Gene Set Enrichment Analysis with GSEA:

You can download the desktop application GSEA from here:
http://www.gsea-msigdb.org/gsea/index.jsp

(published here, for the citation in case you use it:
https://www.pnas.org/doi/10.1073/pnas.0506580102 and
https://www.nature.com/articles/ng1180)

You can download the desktop application GSEA from here: http://www.gsea-msigdb.org/gsea/index.jsp

Two input files are required:

PROJECTNAME_xxsamples_vst_transf_counts_GeneSymbols_forGSEA.txt # data matrix with variance-stabilizing transformed count data. The header and column layout has been formated for GSEA.

PROJECTNAME_xxsamples_yyclasses_forGSEA.cls # Classes file indicating group membership of the samples.

GSEA_settings_screenshot.png # Screenshot with the parameters to be filled in.

For a simple analysis, perform the following steps:

1. Open GSEA

2. Under "Load Data/Browse for files" , upload the two input files. There should be no errors reported after reading in the files.

3. Under "Run GSEA" fill in the details for the analysis, as shown in the example in the screenshot.

Under "gene sets database" you can select gene sets from MSigDB (one or several) but also upload your own gene sets (previously formatted and uploaded).

With the latest update, besides human gene sets, MSigDB also contains gene set collections for mouse. The difference here is in the gene nomenclature for human and murine genes, and of course also in the underlying studies that identified these organism-specific gene sets.

Please use human collections exclusively if your data are human data.

Please use murine collections exclusively if your data are murine data.

If your data are from organisms other than human or mouse, please follow the instructions of the MUW Core Facilities data analyst.

Select the two groups to be compared under "Phenotype labels". (Caution, this analysis is directional, meaning that there is both e.g. control_vs_infected and infected_vs_control; the results have to be read as "enriched in the first group". It is also possible to compare one group against all other samples = REST.

4. "Run".

Caution: When many gene sets have been selected, quite a bit of RAM is required; 32 GB of RAM is usually sufficient; only run one analysis at a time; analysis run time can be one hour or more.

In the .zipped results folder, there is an index.html file which can be opened in your browser. There, you find links to the results, there is also a link to the online user guide. Basically, the "interesting" results are shown in the uppermost section, e.g. "Enrichment in phenotype: infected (5 samples)". Here it says how many gene sets are significantly enriched.
The link "detailed enrichment results in html format" goes to a new page where you can explore in more detail, which gene sets are enriched, and which p- and q-values they have. Alternatively, these results are also available as Excel file.
Clicking on the gene set name links to an online page with information on the origin of the gene set etc.
Clicking on "details" will show a ranked list of genes in the gene set, with those genes highlighted that have contributed to the enrichment of the gene set in the present 2-group comparison.
As always, start from the lowest FDR q-values which are the most significant ones. Stay within the FDR q-val <0.25 range to see significant gene sets.

One more thing to consider is that most gene sets in GSEA/MSigDB have been put together based on human data. Some gene sets do have "murine" in their title, though.
All gene sets use capital letters for the gene symbols (like for human gene symbols).