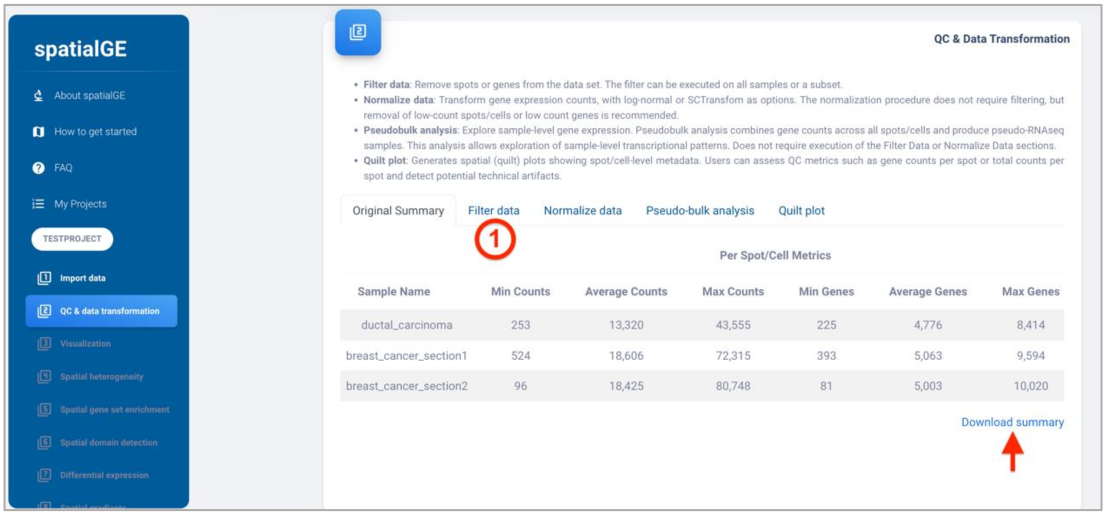
Original summary

Immediately after clicking finishing data import, you will be shown the summary of per spot/cell gene count metrics. This summary shows the number of counts and genes per spot (or cell, depending on the spatial transcriptomics technology), for each sample. Here a gene is counted if at least one read was detected for that gene in at least one spot/cell. This summary can be downloaded to a comma-delimited file if necessary (red arrow).

This tutorial will start by clicking the **Filter data** tab.



QC & Data Transformation

- **Filter data**: Remove spots or genes from the data set. The filter can be executed on all samples or a subset.
- **Normalize data**: Transform gene expression counts, with log-normal or SCTransfom as options. The normalization procedure does not require filtering, but removal of low-count spots/cells or low count genes is recommended.
- **Pseudobulk analysis**: Explore sample-level gene expression. Pseudobulk analysis combines gene counts across all spots/cells and produce pseudo-RNAseq samples. This analysis allows exploration of sample-level transcriptional patterns. Does not require execution of the Filter Data or Normalize Data sections.
- **Quilt plot**: Generates spatial (quilt) plots showing spot/cell-level metadata. Users can assess QC metrics such as gene counts per spot or total counts per spot and detect potential technical artifacts.

Original Summary | Filter data | Normalize data | Pseudo-bulk analysis | Quilt plot

Per Spot/Cell Metrics

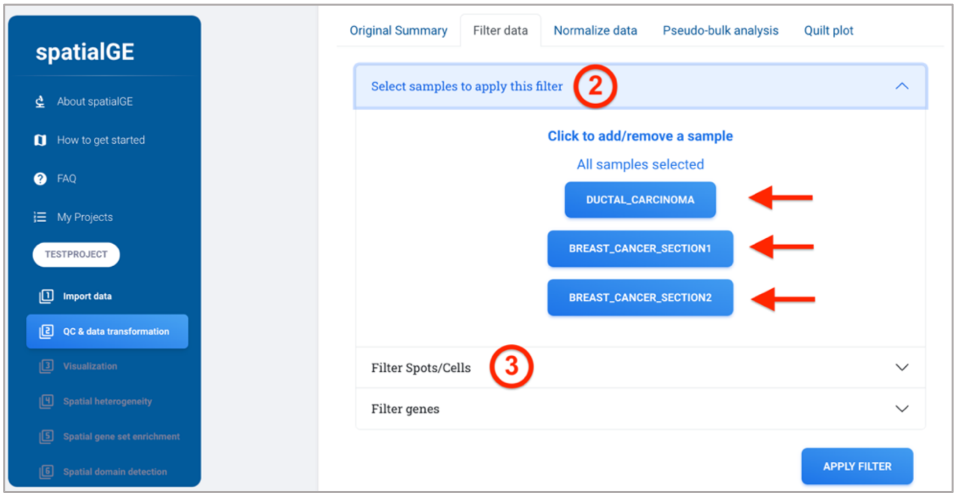
Sample Name	Min Counts	Average Counts	Max Counts	Min Genes	Average Genes	Max Genes
ductal_carcinoma	253	13,320	43,555	225	4,776	8,414
breast_cancer_section1	524	18,606	72,315	393	5,063	9,594
breast_cancer_section2	96	18,425	80,748	81	5,003	10,020

Download summary

Filter data

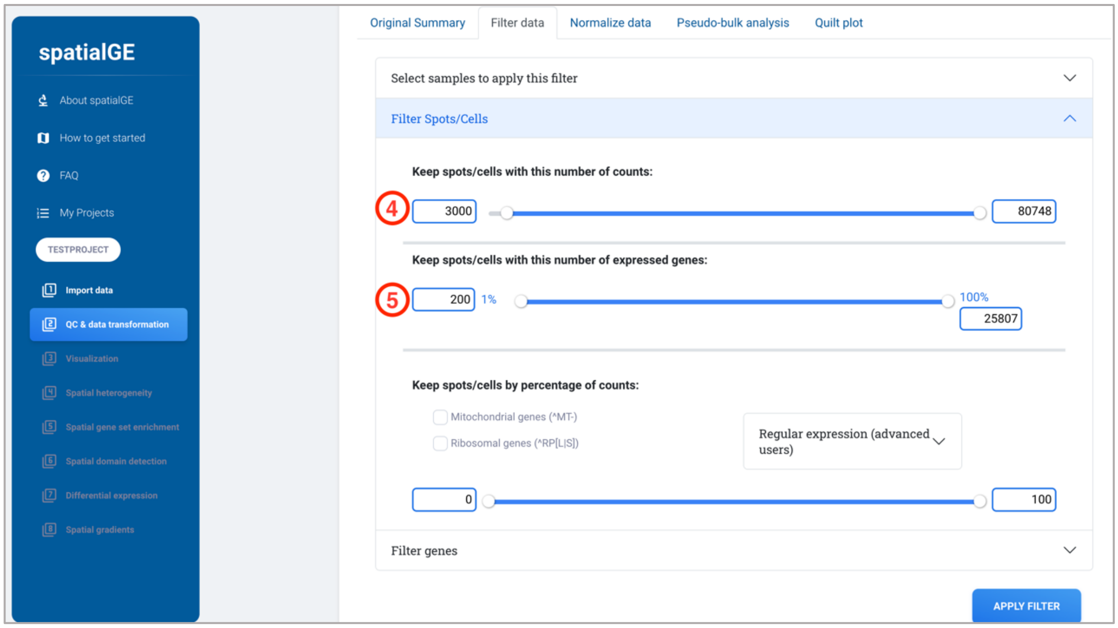
In the **Filter data** tab, please click the "Select samples to apply this filter" text. Blue boxes with the uploaded samples will appear. You can click individual samples that you do not wish to apply the filter (red arrows). In this tutorial, however, no samples will be excluded from the filter as most users will probably want to apply filters to all samples in a project.

Please click in the next section **Filter spots/cells**.



The **Filter spots/cells** section removes Regions of Interest (ROIs), spots, or cells based on the number of counts or genes per ROI/spot/cell. The first filter in this section, **Keep spots/cells with this number of counts** defines the minimum and maximum gene counts an ROI/spot/cell should have to be kept in the data set after filtering. The user is encouraged to think about reasonable values for this filter. Since each project is different, and factors such as quality of the sample, sequencing depth, level of necrosis, affect the number of counts per ROI/spot/cell, establishing default values is difficult. From the **Original summary** tab, it can be seen that the spot each Visium sample has spots with gene counts as low as 96. For the purpose of this example, spots with less than 2000 counts per spot will be removed, which should be enough to remove low-quality spots. For that, the user can either type in the number in the left box or move the left slider button to the right until 2000 shows in the left box. The maximum number of counts per spots will be left untouched. Nevertheless, for technologies that use image segmentation to define cells, the user might want to limit the maximum number of counts per cell in order to remove poorly segmented cells ("doublets").

Next, the **Keep spots/cells with this number of expressed genes** removes ROIs, spots, or cells based on the number of expressed genes. Here, a gene is considered expressed if at least one count was detected in at least one spot. In general, spots with low number of expressed genes are also considered low-quality. From the **Original summary** tab, it can be seen that the sample "breast_cancer_section2" has spots with as low as 81 genes. To remove those spots, please write in the left box the number 200. Similarly, the maximum number of genes will be left untouched, but the same previous considerations apply for image segmentation-based spatial transcriptomics.



Keep spots/cells with this number of counts:

4 3000 80748

Keep spots/cells with this number of expressed genes:

5 200 1% 100% 25807

Keep spots/cells by percentage of counts:

□ Mitochondrial genes (^MT-)
□ Ribosomal genes (^RP[L|S])

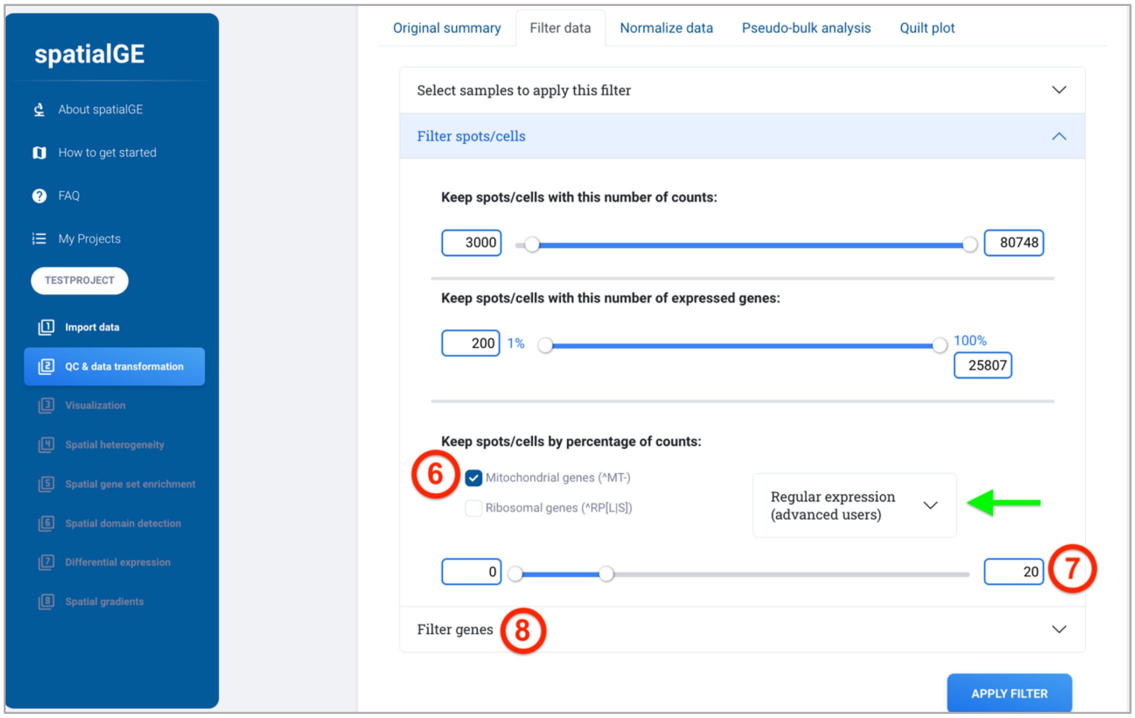
Regular expression (advanced users)

0 100

The spots can also be filtered based on the percentage of mitochondrial or ribosomal counts. **Under Keep spots/cells by percentage of counts**, the user can specify the minimum and maximum percentage of counts from mitochondrial or ribosomal genes. Here, spots will be filtered based only on mitochondrial gene counts by checking the "**Mitochondrial genes (^MT-)**" box. Next, move the slider button on the right to the 20% mark or type 20 in the right box. Keep the left slider button in zero. With this filter, spots that contain 0 to 20% of mitochondrial reads are kept. A 20% cutoff for mitochondrial counts is commonly used in many single-cell studies. Here, this filter aims to remove spots from areas with cellular stress or death which may have higher expression of mitochondrial genes.

Note: The "^MT-" in the **Mitochondrial genes** check box indicates that spatialGE searches for genes with names starting by "MT" followed by a hyphen in the gene names. This filter works in Visium as mitochondrial genes are named that way; however other technologies might not follow that naming pattern. In those cases, the user can input a regular expression or token that captures the names of the genes to be used in the filter in the **Regular expression** box (green arrow).

Now, please click the **Filter genes** section.



Select samples to apply this filter

Filter spots/cells

Keep spots/cells with this number of counts:
3000 — 80748

Keep spots/cells with this number of expressed genes:
200 1% — 100% 25807

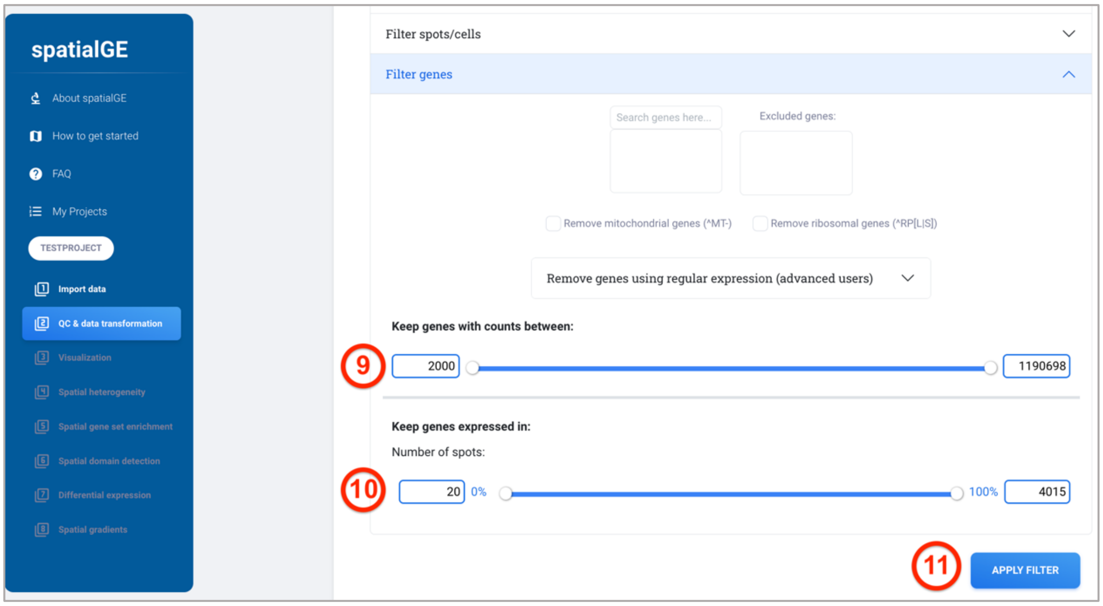
Keep spots/cells by percentage of counts:
☑ Mitochondrial genes (^MT-)
☐ Ribosomal genes (^RP[L|S])
Regular expression (advanced users)
0 — 20

Filter genes

The **Filter genes** section removes genes based on the total number of counts or ROIs/spots/cells where those genes are expressed. The user can also remove genes by name using the two boxes at the top of the section. This functionality is especially useful when removing negative probes or housekeeping genes used for background noise correction. The user can type in the left text box (**Search genes here…),** which will fill the box under it with genes that matching the typed text. By clicking on a gene name, it is passed to the box on the right (**Excluded genes**). If the user decides to remove all mitochondrial or ribosomal genes from the sample, it can be accomplished by clicking one or the two corresponding check boxes (**Remove mitochondrial genes** and/or **Remove ribosomal genes**). Similarly, you can search for regular expression that matches certain genes to be removed (**Remove genes using regular expression**).

In this example, genes will not be removed by gene. Instead, the next two sliders will be used. The **Keep genes with counts between** slider specifies the minimum and maximum number of counts for a gene to be kept. Similar to the filters in the previous sections, the values used in the gene filters are relative to each experiment, and hence defaults or guidelines are difficult to establish. For this example, please type 2000 in the left box, which means that a gene has to have at least 2000 counts across the sample (not in each spot) to be included. The following filter is **Keep genes expressed in**, which controls the number of ROIs/spots/cells a gene must be expressed in (i.e., at least one count) to be kept. Before using this filter, the user is encouraged to think about the particularities of their samples. For example, if a gene is not expressed in a spot, it does not necessarily mean that the technology failed to provide data, given that the same gene can be expressed abundantly in other spots. A reasonable value would be to filter genes expressed in at least 20 spots. The expectancy here is that lowly expressed genes could be removed to avoid problems with statistical testing (e.g., differential expression). Please, type the number 20 in the left box of this filter.

By clicking the **APPLY FILTER** button, the filters selected in all the previous section is executed.



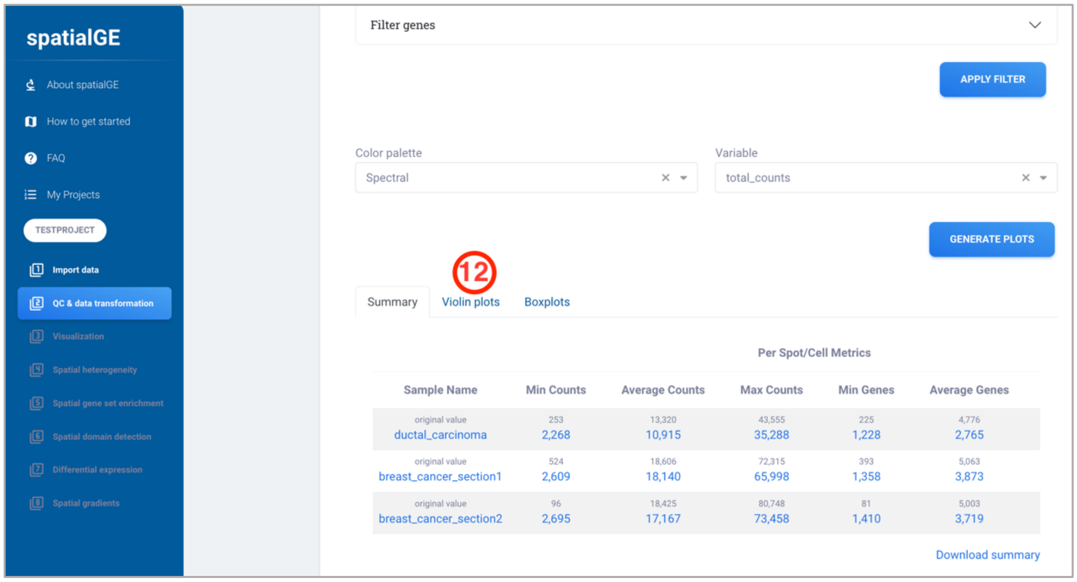
□ Remove mitochondrial genes (^MT-) □ Remove ribosomal genes (^RP[L|S])

Keep genes with counts between: 2000 – 1190698

Keep genes expressed in:
Number of spots: 20 0% – 100% 4015

A pop-up will appear asking if an email notification should be sent once the filter is applied. Then, a new summary of genes and counts is presented below the **APPLY FILTER** button. This summary shows the results of the filtering. The blue numbers indicate the new number of counts and genes. The gray colors show the original values. The minimum number of counts and genes per spot seem to be in the same approximate range, which is result of a filtering procedure focused on the minimum values and aiming to remove low-quality spots. Notice this summary can also be downloaded as a .csv file.

Graphical summaries are available by clicking the **Violin plots** or **Boxplots** tab.



Filter genes

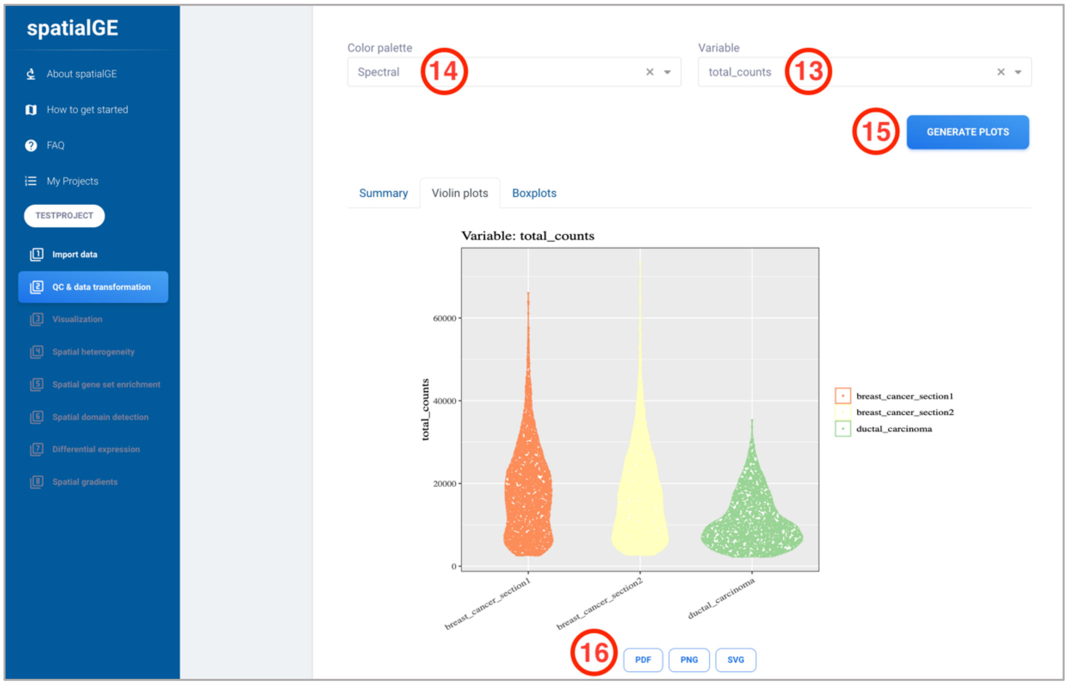
Color palette: Spectral

Variable: total_counts

Per Spot/Cell Metrics

Sample Name	Min Counts	Average Counts	Max Counts	Min Genes	Average Genes
original value ductal_carcinoma	253 2,268	13,320 10,915	43,555 35,288	225 1,228	4,776 2,765
original value breast_cancer_section1	524 2,609	18,606 18,140	72,315 65,998	393 1,358	5,063 3,873
original value breast_cancer_section2	96 2,695	18,425 17,167	80,748 73,458	81 1,410	5,003 3,719

In the **Violin plots** tab, an auto-generated plot shows the distribution of counts. Each point within the violins represents an ROI/spot/cell. With this visual summary is easy to notice that the "breast_cancer_section1" and "breast_cancer_section2" samples have higher counts compared to "ductal_carcinoma". This is expected as the "ductal_carcinoma" data comes from a formalin-fixed paraffin-embedded (FFPE) tissue. This current version of spatialGE allows to plot a second QC metric. By clicking on the **Variable** drop-down, the "total_genes" can be selected. If a different color palette is desired, you can select it in the **Color palette** drop-down. The palettes are provided by the Khroma and RColorBrewer R packages. By clicking the **GENERATE PLOTS** button, and then again, the **Violin plots** drop-down, the violin plots showing the number of genes per spot are presented. Notice that these plots can be downloaded using the **PDF**/**PNG**/**SVG** buttons below the plots.



Color palette: Spectral

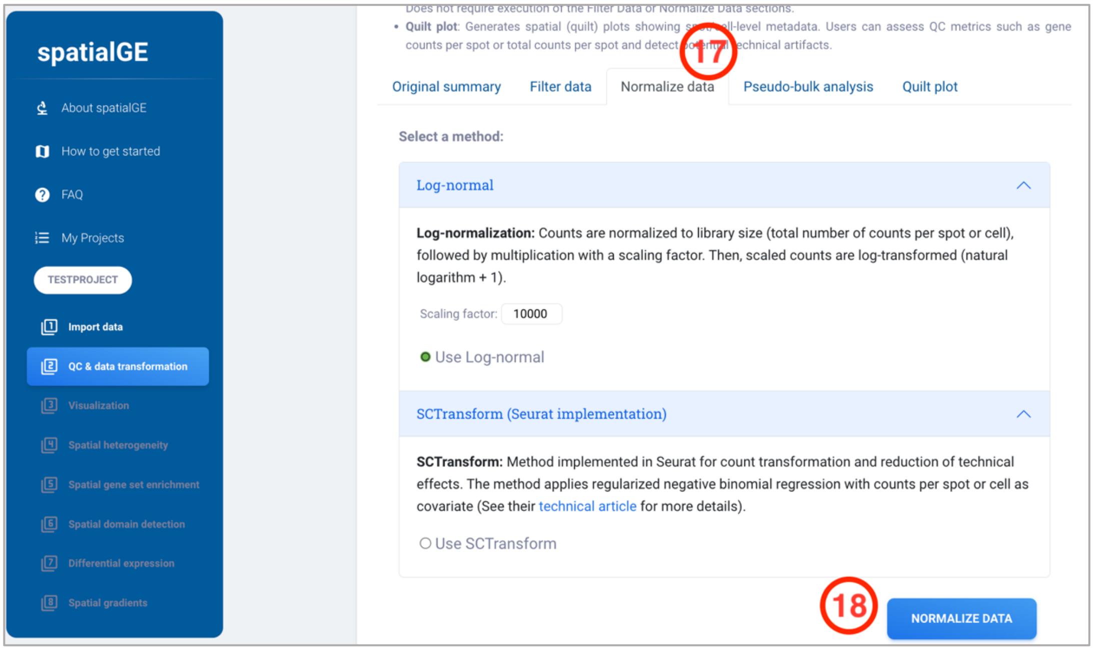
Variable: total_counts

Normalize data

The next step in the analysis is to click on the **Normalize data** tab. The normalization of data is required to enable additional analysis modules in spatialGE. Currently, two normalization methods are available: library-size normalization followed by log-transformation (**Log-normal**) and the variance stabilization method **SCTransform**. The latter uses negative binomial regression residuals to reduce the effect of spot-to-spot total count variation from the gene counts, with the expectation of extracting more biologically meaningful differences. For further details, please refer to the technical article describing the method.

Here, the user selects the method to be used by clicking on the respective radio button. In this example, the default **Log-normal** will be used. Notice that the parameter **Scaling factor** is available for the **Log-normal** transformation. This parameter must not be confounded with the JSON file used in the data import process. Instead, this **Scaling factor** is a rough way to scale ROI/spot/cell counts to a value during normalization. The default 10,000 results from the expectation that most spots have that number of counts. The user can intuit that this factor must be adjusted according to the technology used. For example, in single-cell level spatial transcriptomics, this value will probably need to be set to a lower value.

To perform the data normalization procedure, click the **NORMALIZE DATA** button.



Does not require execution of the Filter Data or Normalize Data sections.

- **Quilt plot**: Generates spatial (quilt) plots showing spot/cell-level metadata. Users can assess QC metrics such as gene counts per spot or total counts per spot and detect potential technical artifacts.

Select a method:

Log-normal

Log-normalization: Counts are normalized to library size (total number of counts per spot or cell), followed by multiplication with a scaling factor. Then, scaled counts are log-transformed (natural logarithm + 1).

Scaling factor: 10000

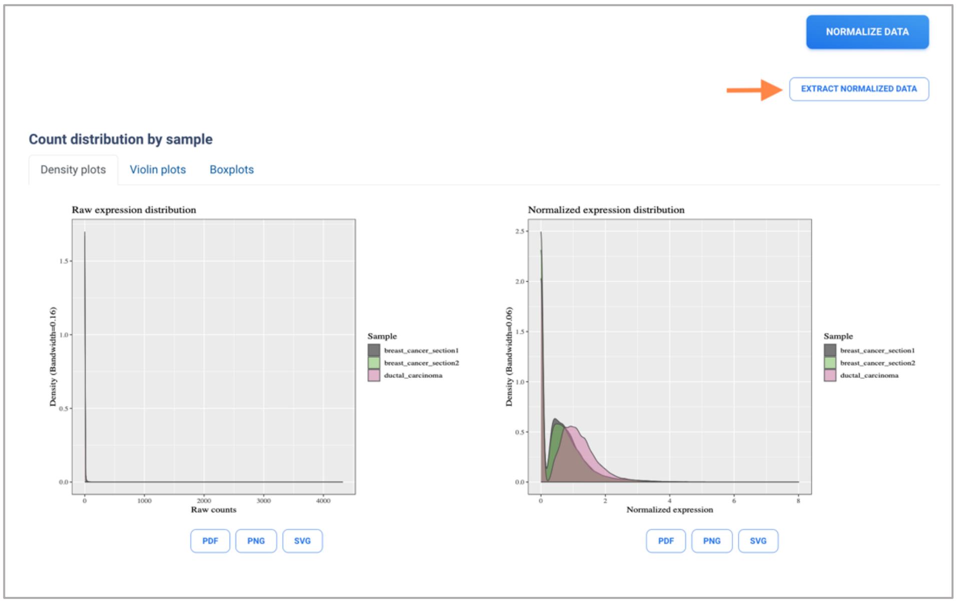
● Use Log-normal

SCTransform (Seurat implementation)

SCTransform: Method implemented in Seurat for count transformation and reduction of technical effects. The method applies regularized negative binomial regression with counts per spot or cell as covariate (See their technical article for more details).

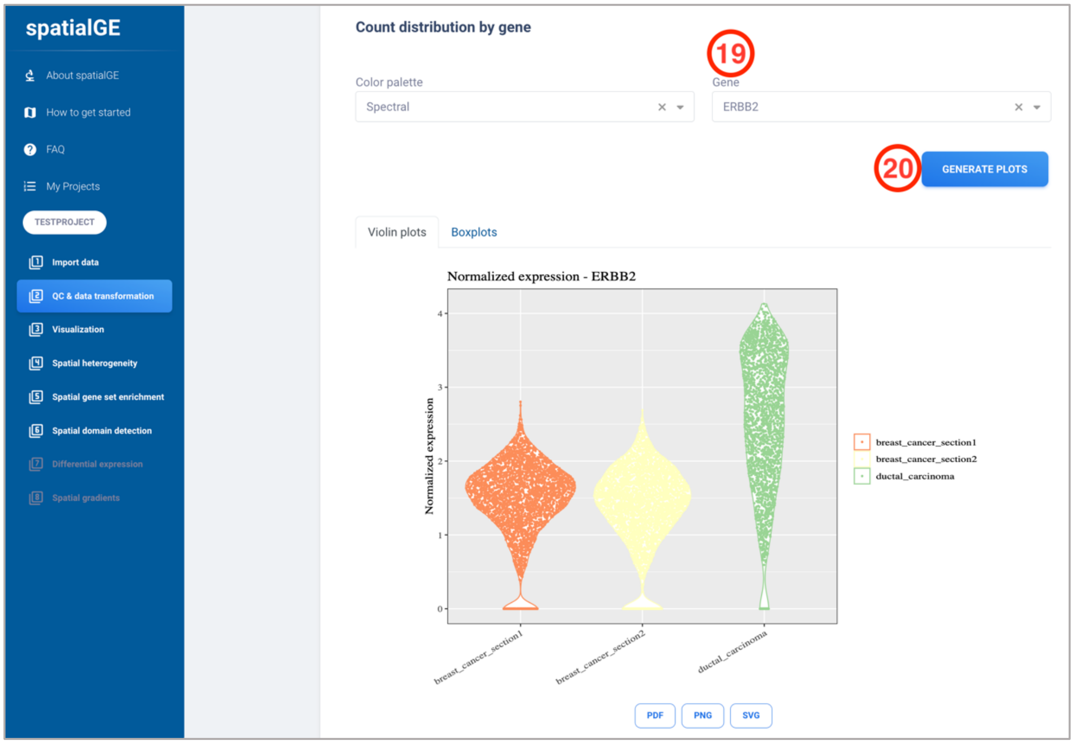
○ Use SCTransform

After the normalization procedure is finished, **Count distribution by sample** plots are displayed under the **NORMALIZE DATA** button. These plots aim to facilitate the user's observation of the effects of normalization. By saving the plots (**PDF/PNG/SVG** buttons), the user can then re-run the normalization with another method and decide which one provides the best results. Of note, you can download the normalized gene counts (**EXTRACT NORMALIZED DATA**, orange arrow) for analysis and plotting outside spatialGE (it is a time-consuming procedure, and it is recommended to activate the email notification).



Count distribution by sample

Below plots in **Count distribution by sample**, the user has the option to plot the count distribution for specific genes. In the **Count distribution by gene** the **Gene** text box allows typing the name of a gene to display **Violin plots** or **Boxplots**. The plots can also be customized by selecting a color palette. In the **Gene** text box, please type ERBB2 and then click **GENERATE PLOTS**. This new plot shows that expression of ERBB2 (also known as HER2) is considerably higher in the "ductal_carcinoma" sample, indicating that HER2 is expressed in a large area this tissue sample.



Count distribution by gene

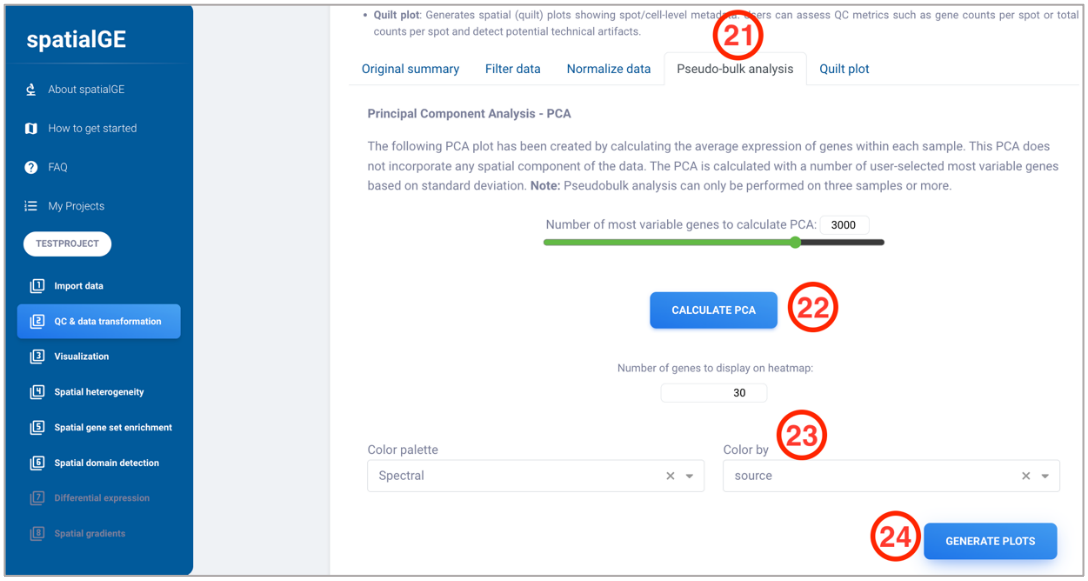
19 Gene: ERBB2

Color palette: Spectral

Next, please click on the **Pseudo-bulk analysis** tab.

Pseudo-bulk analysis

The **Pseudo-bulk analysis** tab provides functionality to assess the overall level of sample-to-sample similarity. Pseudobulk counts are generated by aggregating the spot counts of each, mimicking a "bulk" sample. The consequence is that spatial information is disregarded, and hence this analysis is only intended as a data QC method. Once in the **Pseudo-bulk analysis** tab, start by selecting the **Number of most variable genes to calculate PCA**. Values between 3000 and 5000 (though not limited to this value), can be enough to visually assess sample similarity. Here, the default 3000 will be used. Now, please click **CALCULATE PCA**.



- **Quilt plot**: Generates spatial (quilt) plots showing spot/cell-level metad[illegible]sers can assess QC metrics such as gene counts per spot or total counts per spot and detect potential technical artifacts.

Principal Component Analysis - PCA

The following PCA plot has been created by calculating the average expression of genes within each sample. This PCA does not incorporate any spatial component of the data. The PCA is calculated with a number of user-selected most variable genes based on standard deviation. **Note:** Pseudobulk analysis can only be performed on three samples or more.

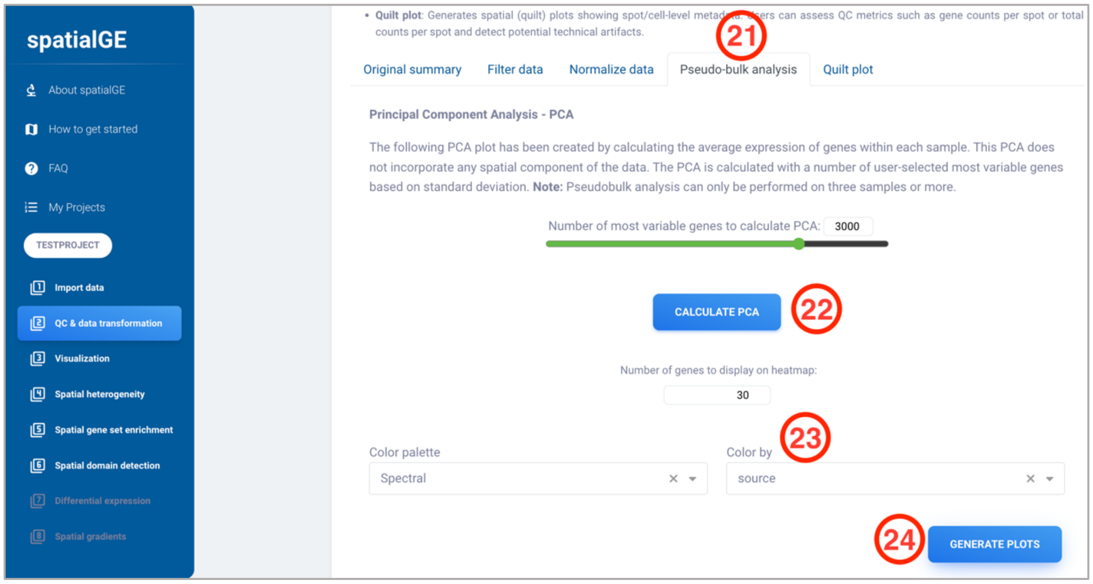
Number of most variable genes to calculate PCA: 3000

Number of genes to display on heatmap: 30

Color palette: Spectral

Color by: source

The **CALCULATE PCA** button performs the aggregation of counts and calculation of a principal component analysis (PCA). Once finished, additional functionality is enabled to visualize the results of pseudo-bulk analysis. Two visualizations are provided a PCA scatterplot or a heatmap. For the heatmap, the **Number of genes to display on heatmap** text box allows the user to input the number of most variable genes (based on pseudo-bulk counts) to be displayed. In this example, the default (30 genes) will be kept. Next, the drop-down **Color by** specifies the variable used to color the samples in the PCA plot. The variables in **Color by** are directly extracted from the sample metadata input during data import. If no sample-level metadata was provided, then the points in the PCA are colored by sample (each sample its own color). Please, click the **Color by** drop-down and select "source". Then, click **GENERATE PLOTS**.



- **Quilt plot**: Generates spatial (quilt) plots showing spot/cell-level metadata. Users can assess QC metrics such as gene counts per spot or total counts per spot and detect potential technical artifacts.

Principal Component Analysis - PCA

The following PCA plot has been created by calculating the average expression of genes within each sample. This PCA does not incorporate any spatial component of the data. The PCA is calculated with a number of user-selected most variable genes based on standard deviation. **Note:** Pseudobulk analysis can only be performed on three samples or more.

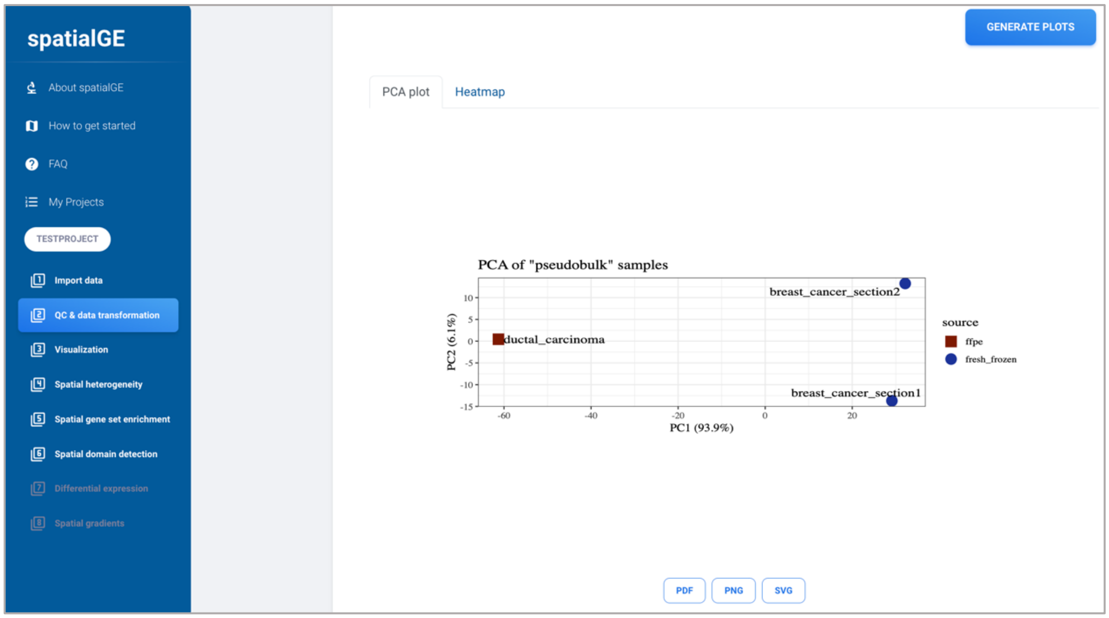
Number of most variable genes to calculate PCA: 3000

Number of genes to display on heatmap: 30

Color palette: Spectral

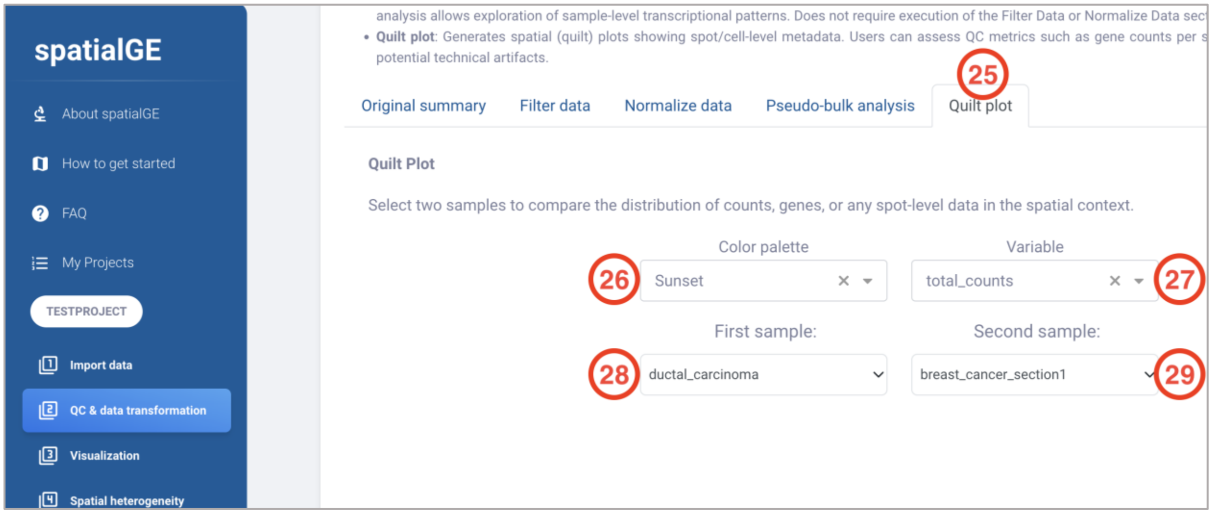
Color by: source

After clicking **GENERATE PLOTS**, two new tabs will appear. The first one, **PCA plot** presents the scatterplot of samples using the calculated principal components. This plot shows that samples coming from fresh frozen tissues are more similar between them compared to the sample extracted from FFPE tissue. This is the expected pattern, and otherwise would require revision of the data for technical effects. The second tab shows the **Heatmap**, with top variable genes in rows and pseudo-bulk samples in columns.



Quilt plot

By clicking on the **quilt plot** tab, the user can generate plots to observe the variation in gene counts at each ROI/spot/cell, organized spatially. The quilt plots are a quick way to identify areas with higher counts and possibly higher cellular density, or conversely, areas with low counts and potentially high necrosis. The **Color palette** drop-down allows you to change the color of the ROIs/spots/cells in the quilt plot. Since the spots in the quilt plot present quantitative data (total counts or total genes), the diverging and sequential palettes are recommended (see the Khroma and RColorBrewer documentation for more information on those types of palettes). For this example, please select the “Sunset” palette. Then, select “total_counts” from the **Variable** drop-down. Next, please select the “ductal_carcinoma” and “breast_cancer_section1” samples from the drop-downs “**First sample**” and “**Second sample**”.



analysis allows exploration of sample-level transcriptional patterns. Does not require execution of the Filter Data or Normalize Data sec

- **Quilt plot**: Generates spatial (quilt) plots showing spot/cell-level metadata. Users can assess QC metrics such as gene counts per s potential technical artifacts.

Quilt Plot

Select two samples to compare the distribution of counts, genes, or any spot-level data in the spatial context.

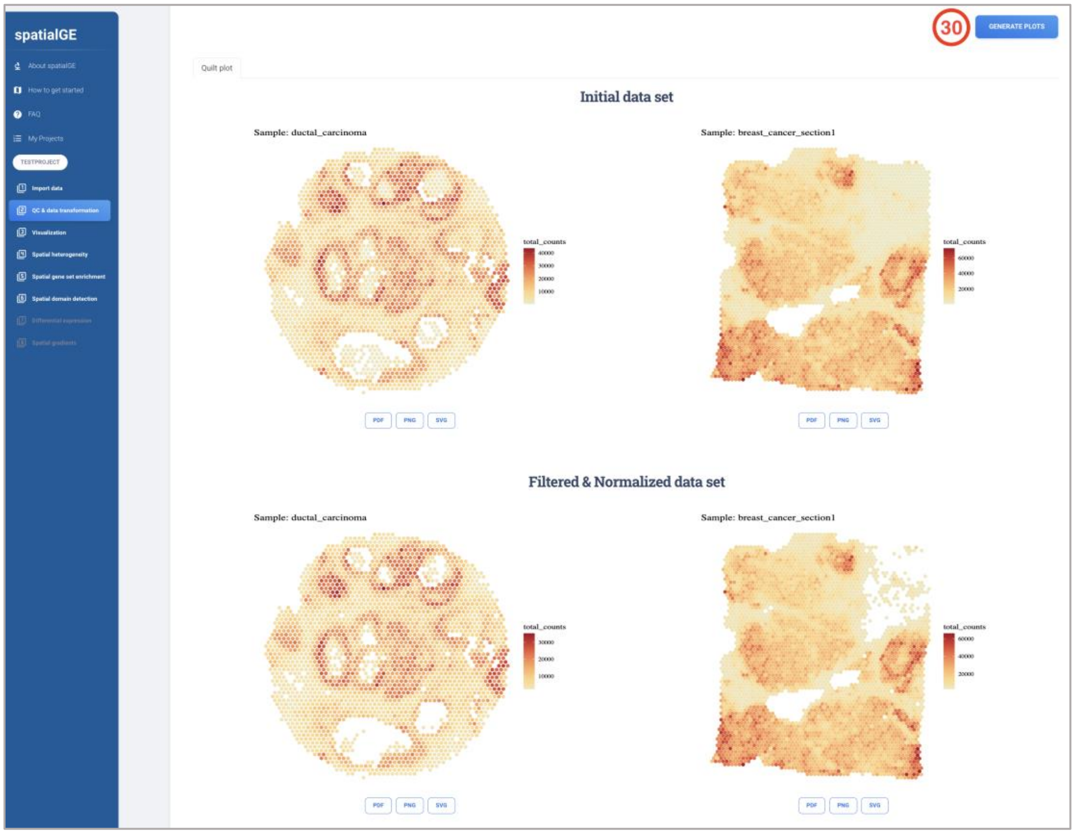
Color palette: Sunset

Variable: total_counts

First sample: ductal_carcinoma

Second sample: breast_cancer_section1

Finally, click the “**GENERATE PLOTS**” button.



The resulting quilt plots show the tissue regions with high number of counts. If filtering and normalization has been performed, the spatialGE produces before and after plots. Here, the user can see that some of the areas with low counts were removed after filtering. These figures can help the user to decide if filtering has been too stringent, or if different filtering options need to be applied.

Note: Currently, spatialGE only generates quilt plots for the total number of counts per spot and the total number of genes per spot (i.e., genes with one count or more per spot). It is the expectation that other QC metrics such as the percentage of mitochondrial genes will be implemented in future versions, or other user-provided per-ROI/spot/cell metrics.