A primer on GSEA + viz

For GSEA, we imagine two analyses. One handles the overview, finding biological pathways of interest given a set of differential expression results. The other gives a more detailed view of an individual pathway.

###High-level overview of GSEA

We are using the "fast" gene set enrichment analysis (fgsea) in the R language. This is effectively a copy of the original GSEA with some algorithmic tweaks to make it run much faster.
https://www.biorxiv.org/content/10.1101/060012v2.full.pdf

The input to a fgsea analysis is the results of a differential gene expression analysis (e.g. DESeq2, etc.). We use those differential expression results to make a ranked list of the genes. Typically, this ranking would be based on something like

$$r_i = -\log(p_i) \cdot \mathrm{sgn}(logFC_i)$$

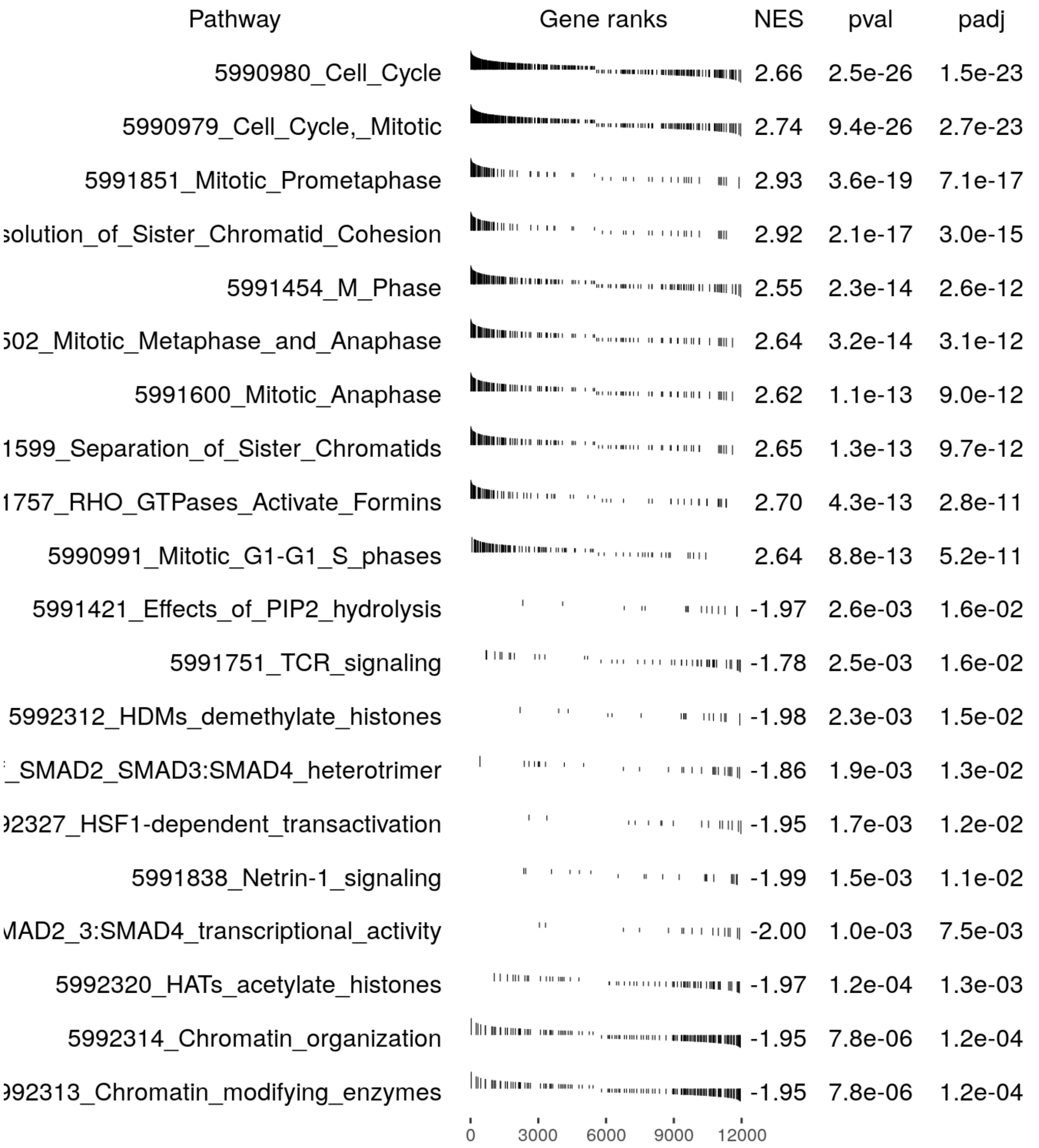
for gene i where p_i is the p-value from the test and $logFC_i$ is the log fold-change; sgn is the "sign" function. This way, we can rank the genes so that the most differentially expressed are at the extremes (top or bottom) of the list.

GSEA also needs a list of "gene sets", which can be things like a set of genes involved in some biological process (i.e. cell death, etc). The idea is that if many genes in those pathways are differentially expressed, we have reason to believe that pathway would be disrupted or changed in some interesting way.

Part 1

For the general analysis, we provide the results of a differential expression analysis, such as DESeq2.

GSEA will then produce a table giving the top pathways, which we could visualize like:



Pathway	Gene ranks	NES	pval	padj
5990980_Cell_Cycle		2.66	2.5e-26	1.5e-23
5990979_Cell_Cycle,_Mitotic		2.74	9.4e-26	2.7e-23
5991851_Mitotic_Prometaphase		2.93	3.6e-19	7.1e-17
solution_of_Sister_Chromatid_Cohesion		2.92	2.1e-17	3.0e-15
5991454_M_Phase		2.55	2.3e-14	2.6e-12
502_Mitotic_Metaphase_and_Anaphase		2.64	3.2e-14	3.1e-12
5991600_Mitotic_Anaphase		2.62	1.1e-13	9.0e-12
1599_Separation_of_Sister_Chromatids		2.65	1.3e-13	9.7e-12
1757_RHO_GTPases_Activate_Formins		2.70	4.3e-13	2.8e-11
5990991_Mitotic_G1-G1_S_phases		2.64	8.8e-13	5.2e-11
5991421_Effects_of_PIP2_hydrolysis		-1.97	2.6e-03	1.6e-02
5991751_TCR_signaling		-1.78	2.5e-03	1.6e-02
5992312_HDMs_demethylate_histones		-1.98	2.3e-03	1.5e-02
_SMAD2_SMAD3:SMAD4_heterotrimer		-1.86	1.9e-03	1.3e-02
92327_HSF1-dependent_transactivation		-1.95	1.7e-03	1.2e-02
5991838_Netrin-1_signaling		-1.99	1.5e-03	1.1e-02
MAD2_3:SMAD4_transcriptional_activity		-2.00	1.0e-03	7.5e-03
5992320_HATs_acetylate_histones		-1.97	1.2e-04	1.3e-03
5992314_Chromatin_organization		-1.95	7.8e-06	1.2e-04
992313_Chromatin_modifying_enzymes		-1.95	7.8e-06	1.2e-04

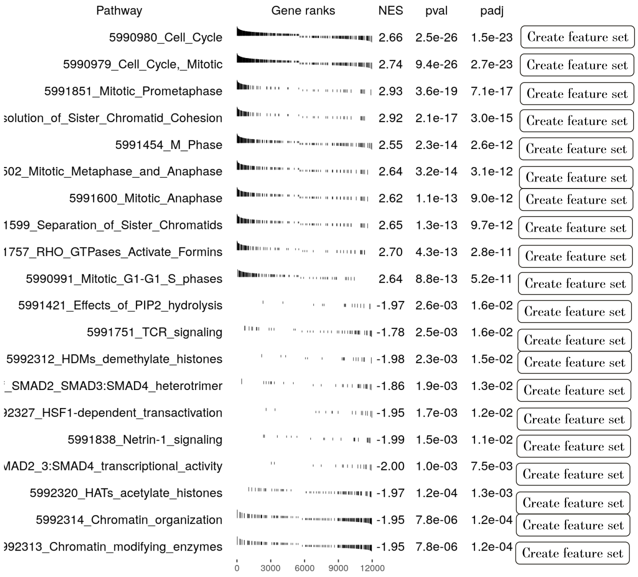
The "rug plots" (in the Gene ranks column) here show where that pathway's genes lie inside the ranked list of ~12,000 genes. The height of the lines is equal to the ranking statistics (the $-\log(p) \cdot \text{sgn}(lfc)$), normalized by the maximum absolute value. Pathways where lots of genes are changing will show a clustering/higher density of lines at the left or right, indicating potential up- or downregulation of that pathway.

The height of the bars really doesn't add anything, so we'll just skip that. The output could be a JSON file with the following structure (or something else that's easier to work with):

```
[
    {
        "pathway": "R-HSA-109581_Apoptosis",
        "pval": 0.494456762749446,
        "padj": 0.733246248679014,
        "log2err": 0.0852884689790502,
        "ES": 0.309457793064953,
        "NES": 0.956948211230173,
        "size": 178,
        "ranks": [
            150,
            256.5,
            267,
            ...
        ],
        "leadingEdge": [
            "51765",
            "5588",
            "3007",
            "3148",
            ...
        ]
    },
    ...
]
```

(The identifiers shown in `leadingEdge` are currently EntrezIDs, but I will work on converting them to 'regular' gene identifiers.)

In addition to displaying the table, we should provide a way to create a `FeatureSet` of the top genes from each pathway (the "leadingEdge" genes). Then, they can use that in combination with an expression matrix to make a boxplot like we have for the DESeq2 output.

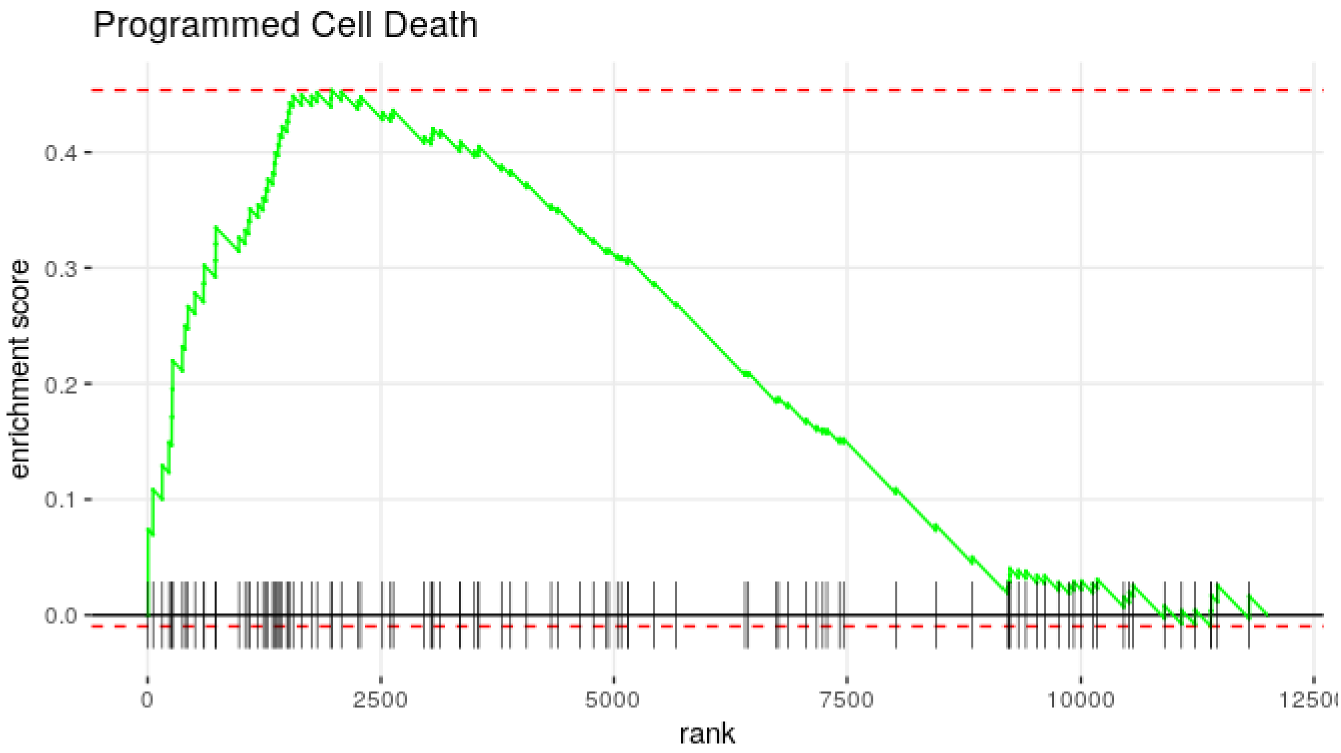


Pathway	Gene ranks	NES	pval	padj	
5990980_Cell_Cycle		2.66	2.5e-26	1.5e-23	Create feature set
5990979_Cell_Cycle,_Mitotic		2.74	9.4e-26	2.7e-23	Create feature set
5991851_Mitotic_Prometaphase		2.93	3.6e-19	7.1e-17	Create feature set
solution_of_Sister_Chromatid_Cohesion		2.92	2.1e-17	3.0e-15	Create feature set
5991454_M_Phase		2.55	2.3e-14	2.6e-12	Create feature set
502_Mitotic_Metaphase_and_Anaphase		2.64	3.2e-14	3.1e-12	Create feature set
5991600_Mitotic_Anaphase		2.62	1.1e-13	9.0e-12	Create feature set
1599_Separation_of_Sister_Chromatids		2.65	1.3e-13	9.7e-12	Create feature set
1757_RHO_GTPases_Activate_Formins		2.70	4.3e-13	2.8e-11	Create feature set
5990991_Mitotic_G1-G1_S_phases		2.64	8.8e-13	5.2e-11	Create feature set
5991421_Effects_of_PIP2_hydrolysis		-1.97	2.6e-03	1.6e-02	Create feature set
5991751_TCR_signaling		-1.78	2.5e-03	1.6e-02	Create feature set
5992312_HDMs_demethylate_histones		-1.98	2.3e-03	1.5e-02	Create feature set
_SMAD2_SMAD3:SMAD4_heterotrimer		-1.86	1.9e-03	1.3e-02	Create feature set
92327_HSF1-dependent_transactivation		-1.95	1.7e-03	1.2e-02	Create feature set
5991838_Netrin-1_signaling		-1.99	1.5e-03	1.1e-02	Create feature set
MAD2_3:SMAD4_transcriptional_activity		-2.00	1.0e-03	7.5e-03	Create feature set
5992320_HATs_acetylate_histones		-1.97	1.2e-04	1.3e-03	Create feature set
5992314_Chromatin_organization		-1.95	7.8e-06	1.2e-04	Create feature set
992313_Chromatin_modifying_enzymes		-1.95	7.8e-06	1.2e-04	Create feature set

###Part 2 (skip for now)

For a more detailed look at an individual pathway, we will provide a second "analysis". The inputs to this will be a differential expression result AND a `FeatureSet`. Since these plots are relatively common in the field, we want to provide a way to produce these.

For a single pathway, the "classic" GSEA figure looks like:



The green line is based on their "enrichment score" (ES), which is kind of a running sum as it goes through the ranked list of all genes. Each time it sees a gene that is in the pathway, it adds to that sum; otherwise, it subtracts. Since many of this pathway's genes are at the top of the ranked list, the green line ascends quickly. The genes to the left of the "peak" are called the "leading edge" genes.

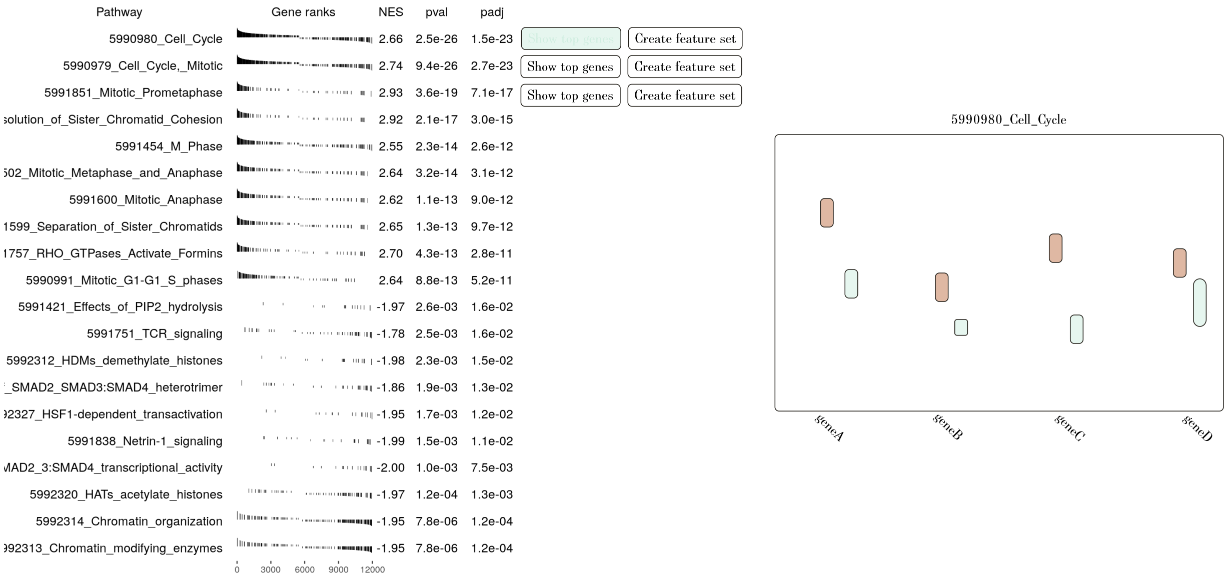
The output from this could look like:

```
{
    "pathway": <str>
    "es_x": [<list of x coords for green ES curve >],
    "es_y": [<list of y coords for green ES curve>]
    "gene_ranks":[<list of x values for the rug plot>]
}
```

Some "nice to have"s:

If we displayed the table shown above, it would be great if we could use that to create a `FeatureSet`. For example, the user could click on a button that says "create new gene set". This would be just like how we can create an `ObservationSet` by highlighting samples in the PCA plot.

Since the GSEA results originate from a specific differential expression analysis, we can use that file to create the plots in the same way we do when we plot after the basic diff. exp. analysis.



Pathway	Gene ranks	NES	pval	padj		
5990980_Cell_Cycle		2.66	2.5e-26	1.5e-23	Show top genes	Create feature set
5990979_Cell_Cycle,_Mitotic		2.74	9.4e-26	2.7e-23	Show top genes	Create feature set
5991851_Mitotic_Prometaphase		2.93	3.6e-19	7.1e-17	Show top genes	Create feature set
solution_of_Sister_Chromatid_Cohesion		2.92	2.1e-17	3.0e-15		
5991454_M_Phase		2.55	2.3e-14	2.6e-12		
502_Mitotic_Metaphase_and_Anaphase		2.64	3.2e-14	3.1e-12		
5991600_Mitotic_Anaphase		2.62	1.1e-13	9.0e-12		
1599_Separation_of_Sister_Chromatids		2.65	1.3e-13	9.7e-12		
1757_RHO_GTPases_Activate_Formins		2.70	4.3e-13	2.8e-11		
5990991_Mitotic_G1-G1_S_phases		2.64	8.8e-13	5.2e-11		
5991421_Effects_of_PIP2_hydrolysis		-1.97	2.6e-03	1.6e-02		
5991751_TCR_signaling		-1.78	2.5e-03	1.6e-02		
5992312_HDMs_demethylate_histones		-1.98	2.3e-03	1.5e-02		
_SMAD2_SMAD3:SMAD4_heterotrimer		-1.86	1.9e-03	1.3e-02		
92327_HSF1-dependent_transactivation		-1.95	1.7e-03	1.2e-02		
5991838_Netrin-1_signaling		-1.99	1.5e-03	1.1e-02		
MAD2_3:SMAD4_transcriptional_activity		-2.00	1.0e-03	7.5e-03		
5992320_HATs_acetylate_histones		-1.97	1.2e-04	1.3e-03		
5992314_Chromatin_organization		-1.95	7.8e-06	1.2e-04		
92313_Chromatin_modifying_enzymes		-1.95	7.8e-06	1.2e-04		