sSNAPPY: an R/Bioconductor package for single-sample directional pathway perturbation analysis

Wenjun Liu[*1], Ville-Petteri Mäkinen[2,3], Wayne D. Tilley[1], and Stephen M. Pederson[1,4,5]

[1]Dame Roma Mitchell Cancer Research Laboratories, Adelaide Medical School, Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, Australia
[2]Computational Medicine, Faculty of Medicine, University of Oulu, Oulu, Finland
[3]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland
[4]Black Ochre Data Laboratories, Telethon Kids Institute, Adelaide, Australia
[5]John Curtin School of Medical Research, Australian National University, Canberra, Australia

Abstract A common outcome of analysing RNA-Seq data is the detection of biological pathways with significantly altered activity between the conditions under investigation. Whilst many strategies test for over-representation within pre-defined gene-sets for genes showing changed expression, these analyses typically do not account for gene-gene interactions encoded by pathway topologies, and are not able to directly predict the directional change of pathway activity. To address these issues, we have deloped a single-sample pathway perturbation analysis method *sSNAPPY*, now available as an R/Bioconductor package, which leverages pathway topology information to compute pathway perturbation scores, and predicts the direction of change across a set of pathways. Here, we demonstrate the use of *sSNAPPY* by applying the method to public scRNA-seq data, derived from ovarian cancer patient tissues collected before and after chemotherapy. Not only were we able to predict the directions of significant perturbations of pathways discussed in the original study, but *sSNAPPY* was also able to detect significant changes of other biological processes, yielding far greater insight into the response to treatment. *sSNAPPY* represents a novel pathway analysis strategy that takes into consideration of pathway topology to predict impacted biology pathways, both within related samples and across treatment groups. In addition to not relying on the detection of differentially expressed genes, the method and associated R package offer important flexibility and provide powerful visualisation tools.

Keywords

RNA-seq, pathway enrichment, R package, topology, KEGG, scRNA-seq

---
*Corresponding Author wenjun.liu@adelaide.edu.au

R version: R version 4.3.0 (2023-04-21)

Bioconductor version: 3.17

Package: 1.4.1

Introduction

Using pathway enrichment analysis to gain biological insights from gene expression data is a pivotal step in the analysis and interpretation of RNA-seq data, for which numerous methods have been developed (reviewed in [1, 2]). Many existing methods tend to view pathways simply as a collection of gene names, as seen in those relying on the detection of differentially expressed genes and applying over-representation analysis (ORA) strategies, and those scoring all genes using functional class scoring (FCS), such as in Gene Set Enrichment Analysis (GSEA) [3], arguably the most widely-used approach. However, databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[4] and WikiPathways[5] capture not only which genes are implicated in a certain biological process but also their interactions, activating or inhibitory roles, and their relative importance within the pathway, all of which are overlooked in ORA- and FCS-based approaches.

To fully utilise that additional information, the latest generation of pathway analysis approaches include many which are topology-based such as SPIA[6], DEGraph[7], NetGSA[8] and PRS[9], as well as others which explicitly model inter-gene correlations[10]. Despite differences in the null hypotheses tested across these approaches, overall, they have demonstrated enhanced sensitivity and specificity due to their abilities to take gene-gene interconnections into account[11, 12]. Nevertheless, most topology-based methods focus only on comparing activities of pathways between two treatment groups and cannot be used to score individual samples (Figure 1). However, in heterogenous data where more than one factor may be influencing observations[13], incorporating scoring within paired samples may be desirable and may be able to reveal more nuanced insights. To address this gap, we present a Single-Sample directioNAl Pathway Perturbation analYsis methodology called *sSNAPPY*, available as an R/Bioconductor package. This article defines how *sSNAPPY* computes changes in gene expression within paired samples, and propagates this through gene-set topologies to predict the perturbation in pathway activities within paired samples, before providing summarised results across an entire dataset (Figure 1). The practical usage of the *sSNAPPY* R/Bioconductor package is illustrated through the analysis of a public scRNA-seq dataset using pseudo-bulk strategies.

[Figure 1 about here.]

Methods

Implementation

sSNAPPY is an R package that has been reviewed and published on the open-source bioinformatics software platform Bioconductor with all source code available via GitHub. The methodology itself is topology-based, designed to compute directional, single-sample, pathway perturbation scores in gene expression datasets with a matched-pair, or nested design (eg. samples collected before and after treatment). This allows for the detection of pathway perturbations within all samples from a treatment group, but also within individual samples.

To run *sSNAPPY*, the only required data is a log-transformed expression matrix (e.g. logCPM) with matching sample metadata describing treatment groups and the nested structure. It is assumed that all pre-processing has been performed beforehand, such as the exclusion of low-signal genes or normalisation to minimise technical artefacts like GC-bias. The first step, performed internally by *sSNAPPY*, is to estimate sample-specific log fold-change ($\delta_{gi} = \mu_{gi} - \mu_{gi'}$) across all genes g within each treated sample i, by subtracting expression estimates for the corresponding baseline samples $\mu_{gi'}$ from those in the treated groups. *sSNAPPY* supports any number of treatment levels and the numbers of samples in each treatment group do not have to be identical. As long as a treated sample is matched to a corresponding control, the sample-specific log fold-changes will be computed for that sample.

Since it has been shown that in RNA-seq data, genes with lower expression tend to have larger variance and larger estimates of change[14], we utilise a gene-level weighting strategy to de-emphasise logFC estimates for low-abundance genes. Gene-level weights w_g are obtained in a treatment-agnostic manner by fitting a loess curve through the relationship between observed gene-level variance (σ_g) and average signal ($\bar{\mu}_{g..}$)(Figure 2), and taking the inverse of the loess-predicted variance as the weight $w_g = a/f(\bar{\mu}_{g..})$, where $f(\bar{\mu}_{g..})$ is the predicted value from the loess curve and the constant a ensures $\sum w_g = 1$. We then use these weighted estimates of logFC ($\delta_{gi}^* = w_g \delta_{gi}$) in the calculation of all pathway perturbation scores.

[Figure 2 about here.]

sSNAPPY was built upon the group-level topology-based scoring algorithm initially proposed in SPIA[6] to propagate genes' changes in expression through pathway topologies to compute a perturbation score for each pathway. In contrast to SPIA, which requires pre-selection of differentially expressed genes, *sSNAPPY* utilises the full expression matrix. By modifying the algorithm to incorporate single-sample, weighted estimates of changes in expression we are able to quantify changes in a pathway within a given sample, and then model these across all samples within a treatment group. Thus, we define the single-sample perturbation score (S_{ip}) for a given pathway p in a treated sample i:

$$S_{ip} = \sum_{g \in G_p} [S_{gip} - \delta_{gi}^*], \text{ where}$$

$$S_{gip} = \delta_{gi}^* + \sum_{g' \in U_{gp}} \beta_{gg'p} \frac{S_{g'ip}}{N_{g'p}}$$

where:

- G_p represents the set of genes in pathway p, such that $g \in G_p$

- S_{gip} is the gene- and sample-specific perturbation score for pathway p

- $\delta_{gi}^* = w_g \delta_{gi}$ is the weighted logFC of gene g in sample i as described above

- U_{gp} is the subset of G_p containing only the genes directly upstream of gene g

- $\beta_{gg'p}$ is the pair-wise gene-gene interactions[6] encoded by the topology matrix for genes g and g'

- N_{gp} is the number of downstream genes from any gene g

- S_{ip} is the accumulated pathway perturbation score for pathway p in sample i

The Bioconductor package graphite[15] provides functions that can be used to retrieve pathway topologies from a database and convert topology information to adjacency matrices. In order to streamline this process we have implemented a convenience function, where users only need to provide the name of the desired database and species to retrieve all topology information in the format required by the scoring algorithm with the correct type of gene identifiers (ie. Entrez ID).

To scale the single-sample pathway perturbation scores (S_{ip}) so they are comparable across pathways and to test for significance of individual scores, null distributions of perturbation scores for each pathway are generated through a sample permutation strategy, retaining the correct correlation structure between genes within a pathway. We recommend users to perform a minimum of 1000 permutations, requiring at least 8 unique samples. Subsequently, the median and median absolute deviation (MAD) of the permuted perturbation scores will be calculated and used to normalise the raw perturbation scores to robust Z-scores ($Z_{ip} = \frac{S_{ip} - median_{n \in N, i \in I}(S_{i^n p})}{MAD_{n \in N, i \in I}(S_{i^n p})}$) and obtain associated two-sided p-values. With each permutation, for example the n-th round, column names (i.e. sample labels) for the expression matrix are randomly shuffled while the rest of the scoring algorithm remains unchanged, resulting in one permuted score being computed for each pathway p in each permuted sample i^n. Performing a total of N rounds of permutations on a dataset where the total number of treated sample equals I, $N \times I$ permutated scores will be generated for each pathway p.

Apart from assessing whether a pathway's activity changed significantly within an individual sample, users may also be interested in detecting changes at the group-level, which can be performed by modelling scores with regression models, incorporating Smyth's moderated t-statistic[16] as implemented in *limma*[17]. The single-sample nature of *sSNAPPY*'s pathway perturbation scores is particularly helpful for datasets with complex experimental designs or known confounding factors as these can also be incorporated into the final regression models.

Operation

The package has been tested on all operating systems, requiring R > 4.2.0, and can be installed using BiocManager as follows.

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("sSNAPPY")
```

Use Cases

Data

To demonstrate the application of *sSNAPPY*, we used pre-processed counts from a publicly available scRNA-seq dataset, retrieved from Gene Expression Omnibus (GEO) with accession code GSE165897. This dataset consists of 11 high-grade serous ovarian cancer (HGSOC) patients samples taken before and after chemotherapy[18]. *sSNAPPY* was used to re-analyse data from the epithelial cells as they were the primary focus of the original study. Since *sSNAPPY* was designed for bulk RNA-seq data, counts of epithelial cells from the same samples were first summed into pseudo-bulk profiles, giving rise to a total of 22 samples. We considered a gene detectable if we observed >1.5 counts per million in >11 samples out of 22, representing all samples from a complete treatment group. 11,101 (33.8%) of the 32,847 annotated genes passed this selection criteria and were included for downstream analyses. Conditional quantile normalisation[19] was then applied to mitigate potential biases introduced by gene length and GC content. The normalised logCPM matrix of the processed dataset and sample metadata can be downloaded from here.

The following packages are required for this workflow

```
library(sSNAPPY)
library(tidyverse)
library(magrittr)
library(ggplot2)
library(patchwork)
library(kableExtra)
library(AnnotationHub)
library(edgeR)
library(patchwork)
library(colorspace)
```

Before running the *sSNAPPY* workflow, we need to load our expression matrix and define our sample-level metadata, then download any pathway information. Firstly we can read in the data, setting the treatment column in the metadata to be a factor. Importantly, to run sSNAPPY, the row names of the expression matrix must be specified as Entrez IDs, for compatibility with pathway databases. Genes without Entrez IDs were excluded, leaving 10,098 genes in the example expression matrix.

```
logCPM <- read_tsv(here::here("data/logCPM.tsv")) %>%
    column_to_rownames("entrezid")
sample_meta <- read_tsv(here::here("data/sample_meta.tsv"), col_types = "cfccncnc")
head(sample_meta)
```

```
## # A tibble: 6 x 8
##   sample        treatment patient_id anatomical_location   Age Stage   PFI CRS
##   <chr>         <fct>     <chr>      <chr>               <dbl> <chr> <dbl> <chr>
## 1 EOC372_treat~ treatmen~ EOC372     Peritoneum             68 IIIC    460 1
## 2 EOC372_post-~ post-NACT EOC372     Peritoneum             68 IIIC    460 1
## 3 EOC443_post-~ post-NACT EOC443     Omentum                54 IVA     177 3
## 4 EOC443_treat~ treatmen~ EOC443     Omentum                54 IVA     177 3
## 5 EOC540_treat~ treatmen~ EOC540     Omentum                62 IIIC    126 2
## 6 EOC540_post-~ post-NACT EOC540     Omentum                62 IIIC    126 2
```

Retrieval of Pathway Topology

Next, pathway topology information needs to be retrieved from a chosen database. Using KEGG as an example, the retrieved topology information will be stored as a list where each element corresponds to a pathway and the numbers in the matrices encode gene-gene interaction.

```
gsTopology <- retrieve_topology(database = "kegg", species = "hsapiens")
```

Instead of downloading the topology matrices of all pathways, it is also possible to provide a restricted set of keywords for a targeted analysis. In the following we'll use the keywords 'metabolism' and 'estrogen' to only return a subset of pathways.

```
# Only retrieve the topology matrices of metabolism- or signalling-related pathways
gsTopology_sub <- retrieve_topology(database = "kegg", species = "hsapiens",
                                    keyword = c("metabolism", "estrogen"))
names(gsTopology_sub)
```

In addition to working with a single database, users can also choose multiple databases for more comprehensive analyses, albeit with an increased analytic and download time.

```
gsTopology_2databases <- retrieve_topology(database = c("kegg", "wiki"),
                                           species = "hs")
```

Score Single-Sample Pathway Perturbation

To compute the single-sample expression changes (i.e. logFC) needed for perturbation scores, samples must be in matched pairs. For instance, biopsies derived pre- vs post-treatment or untreated vs treated cell lines. The factor defining the paired structure needs to be passed to the `weight_ss_fc()` function through the `groupBy` parameter. In our example dataset, pre- and post-treatment samples are matched by patient IDs. Additionally, the sample metadata must include the treatment of all samples. The treatment column must be a factor with the reference level set to be the control treatment.

```
weightedFC <- weight_ss_fc(as.matrix(logCPM), metadata = sample_meta,
 sampleColumn = "sample", groupBy = "patient_id", treatColumn = "treatment")
glimpse(weightedFC)
```

```
## List of 2
##  $ weight: num [1:10098] 9.38e-05 1.19e-04 7.63e-05 1.18e-04 1.21e-04 ...
##  $ logFC : num [1:10098, 1:11] -1.52e-05 -2.39e-05 2.00e-04 -1.53e-04 3.84e-05 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:10098] "ENTREZID:643837" "ENTREZID:26155" "ENTREZID:84069" "ENTREZID:57801" ..
##   .. ..$ : chr [1:11] "EOC372_post-NACT" "EOC443_post-NACT" "EOC540_post-NACT" "EOC3_post-NACT" .
```

The output of `weight_ss_fc()` is a `list` where one element is a matrix of weighted single-sample logFCs (δ^*_{ghi}), with rows corresponding to genes and columns to samples, and the other element is the vector of gene-wise weights (w_g) used to calculate the weighted logFC (δ^*_{ghi}), as described above.

The matrix of δ^*_{ghi} values are then passed to pathway topologies to compute the gene-wise perturbation scores for all genes included in a pathway, before being summed into a single score for each pathway. The string `ENTREZID:` was also added to all row names of the δ^*_{ghi} matrix to be compatible with the format pathway topologies were retrieved in. These gene-wise perturbation scores can also be used in downstream analysis to identify genes playing the most significant roles in each pathway, as will be demonstrated in the visualisation section below. The pathway-level perturbation scores (S_{hip}) are returned as a data.frame containing sample and gene-set names. In the following steps, we first calculate the gene-level contributions to each pathway (S_{ghip}) using the function `raw_gene_pert()` and then obtain pathway-level summaries using `pathway_pert()`. Pathways with zero perturbation scores across all genes and samples will be dropped at this stage.

```
genePertScore <- raw_gene_pert(weightedFC$logFC, gsTopology)
ssPertScore <- pathway_pert(genePertScore)
head(ssPertScore)
```

```
##             sample        score                                  gs_name
## 1 EOC372_post-NACT  0.005134216 kegg.EGFR tyrosine kinase inhibitor resistance
## 2 EOC443_post-NACT -0.001959631 kegg.EGFR tyrosine kinase inhibitor resistance
## 3 EOC540_post-NACT -0.006543325 kegg.EGFR tyrosine kinase inhibitor resistance
## 4   EOC3_post-NACT -0.003840943 kegg.EGFR tyrosine kinase inhibitor resistance
## 5  EOC87_post-NACT -0.003917201 kegg.EGFR tyrosine kinase inhibitor resistance
## 6 EOC136_post-NACT -0.008308897 kegg.EGFR tyrosine kinase inhibitor resistance
```

Sample Permutation for Normalisation and Significance Testing

The values obtained from each pathway will vary greatly due to the variability in topology structures. To determine the significance of individual scores and transform scores to ensure they are comparable across pathways, sSNAPPY utilises a sample-permutation strategy to simulate the null distributions of perturbation scores. Since sample labels will be permuted randomly, sample metadata is not required by the `generate_permuted_scores()` function, instead, users only need to specify the number of treatment groups in the study, including the control level. Since permutation requires a large amount of computational time and memory, sSNAPPY utilises the BiocParallel backend[20], also allowing for customisation by the user. Paralleled with 8 cores, permutating 1000 times took approximately 30 minutes to complete on a local laptop. Whilst permutations can

be performed an a local machine, this strategy also allows for performing the permutation steps on a HPC cluster or similar.

```
set.seed(123)
permutedScore <- generate_permuted_scores(
    as.matrix(logCPM), numOfTreat = 2, NB = 1000,
    gsTopology = gsTopology, weight = weightedFC$weight
)
```

Apart from pathways whose permuted perturbation scores are all zero, the rest of the pathways' empirical distributions should be approximately normally distributed with means equal to zero. To demonstrate that, we randomly selected 6 pathways and visualised the permuted perturbation scores as boxplots (Figure 3).

```
set.seed(234)
permutedScore %>%
    keep(~all(.!=0)) %>%
    .[sample(seq_along(.), 6)] %>%
    as.data.frame() %>%
    pivot_longer(
      cols = everything(), names_to = "gs_name", values_to = "score"
    ) %>%
    mutate(
        gs_name = str_replace_all(gs_name, "\\.", " "),
        gs_name = str_remove_all(gs_name, "kegg ")
    ) %>%
    ggplot(aes(gs_name, score, fill = gs_name)) +
    geom_boxplot() +
    scale_x_discrete(labels = function(x) str_wrap(x, width = 10)) +
    scale_fill_discrete(name = "Gene-set Name") +
    labs(x = "Pathway", y = "Permuted Perturbation Score") +
    theme_bw() +
    theme(
      legend.position = "none",
      axis.title = element_text(size = 16),
      axis.text = element_text(size = 14)
    )
```

[Figure 3 about here.]

Permuted distributions are used to convert each pathway-level score into a scaled robust *Z*-score using the function `normalise_by_permu()`. Robust Z-scores can in turn be transformed into two-sided p-values and corrected for multiple testing using any of the available methods, and returning the FDR adjusted values by default. In our brief example data, no pathways would be considered as significantly perturbed at the single-sample level using an FDR adjustment with $\alpha = 0.05$.

```
normalisedScores <- normalise_by_permu(permutedScore, ssPertScore,
                                        sortBy = "pvalue")
head(normalisedScores)
```

```
##              MAD        MEDIAN                                         gs_name
## 388   0.0275353965  1.538283e-04               kegg.Epstein-Barr virus infection
## 417   0.0002129193  7.921798e-07                 kegg.Autoimmune thyroid disease
## 421   0.0002129193  7.921798e-07                        kegg.Allograft rejection
## 1142  0.0034661258 -1.673836e-06 kegg.C-type lectin receptor signaling pathway
## 505   0.0098582420 -1.139206e-04                   kegg.Th17 cell differentiation
## 1530  0.0036864141  2.507001e-05                           kegg.Autophagy - animal
##                 sample          score     robustZ      pvalue  adjPvalue
## 388   EOC136_post-NACT -0.0817379496 -2.974055 0.002938926 0.2392677
## 417   EOC136_post-NACT -0.0006237555 -2.933261 0.003354221 0.2392677
## 421   EOC136_post-NACT -0.0006237555 -2.933261 0.003354221 0.2392677
## 1142  EOC349_post-NACT  0.0085037555  2.453872 0.014132724 0.9830259
## 505   EOC153_post-NACT  0.0238610587  2.431973 0.015016823 0.7618825
## 1530  EOC443_post-NACT  0.0087762111  2.373890 0.017601798 1.0000000
```

A key question of interest in this dataset is to identify which biological processes were impacted by chemotherapy across the entire group of patients. Using the sample-level output obtained above, we can explore this by applying t-tests or regression models across all samples. In order to minimise spurious results, Smyth's moderated t-statistics[16] are able to be applied across the complete dataset. Given that we have robust Z-score, a constant variance was assumed across all pathways. To perform this analysis, the robust Z-scores were converted to a matrix, and the standard limma methodologies was used with variance set to be constant. For our use case here, no design matrix is required and a simple t-test is appropriate.

```
z_matrix <- normalisedScores %>%
  dplyr::select(robustZ, gs_name, sample) %>%
  pivot_wider(names_from = "sample", values_from = "robustZ") %>%
  column_to_rownames("gs_name") %>%
  as.matrix()
z_fits <- lmFit(z_matrix, design = rep(1, ncol(z_matrix))) %>%
  eBayes(trend = FALSE)
top_table <- topTable(z_fits, number = Inf) %>%
  as_tibble(rownames = "gs_name")
```

Pathways with an FDR < 0.05 in the moderated t-test were considered to be significantly perturbed at the group level. 11 out of the 315 tested KEGG pathways passed the selection threshold (Table 1).

```
table1 <- top_table %>%
  dplyr::filter(adj.P.Val < 0.05) %>%
  mutate(
    Direction = ifelse(logFC < 0, "Inhibited", "Activated"),
    gs_name = str_remove_all(gs_name, "kegg.")
  ) %>%
  dplyr::select(
    Pathway = gs_name, Change = logFC,  P.Value, FDR = adj.P.Val, Direction
  )
```

[Table 1 about here.]

For enrichment analysis in the original study[18], unsupervised clustering was performed on all cells labelled as cancer cells. Clusters were then annotated manually by performing pathway enrichment testing on cluster marker genes. Two clusters, associated with proliferative DNA repair signatures and stress-related markers, contained significantly higher numbers of post-chemotherapy cells than pre-treatment ones[18]. The representative pathways enriched in the stress-associated cluster were *IL6-mediated signaling events*, *TNF signaling pathway*, and *cellular responses to stress*, characterized by marker genes *JUN, FOS, IL6, TNF, CXCR4, SNAI1, VIM, GADD45B*, and *MCL1*. Among the stress-related marker genes reported, 6 of them(*CXCR4, FOS, GADD45B, IL6, JUN*, and *TNF*) were implicated in pathways that were considered to be significantly impacted by sSNAPPY. The other post-chemotherapy cell dominated cluster in the original study was enriched for pathways associated with cell proliferation and DNA repair, such as the Cell cycle, DNA repair, Homology directed repair (HDR) through homologous recombination, and Fanconi anaemia pathway. A key gene involved in those pathways was *CHEK1*, which was also found in significantly perturbed pathways detected by sSNAPPY: the p53 signaling pathway, Cellular senescence and the Human T-cell leukaemia virus 1 infection pathway.

Apart from considering all treated samples as biological replicates, users may elect to perform an analysis incorporating other phenotypic traits that affect patients' responses to chemotherapy, such as disease stages or tumour grades. To do that through the moderated t-statistic strategy and extend the above analysis, all that is required is an appropriate design matrix in the `lmFit` step, or subsetting the samples accordingly

Visualising Perturbed Pathways as Networks

A valuable feature of sSNAPPY is the provision of various visualisation functions to assist in the interpretation of results. Biological pathways are not independent of each other with many genes playing a role across multiple pathways, and as such, visualising pathway analysis results as a network can be a powerful way to intuitively summarise the results and facilitate interpretation of the underlying biology. The `plot_gs_network()` function allows users to easily convert a list of perturbed biological pathways to a network where edges between pathway nodes represent overlapping genes. Defined by the colorBy parameter, pathway nodes can be coloured by either the pathways' predicted direction of changes or the significance levels (Figure 4). The returned plot is a ggplot2 [21] object, meaning that components of the plotting theme and other parameters can be customized as for any other ggplot2 objects.

First we should obtain a subset of pathways to visualise. In the following example, the most highly ranked pathways were inspected, which involved three steps to prepare the data: 1) create a categorical variable with the pathway status, 2) rename the logFC column to reflect the true meaning of the value and, 3) transform the p-values for simpler visualisation.

```r
sigPathway <- topTable(z_fits, number = 8) %>%
  rownames_to_column("gs_name") %>%
  dplyr::rename(Z = logFC) %>%
  mutate(
    status = case_when(
      adj.P.Val > 0.05 ~ "Unchanged",
      Z > 0 ~ "Activated",
      Z < 0 ~ "Inhibited"
    ),
    status = as.factor(status),
    `-log10(p)` = -log10(P.Value)
  )
```

```r
set.seed(123)
# Plot the network structure
p1 <- plot_gs_network(
  normalisedScores = sigPathway, gsTopology = gsTopology, colorBy = "status"
) +
  scale_colour_manual(values = c("red", "blue", "grey30")) +
  theme_void()
set.seed(123)
p2 <- plot_gs_network(
  normalisedScores = sigPathway,
  gsTopology = gsTopology,
  colorBy  = "-log10(p)",
  color_lg_title = expression(paste(-log[10], "p"))
) +
  scale_colour_viridis_c() +
  theme_void()
p1 + p2 + plot_annotation(tag_levels = "A")
```

[Figure 4 about here.]

By examining the network structure, we can see that many of the highly connected pathways playing a central role in the network are immune-related (Figure 4). To summarise related pathways and further enable interpretation, we can apply community detection[22] in order to group related pathways into 'communities'. sSNAPPY's `plot_community()` function is a one-stop shop for applying a community detection algorithm of the user's choice to the network structure and annotating identified communities by the most common pathway category, denoting the main biological processes perturbed in that community. The most recent categories for KEGG pathways were curated from the KEGG website and included as part of sSNAPPY. Annotation of KEGG pathway communities is automatically completed by calling the in-built data object. Analyses involving other pathway databases will require user-provided pathway categories. In the current dataset, the Louvain method was applied to the network of biological pathways and revealed two primary communities, where one was annotated to be endocrine system related and the other one was most closely related to infectious diseases and the immune system (Figure 5).

```r
set.seed(123)
plot_community(
    normalisedScores = sigPathway,
    gsTopology = gsTopology,
    colorBy = "status"

) +
    scale_colour_manual(values = c("red", "blue")) +
    scale_fill_viridis_d() +
    scale_x_continuous(expand = expansion(0.25)) +
    scale_y_continuous(expand = expansion(0.25)) +
    theme_void()
```

[Figure 5 about here.]

A key advantage of sSNAPPY is that it does not require the prior identification of differentially expressed genes, as this is not always possible in clinical datasets. However, knowing which genes are implicated in the perturbation of pathways, particularly those which influence multiple pathways, can provide valuable insights for hypotheses generation about underlying biological mechanisms. Therefore, sSNAPPY provides another visualisation feature called `plot_gs2gene`, which enables the inclusion of select genes from each pathway using network structures. Users can provide a vector of fold-change estimates to visualise genes within pathways, showing their estimated change in expression. As pathways often include hundreds of genes, it is recommended to filter for genes most likely to be playing a significant role. In this example dataset, only genes within the top 500 when ranking by the magnitude of the mean log fold-change were included (Figure 6). An alternative strategy will be to select genes based on test-statistics, however, this decision is up to the individual researcher.

```
meanFC <- rowMeans(weightedFC$logFC) / weightedFC$weight
top500 <- rank(1/abs(meanFC)) <= 500
dirFC <- ifelse(meanFC > 0, "Up-Regulated", "Down-Regulated")
```

Since KEGG pathway topologies were retrieved using EntrezIDs, users can provide a data.frame mapping Entrez IDs to their chosen identifiers, such as gene names, through the mapEntrezID parameter to make the visualisations more informative. A data.frame converting Entrez IDs to ensemble gene names was derived from the Ensembl Release 101[23] and has been made available as part of the package.

```
load(system.file("extdata", "entrez2name.rda", package = "sSNAPPY"))
head(entrez2name)
```

```
## # A tibble: 6 x 2
##   entrezid          mapTo
##   <chr>             <chr>
## 1 ENTREZID:84771    DDX11L1
## 2 ENTREZID:727856   DDX11L1/DDX11L9/DDX11L10
## 3 ENTREZID:100287102 DDX11L1
## 4 ENTREZID:100287596 DDX11L1/DDX11L9
## 5 ENTREZID:102725121 DDX11L1
## 6 ENTREZID:653635   WASH7P
```

```
set.seed(123)
plot_gs2gene(
    normalisedScores = sigPathway,
    gsTopology = gsTopology,
    colorGsBy = "status",
    mapEntrezID = entrez2name,
    geneFC = meanFC[top500],
    edgeAlpha = 1,
    gsNameSize = 4,
    gsNodeSize = 4
) +
    scale_colour_gradient2(name = "logFC") +
    scale_fill_manual(values = c("red", "blue", "grey50")) +
    theme_void()
```

[Figure 6 about here.]

Identifying Key Gene Contributions

To further investigate a specific pathway and elucidate the key genes that contribute to its perturbation, we can generate a heatmap via `plot_gene_contribution()` which shows the gene-level perturbation scores for the top-ranked members of a given pathway. This function takes advantage of the plotting capabilities of the `pheatmap` package[24], and as such, other annotations are also able to be easily included, such as the disease stage, or which general ranges the pathway-level normalised Z-Scores are in. Inclusion of the Z-Scores enabled the assessment of the level of perturbation predicted in each sample and key genes involved (Figure 7).

```r
annotation_df <- normalisedScores %>%
    dplyr::filter(gs_name == "kegg.p53 signaling pathway") %>%
    left_join(dplyr::select(sample_meta, sample, Stage), by = "sample") %>%
    mutate(
      `Z Range` = cut(
        robustZ, breaks = seq(-2, 2, length.out = 6), include.lowest = TRUE
      ),
      sample = str_remove_all(sample, "_post-NACT")
    ) %>%
    dplyr::select(sample, `Z Range`, Stage)
z_levels <- levels(annotation_df$`Z Range`)
annotation_col <- list(
  Stage = c(IIIC = "#4B0055", IVA = "#009B95", IVB = "#FDE333"),
  `Z Range` = setNames(
    colorRampPalette(c("navyblue", "white", "darkred"))(length(z_levels)),
    z_levels
  )
)
plot_gene_contribution(
    genePertMatr = genePertScore$`kegg.p53 signaling pathway` %>%
        set_colnames(str_remove_all(colnames(.), "_post-NACT")),
    color = rev(colorspace::divergex_hcl(100, palette = "RdBu")),
    breaks = seq(-0.0015, 0.0015, length.out = 100),
    annotation_df =  annotation_df,
    topGene = 20, filterBy = "mean",
    mapEntrezID = entrez2name,
    annotation_colors = annotation_col,
    cutree_rows = 3, cutree_cols = 3,
    main = "P53 Signaling Pathway [KEGG]"
)
```

[Figure 7 about here.]

From this heatmap we can identify candidate genes which are likely to be making the biggest contribution to the activation of p53 Signaling Pathway upon chemotherapy, such as *ART* and *ATM*, along with the *MDM4* and members of the GADD45 family. Ataxia-telangiectasia mutated (*ATM*) is a well-established oncosuppressor[25] while haploinsufficient mouse double minute 4 (*MDM4*) is an oncogene that suppresses p53[26]. Mutations of both genes have been observed in many types of cancers[27, 28, 29]. Moreover, the Ataxia telangiectasia and RAD3-related protein kinase (*ATR*) and the growth arrest and DNA damage-inducible gene 45 (*GADD45*) have gained great attention as potential targets for anti-tumor treatments due to their involvement in DNA damage repair[30, 31]. ATR inhibitors, in particular, have shown promising results in clinical trials for HGSOC.[32, 33].

Discussion

In conclusion, we have presented and provided a demonstration for the R/Bioconductor package *sSNAPPY* which offers a novel single-sample pathway perturbation testing approach, tailored for heterogeneous tissue samples in matched-pair design. In contrast to many common enrichment methods, sSNAPPY uses pathway topology information to compute perturbation scores which indicate the likely impact on the activity of a pathway, by predicting direction of change. By applying sSNAPPY to a public scRNA-seq data collected before and after HGSOC patients were subjected to chemotherapy, we demonstrated its ability to detect significant pathway perturbations of various interesting biological processes consistent with, and far beyond what was shown in the original study. *sSNAPPY* addresses the limitations of alternative strategies that fail to account for gene-gene interactions encoded by pathway topologies or predict the directionality of pathway activities. In addition, the single-sample nature of the method can be utilised to address the increasing demand for personalised medicine. Through identifying shared and divergent responses between individuals, *sSNAPPY* can provide valuable insights into the heterogeneous responses across clinical samples. Overall, we believe sSNAPPY represents a valuable addition to the existing body of pathway analysis methods.

Data availability

The dataset analysed in this manuscript are stored in the data directory of this GitHub repository.

Software availability

- Software available from: `https://bioconductor.org/packages/release/bioc/html/sSNAPPY.html`

- Source code available from: `https://github.com/Wenjun-Liu/sSNAPPY`

- Archived source code at time of publication: [DOI (found on right hand side of a Zenodo record)]

- License: MIT

Competing interests

No competing interests were disclosed

Grant information

Acknowledgements

References

[1] Farhad Maleki, Katie Ovens, Daniel J Hogan, and Anthony J Kusalik. Gene set analysis: Challenges, opportunities, and future research. *Front. Genet.*, 11:654, June 2020.

[2] Sarah Mubeen, Alpha Tom Kodamullil, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. On the influence of several factors on pathway enrichment analysis. *Brief. Bioinform.*, 23(3), May 2022.

[3] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102 (43):15545–15550, October 2005.

[4] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes.

[5] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, January 2021. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaa1024.

[6] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, January 2009. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btn577.

[7] Laurent Jacob, Pierre Neuvial, and Sandrine Dudoit. More power via graph-structured tests for differential expression of gene networks. *The Annals of Applied Statistics*, 6(2):561 – 600, 2012. doi: 10.1214/11-AOAS528. URL `https://doi.org/10.1214/11-AOAS528`.

[8] Jing Ma, Ali Shojaie, and George Michailidis. Network-based pathway enrichment analysis with incomplete network information. *Bioinformatics*, 32(20):3165–3174, 06 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw410. URL `https://doi.org/10.1093/bioinformatics/btw410`.

[9] Maysson Al-Haj Ibrahim, Sabah Jassim, Michael Anthony Cawthorne, and Kenneth Langlands. A topology-based score for pathway enrichment. *Journal of Computational Biology*, 19(5):563–573, 2012. doi: 10.1089/cmb.2011.0182. URL `https://doi.org/10.1089/cmb.2011.0182`. PMID: 22468678.

[10] Di Wu and Gordon K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133–e133, 05 2012. ISSN 0305-1048. doi: 10.1093/nar/gks461. URL `https://doi.org/10.1093/nar/gks461`.

[11] Tuan-Minh Nguyen, Adib Shafi, Tin Nguyen, and Sorin Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.*, 20(1):203, October 2019.

[12] Jing Ma, Ali Shojaie, and George Michailidis. A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics*, 20(1):546, December 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3146-1.

[13] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1):7, December 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-7.

[14] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, 15(2):R29, 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-2-r29.

[15] Gabriele Sales, Enrica Calura, Duccio Cavalieri, and Chiara Romualdi. graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):20, December 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-20.

[16] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. doi: doi:10.2202/1544-6115.1027. URL https://doi.org/10.2202/1544-6115.1027.

[17] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi: 10.1093/nar/gkv007.

[18] Kaiyang Zhang, Erdogan Pekcan Erkan, Sanaz Jamalzadeh, Jun Dai, Noora Andersson, Katja Kaipio, Tarja Lamminen, Naziha Mansuri, Kaisa Huhtinen, Olli Carpén, Sakari Hietanen, Jaana Oikkonen, Johanna Hynninen, Anni Virtanen, Antti Häkkinen, Sampsa Hautaniemi, and Anna Vähärautio. Longitudinal single-cell RNA-seq analysis reveals stress-promoted chemoresistance in metastatic ovarian cancer. *Sci. Adv.*, 8(8):eabm1831, February 2022. ISSN 2375-2548. doi: 10.1126/sciadv.abm1831.

[19] K. D. Hansen, R. A. Irizarry, and Z. Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, April 2012. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxr054.

[20] Martin Morgan, Jiefei Wang, Valerie Obenchain, Michel Lang, Ryan Thompson, and Nitesh Turaga. *BiocParallel: Bioconductor facilities for parallel evaluation*, 2022. URL https://github.com/Bioconductor/BiocParallel. R package version 1.32.5.

[21] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, New York, NY, 2009. ISBN 978-0-387-98140-6. doi: 10.1007/978-0-387-98141-3. URL https://link.springer.com/10.1007/978-0-387-98141-3.

[22] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, February 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.026113.

[23] Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanre-waju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, Andrew Berry, Jyothish Bhai, Alexandra Bignell, Konstantinos Billis, Sanjay Boddu, Lucy Brooks, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Jose Gonzalez Martinez, Cristina Guijarro-Clarke, Arthur Gymer, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, José Carlos Marugán, Shamika Mohanan, Aleena Mushtaq, Marc Naven, Denye N Ogeh, Anne Parker, Andrew Parton, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, José G Pérez-Silva, William Stark, Emily Steed, Kyösti Sutinen, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Michal Szpak, Anja Thormann, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Natalie Willhoft, Andrea Winterbottom, Elizabeth Wass, Marc Chakiachvili, Bethany Flint, Adam Frankish, Stefano Giorgetti, Leanne Haggerty, Sarah E Hunt, Garth R IIsley, Jane E Loveland, Fergal J Martin, Benjamin Moore, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Sarah Dyer, Peter W Harrison, Kevin L Howe, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995, November 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1049. URL https://doi.org/10.1093/nar/gkab1049. tex.eprint: https://academic.oup.com/nar/article-pdf/50/D1/D988/42058123/gkab1049.pdf.

[24] Raivo Kolde. *pheatmap: Pretty Heatmaps*, 2019. URL https://CRAN.R-project.org/package=pheatmap. R package version 1.0.12.

[25] Masoumeh Moslemi, Yousef Moradi, Hojat Dehghanbanadaki, Hamed Afkhami, Mansoor Khaledi, Najmeh Sedighimehr, Javad Fathi, and Ehsan Sohrabi. The association between ATM variants and risk of breast cancer: a systematic review and meta-analysis. *BMC Cancer*, 21(1):27, December 2021. ISSN 1471-2407. doi: 10.1186/s12885-020-07749-6.

[26] Jamil Momand, P Zambetti, David C Olson, Donna George, and J Levine'. The mdm-2 Oncogene Product Forms a Complex with the ~53 Protein and Inhibits p53-Mediated Transactivation.

[27] Michael Choi, Thomas Kipps, and Razelle Kurzrock. ATM Mutations in Cancer: Therapeutic Implications. *Molecular Cancer Therapeutics*, 15(8):1781–1791, August 2016. ISSN 1535-7163, 1538-8514. doi: 10.1158/1535-7163.MCT-15-0945.

[28] Liv B. Gansmo, Merete Bjørnslett, Mari Kyllesø Halle, Helga B. Salvesen, Anne Dørum, Einar Birkeland, Kristian Hveem, Pål Romundstad, Lars Vatten, Per Eystein Lønning, and Stian Knappskog. The MDM4 SNP34091 (rs4245739) C-allele is associated with increased risk of ovarian—but not endometrial cancer. *Tumor Biology*, 37(8):10697–10702, August 2016. ISSN 1010-4283, 1423-0380. doi: 10.1007/s13277-016-4940-2.

[29] Gurinder Singh Atwal, Tomas Kirchhoff, Elisabeth E. Bond, Marco Montagna, Chiara Menin, Roberta Bertorelle, Maria Chiara Scaini, Frank Bartel, Anja Böhnke, Christina Pempe, Elise Gradhand, Steffen Hauptmann, Kenneth Offit, Arnold J. Levine, and Gareth L. Bond. Altered tumor formation and evolutionary selection of genetic variants in the human MDM4 oncogene. *Proceedings of the National Academy of Sciences*, 106(25):10236–10241, June 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0901298106.

[30] Dan Liebermann, Jennifer S. Tront, Xiogen Sha, Ksushiki Mukherjee, Alisha Mohamed-Hadley, and Barbara Hoffman. Gadd45 Stress Sensors in Malignancy and Leukemia. *Critical Reviews™ in Oncogenesis*, 16(1-2):129–140, 2011. ISSN 0893-9675. doi: 10.1615/CritRevOncog.v16.i1-2.120.

[31] Suwen Hu, Zi Hui, Jilong Duan, Carmen Garrido, Tian Xie, and Xiang-Yang Ye. Discovery of small-molecule ATR inhibitors for potential cancer treatment: A patent review from 2014 to present. *Expert Opinion on Therapeutic Patents*, 32(4):401–421, April 2022. ISSN 1354-3776, 1744-7674. doi: 10.1080/13543776.2022.2027911.

[32] Siyu Li, Tao Wang, Xichang Fei, and Mingjun Zhang. ATR Inhibitors in Platinum-Resistant Ovarian Cancer. *Cancers*, 14(23):5902, November 2022. ISSN 2072-6694. doi: 10.3390/cancers14235902.

[33] Lanqing Gong, Liqiong Cai, Guodong Li, Jing Cai, and Xiaoqing Yi. GADD45B Facilitates Metastasis of Ovarian Cancer Through Epithelial–Mesenchymal Transition. *OncoTargets and Therapy*, Volume 14:255–269, January 2021. ISSN 1178-6930. doi: 10.2147/OTT.S281450.

List of Figures

Figure 1. Schematic illustration of the differences betweeen conventional pathway analysis methods and *sSNAPPY*. Instead of being limited to treatment-level analyses, *sSNAPPY* allows the detection of pathway perturbation in individual samples by using sample-specific estimates of fold-change instead of experiment-wide estimates. (Created with BioRender.com).
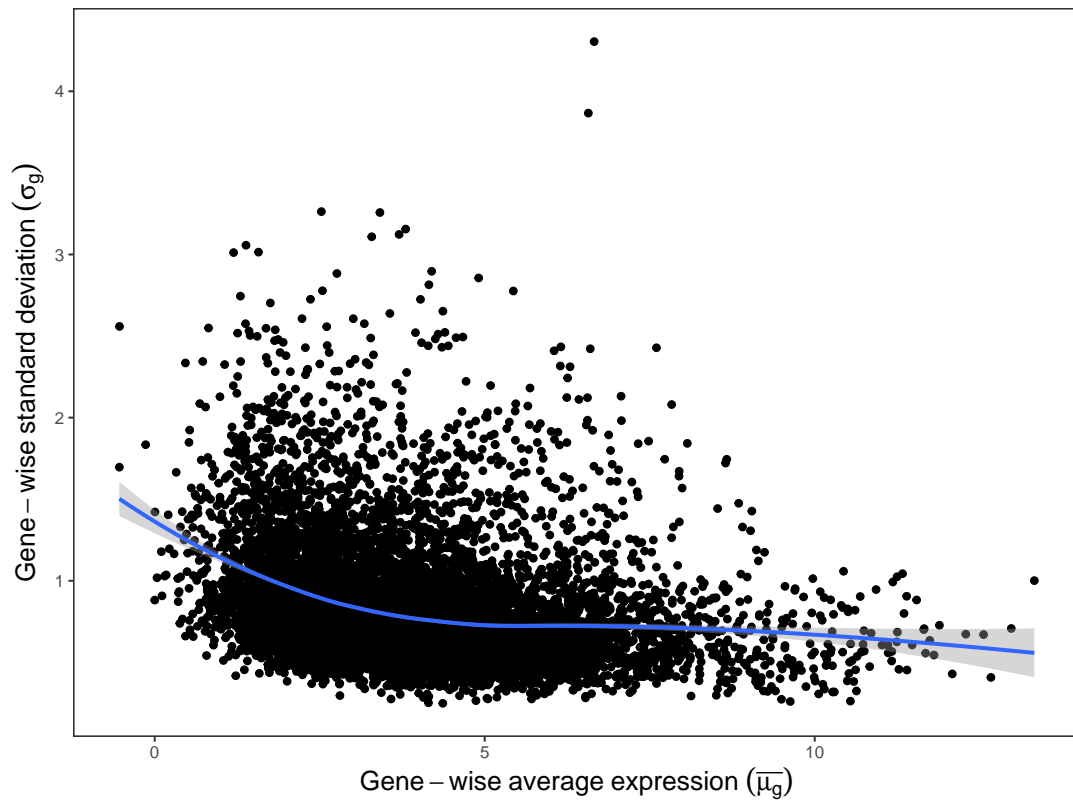
Figure 2. Gene-wise standard deviations are plotted against the mean logCPM values with mean-variance trend modelled by a loess fit. Genes with low expression values tend to have a larger variance.
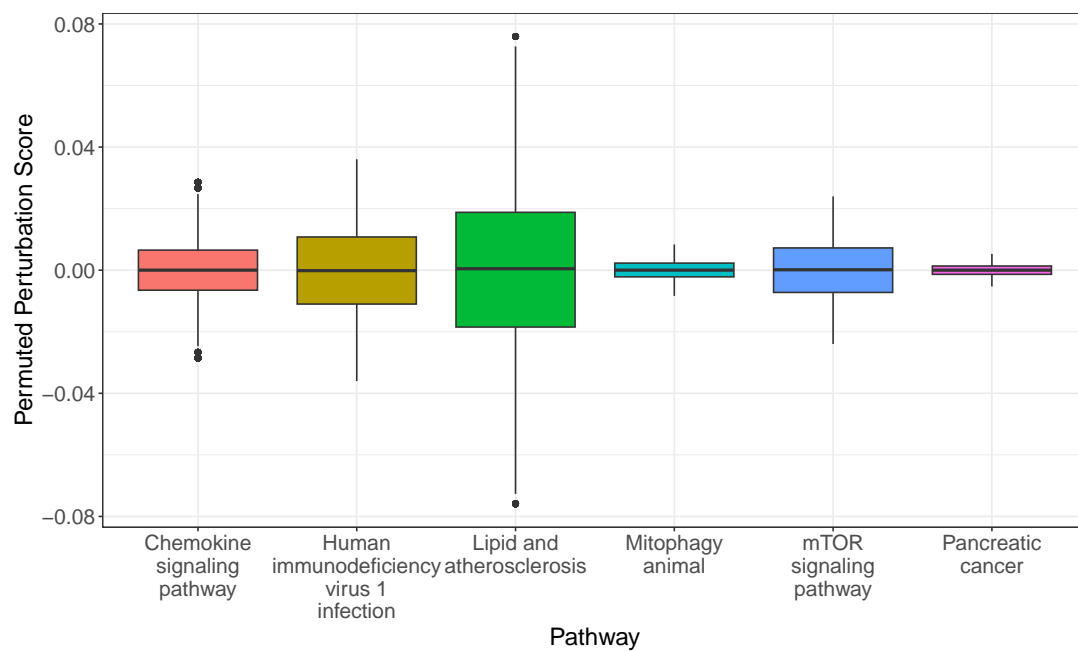
Figure 3. Permuted perturbation scores of six randomly selected pathways. All sampled empirical distributions are approximately normally distributed with a mean of zero.
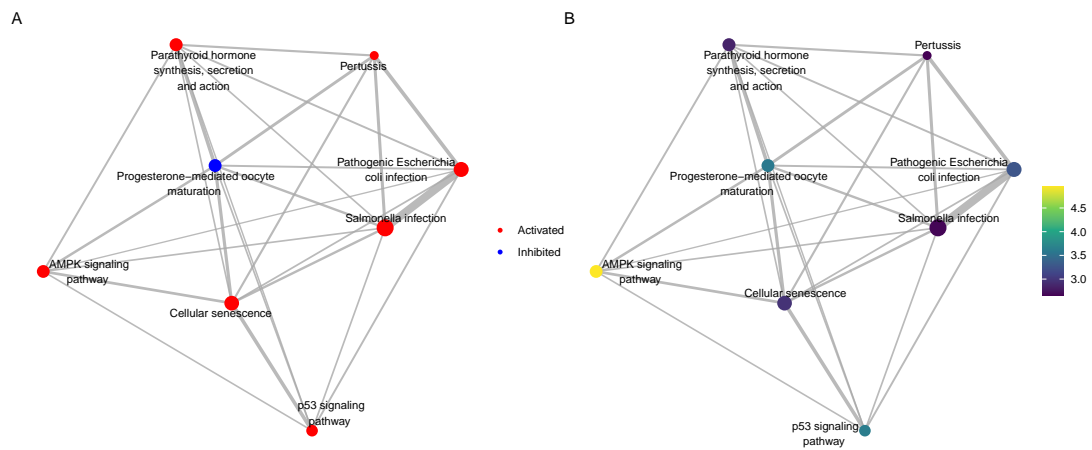
Figure 4. Significantly perturbed KEGG pathways identified among post-chemotherapy samples using sSNAPPY, colored by (A) pathways' predicted direction of changes and (B) pathways' -log10(p-values). Pathways with a FDR < 0.05 in the moderated t-test were included.
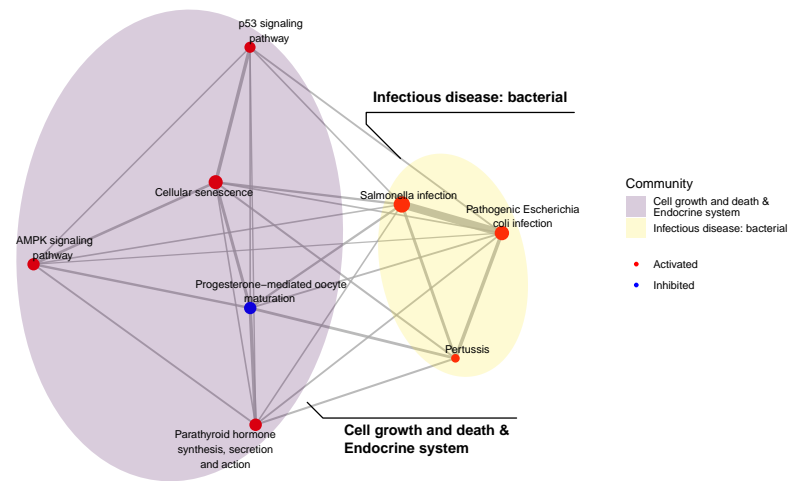
Figure 5. **Significantly perturbed KEGG pathways identified among post-chemotherapy samples using sSNAPPY, colored by community structures detected through the louvain algorithm. The main biological processes perturbed by the chemo-therapy were shown.**
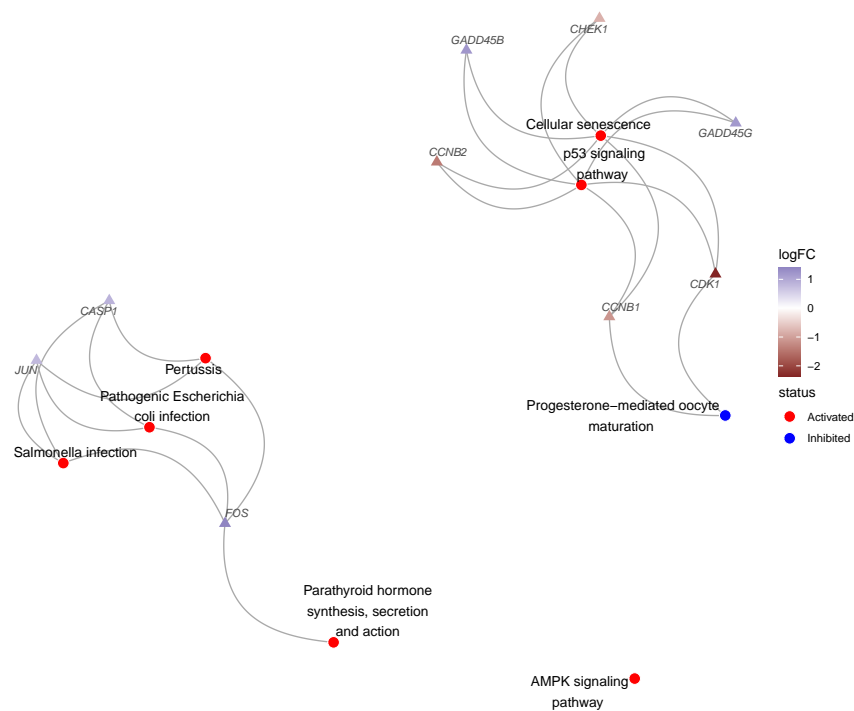
Figure 6. Significantly perturbed KEGG pathways identified among post-chemotherapy samples using sSNAPPY, showing any genes in the top 500 ranked for differential expression, and which pathways they are likely contributing to. No genes in the top500 were found to contribute to the perturbation of the AMPK signaling pathway.
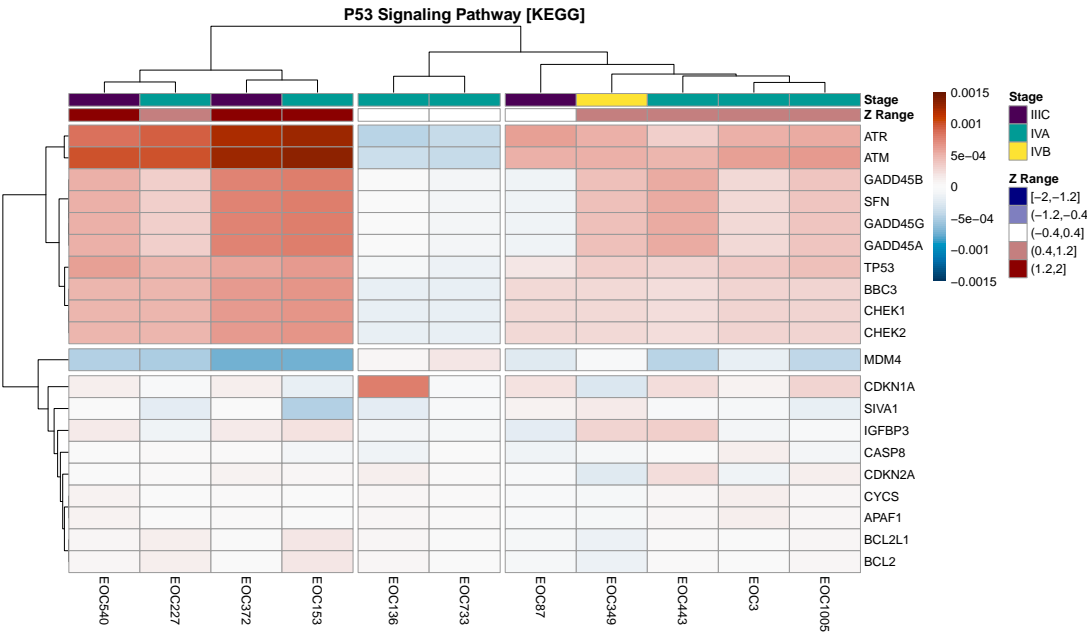
Figure 7. Gene-level perturbation scores of the top 20 genes in the "p53 signalling pathway" ranked by their average contribution to the perturbation score. Samples were annotated by the stage of cancers, along with robust Z-scores which indicate the extent of perturbation within each sample. The genes ATR and ATM were identified as possible key drivers of the activation of p53 signalling pathway, with downregulation of MDM4 also appearing likely to be playing a role.

List of Tables

Table 1. **Significantly impacted KEGG pathways identified among post-chemotherapy samples using sSNAPPY**

Pathway	Change	PValue	FDR	Direction
AMPK signaling pathway	0.88	1.1e-05	0.0024	Activated
p53 signaling pathway	0.88	2.4e-04	0.0186	Activated
Progesterone-mediated oocyte maturation	-0.77	2.6e-04	0.0186	Inhibited
Pathogenic Escherichia coli infection	0.66	5.3e-04	0.0286	Activated
Cellular senescence	0.68	1.1e-03	0.0455	Activated
Parathyroid hormone synthesis, secretion and action	0.63	1.4e-03	0.0455	Activated
Salmonella infection	0.59	2.1e-03	0.0455	Activated
Pertussis	0.59	2.2e-03	0.0455	Activated
Human T-cell leukemia virus 1 infection	0.62	2.2e-03	0.0455	Activated
Yersinia infection	0.62	2.3e-03	0.0455	Activated
Toxoplasmosis	0.63	2.3e-03	0.0455	Activated