood sizes (e.g., healthy tissues with highly organized structure). On the other hand, if tissue domains are fragmented (e.g., cancer tissues with scattered tumor clones), smaller neighborhood sizes might be a better option. In the latter case, **Minimum number of neighbors** between three to six might be good choices. For this tutorial, the default (3) will be kept. Another important parameter is the **Distance summary metric** to be used. Due to easier interpretability, the "Minimum distance" option will be used in this tutorial. Nevertheless, we encourage the user to run the analysis using both metrics and look for consistent patterns.

To execute the analysis, click the **RUN STGRADIENT** button.


TESTPROJECT
1 Import data
2 QC & data transformation
3 Visualization
4 Spatial heterogeneity
5 Spatial gene set enrichment
6 Spatial domain detection
7 Differential expression
8 Spatial gradients
RUN STGRADIENT
5

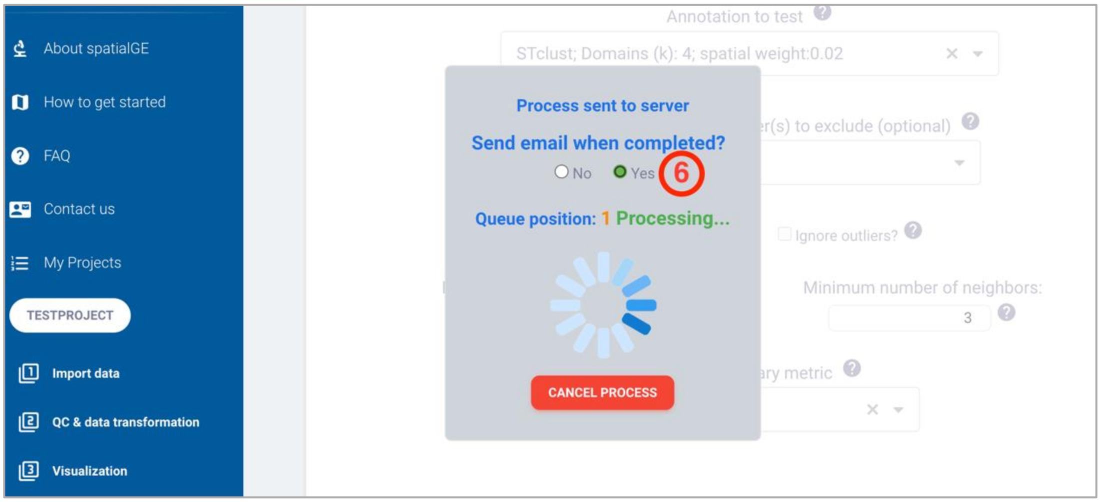
☑ Robust regression? ☐ Ignore outliers?

Restrict correlation to this limit:

Minimum number of neighbors: 3

Distance summary metric: Minimum distance

Depending on the number of samples and genes to be tested (**Number of most variable genes to use**), this analysis can be time-consuming. It is recommended to use the email notification functionality from the pop-up showing after initializing the analysis. An email notification can be received once the analysis has finished by selecting the **Yes** radio button under **Send email when completed?**



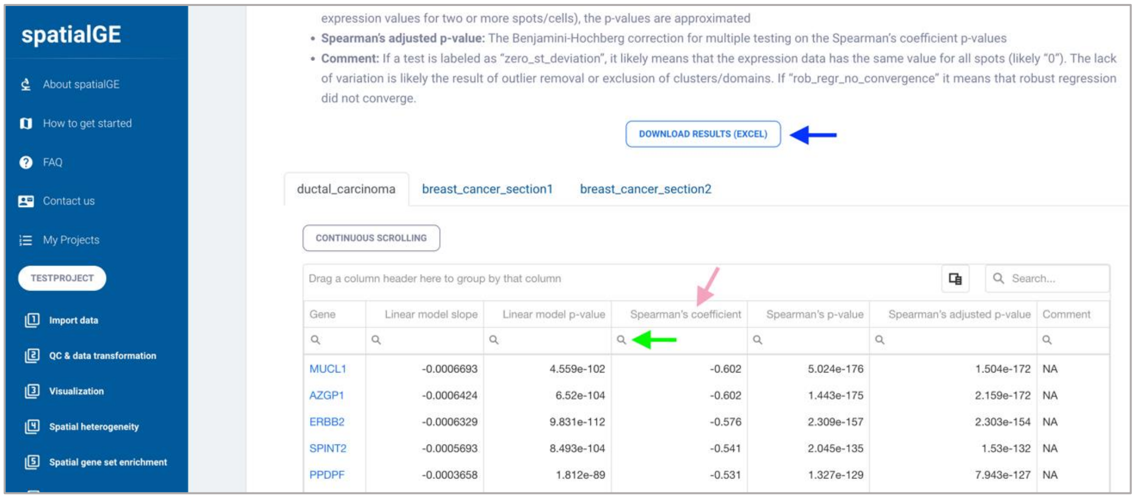
Process sent to server

Send email when completed?

○ No ● Yes

Queue position: 1 Processing...

Once the analysis is finished, explanation of the columns in the results table is presented. The results table has filtering and sorting capabilities by clicking on the loupe icons (green arrow) or the column name (pink) respectively. A result table is generated for each sample, which can be downloaded as spreadsheets in a workbook by clicking the **DOWNLOAD RESULTS (EXCEL)** button (blue arow). This functionality is especially useful when running the analysis using different parameters. Finally, by clicking on a gene name, the corresponding GeneCards record is opened.



expression values for two or more spots/cells), the p-values are approximated

- **Spearman's adjusted p-value:** The Benjamini-Hochberg correction for multiple testing on the Spearman's coefficient p-values
- **Comment:** If a test is labeled as "zero_st_deviation", it likely means that the expression data has the same value for all spots (likely "0"). The lack of variation is likely the result of outlier removal or exclusion of clusters/domains. If "rob_regr_no_convergence" it means that robust regression did not converge.

Gene	Linear model slope	Linear model p-value	Spearman's coefficient	Spearman's p-value	Spearman's adjusted p-value	Comment
MUCL1	-0.0006693	4.559e-102	-0.602	5.024e-176	1.504e-172	NA
AZGP1	-0.0006424	6.52e-104	-0.602	1.443e-175	2.159e-172	NA
ERBB2	-0.0006329	9.831e-112	-0.576	2.309e-157	2.303e-154	NA
SPINT2	-0.0005693	8.493e-104	-0.541	2.045e-135	1.53e-132	NA
PPDPF	-0.0003658	1.812e-89	-0.531	1.327e-129	7.943e-127	NA