GSEA GUIDELINES

Guidelines for Gene Set Enrichment Analysis Interpretation
Serrano Lab

GSEA is a powerful computational method that identifies whether a predefined set of genes (gene set) shows a statistically significant, coordinated difference in expression between two biological states or phenotypes. Below are key guidelines to help you perform and interpret GSEA analyses effectively.

Main Goal of the Analysis

The main goal of GSEA is to move beyond examining individual genes in isolation and instead consider the collective behavior of functionally related genes. By focusing on biological pathways or processes (gene sets), GSEA can provide insights into the underlying biology of your system, showing if certain functions, pathways, or regulatory programs are upregulated or downregulated under different conditions.

Guidelines

Input Gene List and Ranking Criteria

Input:
GSEA typically requires a ranked list of **all genes** tested in the differential expression (DE) analysis. These genes should encompass the entire transcriptome analyzed, not just the significantly differentially expressed ones. The rationale for including all genes is that GSEA computes enrichment scores by walking down the entire ranked list, focusing on the distribution of genes from a set among the up- or down-regulated genes.

Ranking Genes:
You can rank genes by either statistical significance (p-value), magnitude of change (log2 fold change), or a combined metric. Each method has a biological rationale:

- **Ranking by p-value:**
  Prioritizes genes that show the most statistically reliable differences between conditions. The assumption is that highly significant genes (even with smaller effect sizes) might represent key regulatory shifts reliably detected by your experiment.
- **Ranking by log2 fold change (LFC):**
  Emphasizes genes with the greatest magnitude of change. The rationale is that large changes in gene expression might be more biologically meaningful, even if less statistically robust. This approach can highlight strong perturbations in specific pathways.

In practice, many workflows use a statistic that incorporates both significance and effect size, such as a signed log p-value or a moderated statistic from tools like **DESeq2**.

Normalized Enrichment Score (NES)

The **Normalized Enrichment Score (NES)** is a key metric from GSEA. It adjusts the raw enrichment score (which measures how much a gene set is overrepresented at the top or bottom of the ranked list) by accounting for the size of the gene set. This normalization allows comparisons across gene sets of different sizes.

Biological Interpretation:
If the NES is positive and large, the leading genes of that gene set are predominantly found at the top of your ranked list (i.e., they are upregulated in your condition of interest). Conversely, a large negative NES

indicates the leading genes are concentrated at the bottom (downregulated). The "leading edge" subset of genes identified by GSEA is often the most biologically relevant portion of the gene set driving the enrichment signal.

Statistical Thresholds for Significance (Nominal p-value and q-value)

GSEA provides two key significance measures:

- **Nominal p-value:**
  Often a threshold of 0.05 is used for the nominal p-value. This provides an initial indication that the enrichment is not due to random chance.

- **Adjusted q-value (FDR):**
  GSEA often uses a false discovery rate (FDR) cutoff of $q < 0.25$. This is more lenient than typical RNA-seq analyses. The reason is that GSEA is often used as a hypothesis-generation tool. By setting a higher q-value threshold, you allow for broader discovery of potentially interesting pathways that warrant further investigation, rather than prematurely excluding sets that might hold biological relevance.

In other words, GSEA is typically exploratory, so an FDR of 0.25 is accepted (as per the original GSEA publications) to identify candidate gene sets for follow-up analyses.

Using Custom Gene Sets

While GSEA can use established gene set databases such as MSigDB, you can also create custom gene sets to suit your specific biological questions. Custom gene sets might be based on literature searches, known biomarkers, or pathways not represented in standard databases.

Parameters and Guidelines for Custom Gene Sets:

- **Size:** Include at least 10-15 genes to ensure robust statistical enrichment detection. Very small gene sets may yield unstable results, while very large sets may dilute the signal.
- **Biological Relevance:** Curate genes known to participate in a specific biological process, pathway, or cellular compartment. Ensure these genes are experimentally validated or supported by strong evidence.
- **Update and Validate:** Regularly update and validate your gene sets as new data and literature emerge.
- **Formatting:** Provide gene sets as lists of gene identifiers that match the gene naming convention used in your dataset.

Interpreting the Results and Suggested Next Steps

Once you have a list of enriched gene sets with corresponding NES, p-values, and q-values, consider the following next steps:

- **Biological Contextualization:**
  Map the enriched gene sets back to the biological question you are asking. Are the enriched pathways related to cell cycle, immune response, metabolism, or differentiation? Place the results in the context of the known biology, the experimental design, and the phenotype of interest.

- **Leading Edge Analysis:**
  Examine the subset of genes contributing most to the enrichment score, known as the "leading edge" genes. These may represent key regulators or effectors driving the observed phenotype.

- **Validation Through Literature and Databases:**
  Cross-reference your significantly enriched gene sets with known biological databases and literature. Are these gene sets previously implicated in similar conditions or phenotypes?

- **Experimental Follow-Up:**
  Design targeted experiments to validate the role of these pathways or genes. This might include qPCR or Western blotting of key regulators, knockdown or overexpression studies, or flow cytometry to verify cellular states.

- **Compare Across Conditions or Time Points:**
  If you have multiple conditions or time points, compare the enriched gene sets across these contexts. Identify persistent or condition-specific pathways, which may yield insights into dynamic regulatory changes.

- **Integration with Other Omics:**
  Consider integrating GSEA results with other omics data types (e.g., proteomics, metabolomics) to build a more comprehensive picture of the underlying biology.

- **Refinement of Hypotheses:**
  Use the GSEA results to refine or generate new hypotheses. The goal is to move from broad pathway-level insights to more focused, mechanistic experiments.

Additional Considerations and Feedback

- **Data Quality Control:**
  Ensure input data quality (e.g., normalization, batch correction) before performing GSEA to reduce false signals.

- **Multiple Contrasts:**
  If performing GSEA on multiple contrasts (e.g., multiple clusters or conditions), organize and compare results to identify shared or distinct pathways.

- **Reporting and Reproducibility:**
  Keep records of the version of gene sets used, the tool and parameters, and the exact code for GSEA to ensure reproducibility. Consider using public repositories or notebooks (e.g., GitHub, Quarto, R Markdown) for transparent reporting.

References and Tools

- Subramanian A, Tamayo P, et al. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.” *PNAS* (2005). Link to the article

- Liberzon A, Subramanian A, et al. “Molecular signatures database (MSigDB) 3.0.” *Bioinformatics* (2011). Link to the article

- Gene Set Enrichment Analysis (GSEA) User Guide. Link to website

- Christopher Weidner, Matthias Steinfath, et al. “A Protocol for Using Gene Set Enrichment Analysis to Identify the Appropriate Animal Model for Translational Research.” *J Vis Exp* (2017) Link to the article

- “GSEA Enrichment Analysis: A Quick Guide to Understanding and Applying Gene Set Enrichment Analysis.” *Metwarebio* Link to the blog

Example

Below is an example of setting the q-value threshold to 0.25 and performing GSEA in R, as shown in your provided code block. This code uses `fgsea` for enrichment and demonstrates how to create and plot your results, as well as how you might define your own gene set:

```
# Example: Setting FDR threshold to 0.25
fgsea_res0_hallmark = fgsea_all(res = sc_results0,
                                gsets = msig_hallmark,
                                FDR = 0.25, nperm = 10000)

# Checking results and adding a significance column
fgsea_res0_hallmark$sig = fgsea_res0_hallmark$padj < 0.05

# Example plot of NES scores using ggplot2
library(RColorBrewer)
pal <- brewer.pal(n = 2, name = "Accent")

p1 <- ggplot(fgsea_res0_hallmark, aes(reorder(pathway, NES),
              NES, fill = sig)) +
      geom_col(colour = "black", size=0.2, width=0.9) +
      scale_fill_manual(values = pal) +
      coord_flip() +
      labs(x = "Pathway",
           y = "Normalized Enrichment Score (NES)",
           fill = "p adj < 0.05",
           title = "Hallmark pathways NES from GSEA") +
      theme_linedraw() +
      geom_text(aes(label = size),
      vjust = "center",
      hjust = "outward", size = 3, colour = "gray30")

print(p1)

# Creating a custom gene set
gene.sets <- list(My_Custom_Pathway = c("GENE1",
                                        "GENE2",
                                        "GENE3",
                                        "GENE4",
                                        "GENE5"))
```

Note: If you would like to run this analysis, you can find the complete code with example dataset in *Session 5 of the CReM scRNAseq Challenge.*