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

Wenjun Liu*[1], Ville-Petteri Mäkinen[2,3,4,5], Wayne D. Tilley[1], and Stephen M. Pederson[1,6,7]

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

Abstract Detecting biological pathways whose activities were significantly different between biological conditions is a key interest in analysing RNA-seq data. Majority of current strategies function by testing for over-representation of differentially expressed genes (DEGs) among pre-defined gene-set, failing to account for gene-gene interactions encoded by pathway topologies or predicting the directionality of pathway activities. To address these issues, a single-sample pathway perturbation analysis method that leverages pathway topology information to compute perturbation scores that predict pathways' potential directions of changes was developed and made into a R/Bioconductor package *sSNAPPY*. We demonstrated the use of *sSNAPPY* by applying it to public scRNA-seq data derived from ovarian cancer patient tissues collected before and after chemotherapy. Not only replicated results reported in the original study, *sSNAPPY* was also able to detect significant pathway perturbation of other interesting biological processes. *sSNAPPY* presents a novel pathway analysis strategy that takes into consideration of pathway topology to predict biology pathways' activities in both individual samples and treatment groups. In addition to not relying on the detection of differentially expressed genes, the method also offers great flexibility and provides access to powerful visualisation functions.

Keywords

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

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

R version: R version 4.2.2 (2022-10-31)

Bioconductor version: 3.16

Introduction

Using pathway enrichment analysis to gain biological insights from changes in gene expression is a pivotal step in analysing an interpreting RNA-seq data, for which numerous methods have been developed. Most of the existing methods, regardless of whether relying on the detection of differentially expressed genes (over-representation analysis ORA)[1] or scoring all genes (functional class scoring methodologies FCS)[2], tend to view pathways simply as a collection of gene names. However, databases such as Kyoto Encyclopaedia of Genes and Genomes (KEGG)[3] and WikiPathways[4] capture not only which genes are implicated in a certain biological process but also their interactions and relevant importance, both of which are overlooked in ORA and FCS.

To fully utilise that additional information, the latest generation of pathway analysis approaches have predominantly been topology-based[5]. 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 account6. Nevertheless, most of those topology-based methods focus only on comparing activities of pathways between two groups (eg. treatment) and cannot be used to score individual samples[6, 7]. In heterogenous data where more than one factors could affect pathway activities, however, individual sample scoring could be desirable[8]. To address this gap, we present a single-sample directional pathway perturbation analysis strategy called *sSNAPPY*. This article explains how *sSNAPPY* computes changes in gene expression in individual samples and propagates that through gene-set topologies to predict the perturbation in pathway activities. The implication of the R/Bioconductor package will be illustrated through a public scRNA-seq dataset that has been pseudo-bulked.

Methods

Implementation

sSNAPPY is an R package that has been reviewed and published on the open-source bioinformatics software platform Bioconductor with its source codes deposited to GitHub. It is a topology-based pathway analysis method designed to compute directional single-sample pathway perturbation scores in RNA-seq datasets with matched-paired design (eg. samples collected before and after treatment), allowing the detection of significant pathway perturbation on not only treatment but also single-sample level.

To implement *sSNAPPY*, the only inputs required are a log-transformed counts per million (CPM) matrix and matching sample metadata data.frame describing the treatment of each sample and the categorical variable grouping samples into matched pairs. It is assumed that genes with low counts have been filtered out and the matrix has been normalised to correct for technical artefacts such as gene length or GC contents. Next, knowing what the baseline treatment level is, the single-sample log fold-changes (logFC) of each treated sample will be calculated by subtracting the logCPM of the corresponding baseline samples from the treated ones. Since it has been shown previously that in RNA-seq data, genes with lower expressions turn to have larger variance[9] , we introduced a weighting strategy to down-weight the raw single-sample logFCs of those genes. Our weighting strategy functions by estimating the relationships between the variances in single-sample logFCs and mean logCPMs for each gene with loess regression and defining the gene-wise weight to be the inverse of the predicted variance derived from applying the regression model to genes' mean logCPM values.

sSNAPPY was built upon the group-level topology-based scoring algorithm initially proposed in R package SPIA[10] to propagate genes' changes in expression through pathway topologies to compute one perturbation score for each treated sample in each pathway, where the signs of the scores predict biological processes' directions of changes. The single-sample perturbation score of a given pathway t_{A_s} is defined to be:

$$S_{hi} = \Delta mu_{hi} + \sum_{j>i}^{N_U(i)} B_{ij} \frac{S_{hj}}{N_D(j)}$$

$$S_h = \sum (S_{hi} - \Delta mu_{hi})$$

where:

- S_{hi} is the gene-wise perturbation score of gene i in sample h
- Δmu_{hi} is the weighted logFC of gene i in sample h
- g_j are the genes directly upstream g_i

- $N_U(i)$ is the number of genes directly upstream g_i
- $N_D(j)$ is the number of genes directly downstream g_j
- B_{ij} is the gene-gene interaction encoded by the topology matrix
- S_h is the pathway-level perturbation score in sample h

Bioconductor package graphite[11] provides functions that can be used to retrieve pathway topologies from a database and convert topology information to adjacency matrices, but to streamline this process, we have chained multiple graphite functions into one, where users only need to provide the name of the desired database to retrieve all topology information in the format required by the scoring algorithm with the correct type of gene identifiers.

To scale the single-sample pathway perturbation scores so they are comparable across pathways and to test the significances of individual scores, null distributions of perturbation scores for each pathway will be generated through a sample permutation strategy. With each permutation round, the column names (ie. sample labels) for the logCPM matrix will be randomly shuffled while the rest of the scoring algorithm remains unchanged. We recommend users to perform a minimum of 1000 rounds of permutation, 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 and obtain associated two-sided p-values. Since the method is single sample-based, the permutation strategy remains applicable regardless of experimental design.

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 could be achieved by modelling scores with regression models. The single-sample nature of *sSNAPPY*'s pathway perturbation scores is particularly helpful for datasets with complex experimental designs or known confounding factors.

Operation

The package has been tested and will function on all operating systems. To install and operate the package, R version 4.2.0+ is required. The package needs to be downloaded using BiocManager.

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

Use Cases

Data

We used A publicly available scRNA-seq data of patient-derived ovarian tissue collected prior to and after 11 homogeneously treated high-grade serous ovarian cancer (HGSOC) patients were subjected to chemotherapy[12] to demonstrate the use of *sSNAPPY* here. Pre-processed count data were retrieved from Gene Expression Omnibus (GEO) with accession code GSE165897.

In the original study, cells were classified into epithelial, stromal, and immune cells but we have chosen to only focuses on epithelial cells as they were what the original study primarily focused on. 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. A total of 11,101 (33.8%) out of 32,847 annotated genes passed the selection criteria and were included in downstream analyses. Conditional quantile normalisation[13] was applied to mitigate potential biases introduced by gene length and GC content. The logCPM matrix of the processed dataset and sample metadata can be downloaded from here.

To start, firstly load all the packages that will be used in this workflow:

```
library(sSNAPPY)
library(tidyverse)
library(magrittr)
library(ggplot2)
library(cowplot)
library(kableExtra)
library(AnnotationHub)
library(edgeR)
```

To read in the data:

```
logCPM <- readRDS(here::here("data/logCPM.rds"))
sample_meta <- readRDS(here::here("data/sample_meta.rds"))
head(sample_meta)
```

```
## # A tibble: 6 x 8
##   sample                 treatment       patie~1 anato~2   Age Stage   PFI   CRS
##   <chr>                  <chr>           <chr>   <chr>   <dbl> <chr> <dbl> <dbl>
## 1 EOC372_treatment-naive treatment-naive EOC372  Perito~    68 IIIC    460     1
## 2 EOC372_post-NACT       post-NACT       EOC372  Perito~    68 IIIC    460     1
## 3 EOC443_post-NACT       post-NACT       EOC443  Omentum    54 IVA     177     3
## 4 EOC443_treatment-naive treatment-naive EOC443  Omentum    54 IVA     177     3
## 5 EOC540_treatment-naive treatment-naive EOC540  Omentum    62 IIIC    126     2
## 6 EOC540_post-NACT       post-NACT       EOC540  Omentum    62 IIIC    126     2
## # ... with abbreviated variable names 1: patient_id, 2: anatomical_location
```

Data preparation and retrieval pathway topology

To apply sSNAPPY, the rownames of the logCPM matrix must be converted to Entrez IDs. Genes without an Entrez IDs were removed.

```
ah <- AnnotationHub() %>%
  AnnotationHub::subset(rdataclass == "EnsDb") %>%
  AnnotationHub::subset(str_detect(description, "101")) %>%
  AnnotationHub::subset(genome == "GRCh38")
stopifnot(length(ah) == 1)
ensDb <- ah[[1]]
rownames(logCPM) <- mapIds(ensDb,rownames(logCPM),
                           "ENTREZID", keytype = "GENENAME")
```

```
## Warning: Unable to map 210 of 10311 requested IDs.
```

```
# Remove genes that couldn't be matched to entrez IDs
logCPM <- logCPM[!is.na(rownames(logCPM)),]
```

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")
```

Instead of downloading the topology matrices of all pathways, it is also possible to specify specific pathways' names to focus on. Customised weights could be assigned to different types of gene-gene interaction type by providing a named numeric vector.

```
# Only retrieve the topology matrices of 3 specific pathways
gsTopology_sub <- retrieve_topology(
  database = "kegg",
  pathwayName = c(
    "Glycolysis / Gluconeogenesis",
    "Citrate cycle (TCA cycle)",
    "Pentose phosphate pathway"
))
```

Score single-sample pathway perturbation

To compute the single-sample logFCs needed for perturbation scores, samples much be in matched pairs and the factor defining those pairs must be specified in the `weight_ss_fc()` function. 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, one level of which must be the control level indicated as a parameter of the `weight_ss_fc()` function.

```
weightedFC <- weight_ss_fc(
    logCPM, sample_meta, factor = "patient_id",
    control = "treatment-naive"
)
names(weightedFC)
```

```
## [1] "weight" "logFC"
```

The output of the weight_ss_fc function is a `list` where one element is a matrix of single-sample logFCs with rows corresponding to genes and columns to treated samples and the other element is a vector of gene-wise weights that will be used to alleviate the influence of lowly expressed but highly variable genes.

Single-sample logFCs will then be piped through pathway topologies to compute the gene-wise perturbation scores for all genes included in a pathway. Apart from being summed into pathway-level perturbation scores, gene-wise perturbation scores can also be ranked to identify genes playing the most significant roles in each pathway, which will be further elaborated in the visualisation section below. The pathway-level perturbation scores will be returned as a data.frame containing sample and gene-set names.

```
# Compute perturbation scores at the gene-level within
# each pathway each treated sample
genePertScore <- raw_gene_pert(weightedFC$logFC, gsTopology)
# Sum gene-level scores to derive pathway-level scores
ssPertScore <- pathway_pert(genePertScore)
head(ssPertScore)
```

```
##   sample           tA                                    gs_name
## 1 EOC372  0.005332217 EGFR tyrosine kinase inhibitor resistance
## 2 EOC443 -0.001506460 EGFR tyrosine kinase inhibitor resistance
## 3 EOC540 -0.006199438 EGFR tyrosine kinase inhibitor resistance
## 4   EOC3 -0.004205656 EGFR tyrosine kinase inhibitor resistance
## 5  EOC87 -0.003840786 EGFR tyrosine kinase inhibitor resistance
## 6 EOC136 -0.008354838 EGFR tyrosine kinase inhibitor resistance
```

Sample permutation for normalisation and significance testing

To estimate the significance of individual scores and transform the scores so 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 parallelises this step through functions provided by BiocParallel. Users can choose to customize the parallel back-end or the default one returned by `BiocParallel::bpparam( )` will be used. If the number of samples or the size of the chosen pathway database is large, it is recommended to perform the permutation step on a high-performance computer.

```
permutedScore <- generate_permuted_scores(
    logCPM, numOfTreat = 2, NB = 1000,
    gsTopology = gsTopology, weight = weightedFC$weight
)
```

Next, using the function normalise_by_permu, the raw perturbation scores will be normalised to robust z-scores using the median and MAD of permuted scores and further converted to two-sided p-values. Defaulted to using the false-discovery rate, the p-values will be corrected for multiple testings using a user-defined approach. In this case, using an FDR of 0.05 as a cut-off, none of the pathways was considered to be significantly perturbed at the individual sample level.

```
normalisedScores <- normalise_by_permu(permutedScore, ssPertScore)
head(normalisedScores)
```

```
##                                   gs_name          MAD        MEDIAN  sample
## 1 EGFR tyrosine kinase inhibitor resistance 0.0073873075 -1.837123e-04 EOC1005
## 2                     Endocrine resistance 0.0109646845  1.323193e-04 EOC1005
```

```
## 3                     Antifolate resistance 0.0017190861  2.189169e-05 EOC1005
## 4                   Platinum drug resistance 0.0085760397  9.387240e-05 EOC1005
## 5                  mRNA surveillance pathway 0.0001001364 -1.095895e-06 EOC1005
## 6                            RNA degradation 0.0015561033  1.181997e-06 EOC1005
##               tA    robustZ    pvalue adjPvalue
## 1 -0.0044946212 -0.5835562 0.5595190 0.9996697
## 2  0.0051022014  0.4532627 0.6503596 0.9996697
## 3  0.0011010089  0.6277272 0.5301826 0.9996697
## 4 -0.0065039960 -0.7693374 0.4416930 0.9996697
## 5 -0.0001183863 -1.1713060 0.2414758 0.9996697
## 6 -0.0001092463 -0.0709646 0.9434259 0.9996697
```

A key biological question to answer in this study was what biological processes were impacted by chemotherapy across all patients, the answer to which could be obtained by applying t-tests to normalised scores for each pathway, with the null hypothesis being the mean of normalised perturbation scores equals to 0 for a given pathway. To avoid under-estimation of sample variance, the moderated t-statistics approach proposed by Smyth 200414 was adopted here. Being an empirical Bayes strategy, in moderated t-test, estimated sample variances will be adjusted towards the expected variances estimated by pooling information across all genes/pathways. Performances of the moderated t-statistics were evaluated through simulation studies, and the moderated approach demonstrated more advantageous performance in controlling the false discovery rate.

```
# Normalised perturbation scores were converted to a matrix,
# with rows corresponding to pathways and columns to samples
pert_matrix <- normalisedScores %>%
    dplyr::select(robustZ, gs_name, sample) %>%
    pivot_wider(
        names_from = sample,
        values_from = robustZ
    ) %>%
    column_to_rownames("gs_name") %>%
as.matrix()
# Linear models were fitted for each pathway. No design matrix was
# specified as all samples were replicates of the same treatment group
fit_kg <- lmFit(pert_matrix)
# Moderated t-statistics are calculated using the eBayes function
fit_kg <- eBayes(fit_kg, trend = abs(rowMeans(pert_matrix)))
```

Pathways with an FDR < 0.05 in the moderated t-test were considered to be significantly perturbed at the group level. 22 out of the 315 tested KEGG pathways passed the selection threshold (Table 1).

```
# Use topTable to summarise the linear model fit and correct the p-values with FDR
table1 <- topTable(fit_kg,
                   number = Inf) %>%
    rownames_to_column("gs_name") %>%
    mutate(
        logFC = round(logFC, 4),
        gs_name = as.factor(gs_name),
        Direction = ifelse(logFC < 0, "Inhibited", "Activated"),
        Significant = ifelse(adj.P.Val < 0.05, TRUE, FALSE)) %>%
    dplyr::select(
        `Gene-set Name` = gs_name, `Change in Perturbation Score` = logFC,
        P.Value, FDR = adj.P.Val, Direction, Significant
    )
```

[Table 1 about here.]

In the original study[12], to study the effect of chemotherapy, unsupervised clustering was performed on all cells labelled to be cancer cells. Identified cancer clusters were labelled by performing pathway enrichment testing on cluster marker genes. Two clusters, associated with proliferative DNA repair signatures and stress-related markers, were found to contain significantly higher numbers of post-chemotherapy cells then pre-treatment ones (Ta in Zhang et al. [12]). Marker genes reported for those two clusters were observed in pathways that were detected to be significantly perturbed by sSNAPPY(Table 2).

[Table 2 about here.]

Apart from treating all treated samples as biological replicates, users might also wish to subset samples into groups by phenotypic traits known to affect patients' responses to chemotherapy, such as the stages of their cancers. To do that through the moderated t-statistic strategy, we simply need to provide a design matrix in the `lmFit` step.

```
X <- model.matrix(
    ~0 + Stage,
    data = sample_meta %>%
        dplyr::filter(treatment == "post-NACT") %>%
        .[match(colnames(pert_matrix), .$patient_id),] %>%
        mutate(
            Stage = ifelse(Stage == "IVA", "IVA", "IIIC/IVB")
        ) )  %>%
    set_colnames(str_remove_all(colnames(.), "Stage")) %>%
    .[,colSums(.) != 0]
fit_kg2 <- lmFit(pert_matrix, design = X)
```

Visualise perturbed pathways as networks

sSNAPPY provides various visualisation functions to assist in the interpretation of results. Since biological pathways are not independent of each other and often contain redundant genes, visualising pathway analysis results as a network is a powerful way to not only intuitively summarise the results but also to facilitate the interpretations of the underlying biology. The `plot_gs_network()` function allows users to easily convert a list of significantly 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 1). The returned plot is a ggplot2 [14] object, meaning that components of its theme could be customized just as any other ggplots using the `ggplot2::theme` function.

```
# Extract significantly perturbed pathway. To colour the nodes by directions
# of changes, the column name of the average perturbation scores must be robustZ.
sigPathway <- topTable(fit_kg, number = Inf) %>%
    dplyr::filter(adj.P.Val < 0.05) %>%
    rownames_to_column("gs_name") %>%
    dplyr::rename(
        robustZ = AveExpr,
        pvalue = P.Value
    )
# Plot the network structure
p1 <- sSNAPPY::plot_gs_network(
    normalisedScores = sigPathway,
    gsTopology = gsTopology,
    colorBy = "robustZ"
) +
    theme(
        panel.border = element_blank(),
        panel.background = element_blank()
    )
p2 <- sSNAPPY::plot_gs_network(
    normalisedScores = sigPathway,
    gsTopology = gsTopology,
    # or color nodes by significance levels
    colorBy  = "pvalue"
) +
    theme(
        panel.border = element_blank(),
        panel.background = element_blank()
    )
plot_grid(
    p1, p2,
    labels = c("A", "B")
)
```

[Figure 1 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. To confirm it and further condense the information, we can summarise the network structures into key biological groups by performing community detection. Widely used in network analysis, community detection is a technique used to identify groups of nodes that are more densely connected than to any other nodes in the network[15]. 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. Retrieved directly from the KEGG website, we have curated the most recent categorisations of KEGG pathways and included it as part of the sSNAPPY package. Annotation of KEGG pathway communities will be automatically completed by calling the in-built database, while analyses involving other pathway databases require user-provided pathway categorisations.

The defaulted Louvain method was applied to the network of significantly perturbed biological pathways and revealed two community structures, where one was annotated to be endocrine system related and the other one was infectious disease-related (Figure 2).

```
sSNAPPY::plot_community(
    normalisedScores = sigPathway,
        gsTopology = gsTopology,
        colorBy = "robustZ",
    ) +
    theme(
        panel.border = element_blank(),
        panel.background = element_blank()
    )
```

[Figure 2 about here.]

Inferred directly from the expression matrix, a key advantage of sSNAPPY is that it does not require the prior definition of differentially expressed genes, which is not always detectable in clinical datasets. However, knowing the genes that are implicated in perturbing biological pathways, particularly those that affect multiple gene-sets, can provide valuable insights for biologists seeking to formulate hypotheses about underlying biological mechanisms. Therefore, sSNAPPY presents another visualisation feature called `plot_gs2gene`, which enables the inclusion of pathway genes in network structures. Although defaulted to colour all gene nodes in grey, users can provide a vector of logFCs to colour the genes by their changes in expression. As pathways are often made of hundreds of genes, it is recommended not to plot all genes included in perturbed pathways but filter for genes more likely to be playing a significant role, achieved through only providing logFCs of genes to plot. In this example dataset, we chose to only include genes with the top 500 magnitudes of mean logFCs (Figure 3).

```
# Calculate gene-wise mean logFCs
meanFC <- apply(weightedFC$logFC, 1, mean )
# Extract the top 500 meanFCs
top500_FC <- meanFC %>%
    abs() %>%
    sort(decreasing = TRUE, ) %>%
    .[1:500]
```

Since pathway topologies were retrieved in Entrez IDs, by default, genes' Entrez IDs will be used to annotate gene nodes in the plot. However, users can provide a data.frame mapping Entrez IDs to their chosen identifiers through the mapRownameTo parameter. A data.frame converting Entrez IDs to ensemble gene names has been made available as part of the package.

```
# Read in built-in data.frame entrez2name that matches genes'
# Entrez IDs to gene names
load(system.file("extdata", "entrez2name.rda", package = "sSNAPPY"))
head(entrez2name)
```

```
## # A tibble: 6 x 3
##   gene_id         mapTo   entrezid
##   <chr>           <chr>   <chr>
## 1 ENSG00000223972 DDX11L1 ENTREZID:84771
## 2 ENSG00000223972 DDX11L1 ENTREZID:727856
## 3 ENSG00000223972 DDX11L1 ENTREZID:100287102
```

```
## 4 ENSG00000223972 DDX11L1 ENTREZID:100287596
## 5 ENSG00000223972 DDX11L1 ENTREZID:102725121
## 6 ENSG00000227232 WASH7P  ENTREZID:653635
```

```
# Plot the pathway-gene network for genes with top 500
#magnitudes of mean FCs and label gene nodes by gene names
sSNAPPY::plot_gs2gene(
    normalisedScores = sigPathway,
    gsTopology = gsTopology,
    colorGS_By = "robustZ",
    mapEntrezID = entrez2name,
    geneFC = top500_FC,
    edgeAlpha = 0.3,
    GsName_size = 4
) +
    theme(
        panel.border = element_blank(),
        panel.background = element_blank()
)
```

[Figure 3 about here.]

Identify hub genes contributing to pathway perturbation

If we would like to further investigate a specific pathway and elucidate the key genes that contributed to its perturbation, such as the activation of the "p53 signalling pathway," we can employ a heatmap to display the gene-level perturbation scores of all the genes within the pathway and annotate each column (ie. each sample) by the direction of pathway perturbation in that sample or any other sample metadata using the plot_gene_contribution function (Figure 4).

```
plot_gene_contribution(
    genePertScore = genePertScore,
    gsToPlot = "p53 signaling pathway",
    metadata = dplyr::filter(sample_meta, treatment == "post-NACT") %>%
        dplyr::select(sample = patient_id, Stage),
    annotation_attribute = c("pathwayPertScore", "Stage"),
    pathwayPertScore = ssPertScore,
    mapEntrezID = entrez2name
)
```

[Figure 4 about here.]

From this heatmap we can easily identify that the hub genes making the biggest contribution to the activation of p53 signalling pathway upon chemotherapy were gene ART and gene ATM. The Ataxia-telangiectasia mutated (ATM) gene is a well-established oncosuppressor[16], mutation of which has been observed in many types of cancers[17]. Also involved in DNA damage repair, the Ataxia telangiectasia and RAD3-related protein kinase (ATR) gene has been shown to be a promising therapeutic target for HGSOC[18].

Discussion

In conclusion, the paper showcased an R/Bioconductor package that offers a novel single-sample pathway perturbation testing approach. sSNAPPY utilizes pathway topology information to compute perturbation scores that predict pathways' potential directions of changes in individual samples. This approach addresses the limitations of current strategies that fail to account for gene-gene interactions encoded by pathway topologies or predict the directionality of pathway activities. 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 beyond what were shown in the original studies. Overall, sSNAPPY presents a promising strategy for single sample-based pathway analysis in RNA-seq data.

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

Any grants that supported the work must be listed here, including the grant number.

Acknowledgements

[1] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, 21(13):2988–2993, July 2005. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bti457.

[2] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, December 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-47.

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

[4] 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.

[5] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol*, 8(2):e1002375, February 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002375.

[6] 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.

[7] Michaela Bayerlová, Klaus Jung, Frank Kramer, Florian Klemm, Annalen Bleckmann, and Tim Beißbarth. Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics*, 16(1):334, December 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0751-5.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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`.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

List of Figures

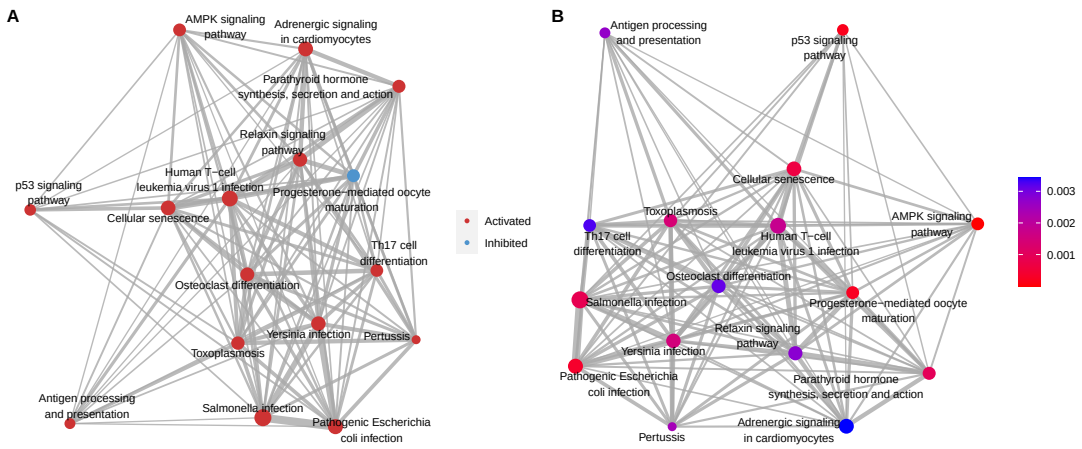

Figure 1. Significantly perturbed KEGG pathways identified among post-chemotherapy samples using sSNAPPY, colored by (A) pathways' predicted directions of changes and (B) pathways' -log10(p-values). Pathways with a FDR < 0.05 in the moderated t-test were included.

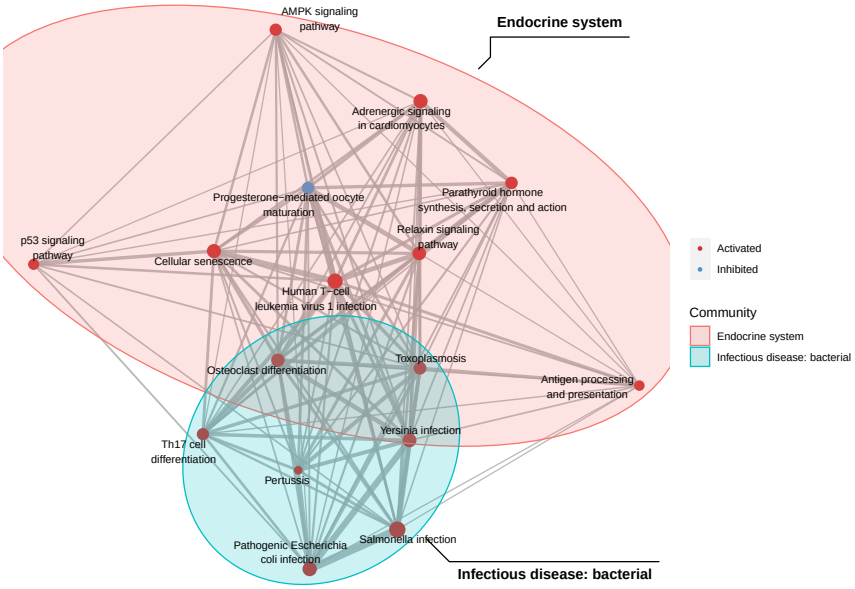

Figure 2. Significantly perturbed KEGG pathways identified among post-chemotherapy samples using sSNAPPY, colored by (A) pathways' predicted directions of changes and (B) pathways' -log10(p-values). Pathways with a FDR < 0.05 in the moderated t-test were included.

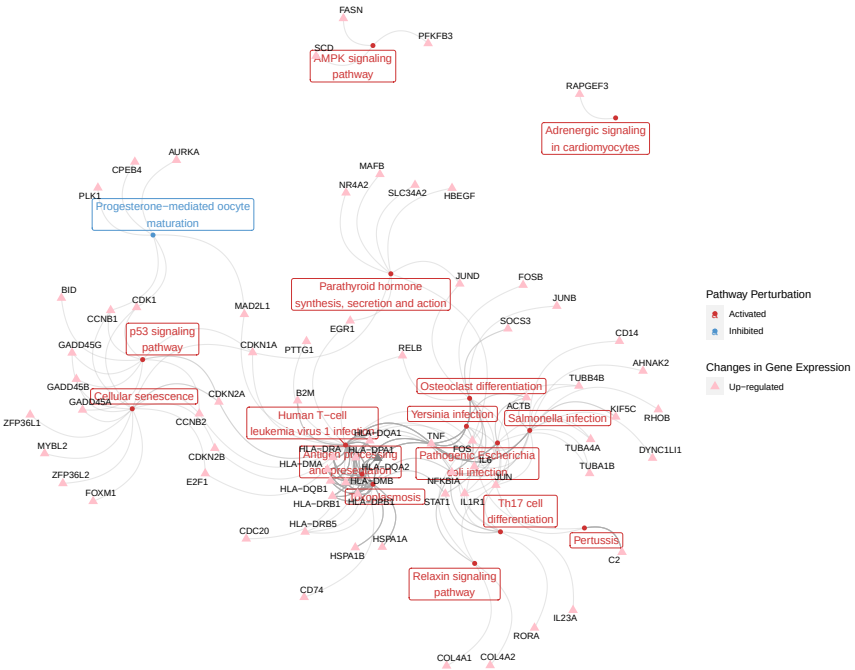

Figure 3. Significantly perturbed KEGG pathways identified among post-chemotherapy samples using sSNAPPY, annotated by community structure identified with the Louvain algorithm.2 communities were formed, both of which were annotated by the pathway category that the majority of the pathways belong to.

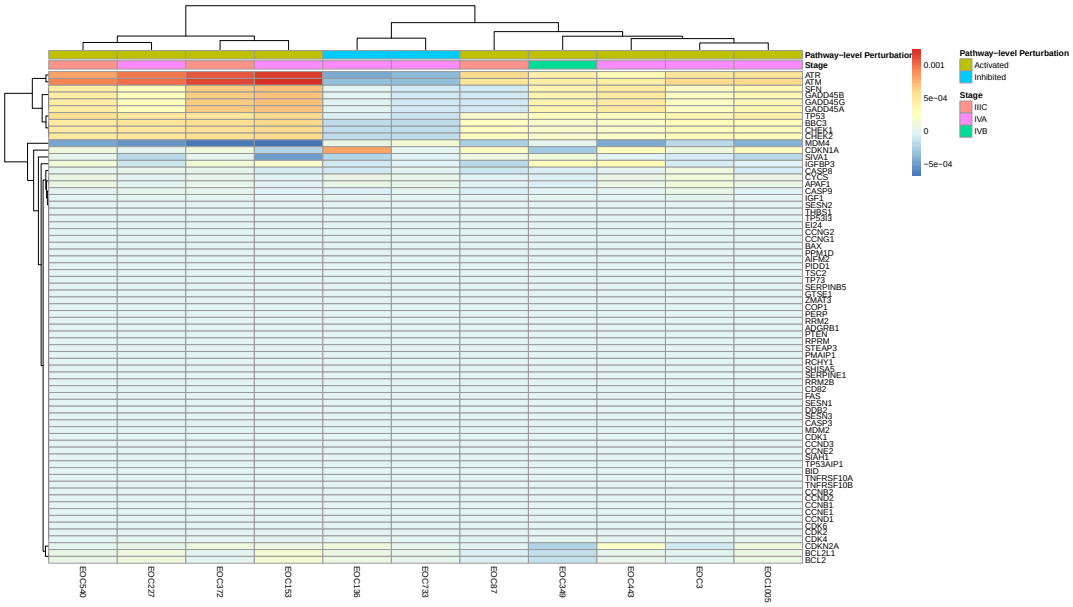

Figure 4. Gene-level perturbation scores of all genes included in the "p53 signalling pathway" pathway in each sample where columns of samples were annotated by pathway-level perturbation and the stages of cancers. Genes ATR and ATM were the key driver of the activation of p53 signalling pathway.

List of Tables

Table 1. Results of KEGG pathways identified among post-chemotherapy samples using sSNAPPY

Gene-set Name	Change in Perturbation Score	PValue	FDR	Direction	Significant
AMPK signaling pathway	0.8339	0.0000053	0.0011315	Activated	TRUE
p53 signaling pathway	0.8218	0.0001878	0.0183811	Activated	TRUE
Progesterone-mediated oocyte maturation	-0.7096	0.0002577	0.0183811	Inhibited	TRUE
Pathogenic Escherichia coli infection	0.6356	0.0003715	0.0198775	Activated	TRUE
Cellular senescence	0.6659	0.0006706	0.0287013	Activated	TRUE
Salmonella infection	0.6165	0.0008808	0.0302502	Activated	TRUE
Parathyroid hormone synthesis, secretion and action	0.6150	0.0009895	0.0302502	Activated	TRUE
Toxoplasmosis	0.6350	0.0014170	0.0357963	Activated	TRUE
Yersinia infection	0.5952	0.0015055	0.0357963	Activated	TRUE
Human T-cell leukemia virus 1 infection	0.5954	0.0018025	0.0385739	Activated	TRUE
Pertussis	0.5647	0.0024215	0.0458335	Activated	TRUE
Antigen processing and presentation	0.5594	0.0026594	0.0458335	Activated	TRUE
Relaxin signaling pathway	0.5105	0.0027843	0.0458335	Activated	TRUE
Osteoclast differentiation	0.5906	0.0031175	0.0459781	Activated	TRUE
Th17 cell differentiation	0.6178	0.0032525	0.0459781	Activated	TRUE
Adrenergic signaling in cardiomyocytes	0.6312	0.0034376	0.0459781	Activated	TRUE
Rheumatoid arthritis	0.5555	0.0044125	0.0509248	Activated	FALSE
Human immunodeficiency virus 1 infection	0.5380	0.0046017	0.0509248	Activated	FALSE
Leukocyte transendothelial migration	0.5401	0.0047306	0.0509248	Activated	FALSE
B cell receptor signaling pathway	0.5605	0.0047593	0.0509248	Activated	FALSE
Fluid shear stress and atherosclerosis	0.5416	0.0050961	0.0515160	Activated	FALSE
Chagas disease	0.4553	0.0052960	0.0515160	Activated	FALSE
Amphetamine addiction	0.4969	0.0066948	0.0600218	Activated	FALSE
Parkinson disease	0.6322	0.0067314	0.0600218	Activated	FALSE
Chemical carcinogenesis - reactive oxygen species	0.4914	0.0076541	0.0636785	Activated	FALSE
Prion disease	0.5744	0.0077366	0.0636785	Activated	FALSE
Transcriptional misregulation in cancer	0.5247	0.0083880	0.0664828	Activated	FALSE
Amyotrophic lateral sclerosis	0.5598	0.0095735	0.0722516	Activated	FALSE
Autophagy - other	0.5089	0.0097911	0.0722516	Activated	FALSE
Endocrine resistance	0.4517	0.0103565	0.0738765	Activated	FALSE
cAMP signaling pathway	0.4566	0.0111216	0.0767749	Activated	FALSE
Estrogen signaling pathway	0.5307	0.0115901	0.0775091	Activated	FALSE
Leishmaniasis	0.5675	0.0130130	0.0843876	Activated	FALSE
C-type lectin receptor signaling pathway	0.6259	0.0145868	0.0891971	Activated	FALSE
Growth hormone synthesis, secretion and action	0.4164	0.0145883	0.0891971	Activated	FALSE
Chemokine signaling pathway	0.5603	0.0154961	0.0921156	Activated	FALSE
Aldosterone-regulated sodium reabsorption	0.4137	0.0192526	0.1113530	Activated	FALSE
T cell receptor signaling pathway	0.4436	0.0205149	0.1155315	Activated	FALSE
Mitophagy - animal	0.4634	0.0224493	0.1231833	Activated	FALSE
Non-alcoholic fatty liver disease	0.4035	0.0270647	0.1447960	Activated	FALSE
Longevity regulating pathway	0.4070	0.0283340	0.1478898	Activated	FALSE
Shigellosis	0.4729	0.0309009	0.1574472	Activated	FALSE
IL-17 signaling pathway	0.4055	0.0352484	0.1754221	Activated	FALSE
Axon guidance	-0.3872	0.0403841	0.1947390	Inhibited	FALSE
EGFR tyrosine kinase inhibitor resistance	-0.3843	0.0409498	0.1947390	Inhibited	FALSE
Olfactory transduction	0.4286	0.0477795	0.2222785	Activated	FALSE
FoxO signaling pathway	-0.4077	0.0508525	0.2315413	Inhibited	FALSE
Colorectal cancer	-0.3727	0.0531799	0.2370936	Inhibited	FALSE
Fc epsilon RI signaling pathway	-0.3842	0.0558730	0.2440167	Inhibited	FALSE
Bladder cancer	-0.3708	0.0648394	0.2775128	Inhibited	FALSE
Spinocerebellar ataxia	0.3255	0.0674275	0.2829310	Activated	FALSE
Viral protein interaction with cytokine and cytokine receptor	0.4174	0.0705270	0.2902456	Activated	FALSE
Signaling pathways regulating pluripotency of stem cells	0.3469	0.0733669	0.2940451	Activated	FALSE
Endometrial cancer	-0.3182	0.0753600	0.2940451	Inhibited	FALSE
Chronic myeloid leukemia	-0.3675	0.0774357	0.2940451	Inhibited	FALSE
Coronavirus disease - COVID-19	0.3324	0.0776344	0.2940451	Activated	FALSE
Cushing syndrome	0.3756	0.0789040	0.2940451	Activated	FALSE
Necroptosis	-0.3850	0.0799111	0.2940451	Inhibited	FALSE
Rap1 signaling pathway	0.3348	0.0810685	0.2940451	Activated	FALSE
Fanconi anemia pathway	0.3371	0.0891791	0.3145572	Activated	FALSE
Non-small cell lung cancer	-0.3177	0.0896635	0.3145572	Inhibited	FALSE
Chemical carcinogenesis - receptor activation	0.2794	0.0923548	0.3187729	Activated	FALSE
Vasopressin-regulated water reabsorption	0.3015	0.0952416	0.3235191	Activated	FALSE
Ferroptosis	0.3148	0.0974137	0.3255310	Activated	FALSE
Apoptosis	0.3510	0.0988762	0.3255310	Activated	FALSE
Insulin signaling pathway	-0.3454	0.1005605	0.3260598	Inhibited	FALSE
Gap junction	-0.2715	0.1052140	0.3360568	Inhibited	FALSE
ErbB signaling pathway	-0.3217	0.1070099	0.3367664	Inhibited	FALSE
Prolactin signaling pathway	0.3137	0.1090413	0.3381862	Activated	FALSE
Natural killer cell mediated cytotoxicity	-0.3447	0.1143208	0.3396919	Inhibited	FALSE
Longevity regulating pathway - multiple species	0.2782	0.1152363	0.3396919	Activated	FALSE
Oocyte meiosis	-0.2866	0.1168138	0.3396919	Inhibited	FALSE
Systemic lupus erythematosus	0.2844	0.1169634	0.3396919	Activated	FALSE
Ras signaling pathway	0.3548	0.1174636	0.3396919	Activated	FALSE
Legionellosis	0.3318	0.1206566	0.3416993	Activated	FALSE
Regulation of actin cytoskeleton	-0.3046	0.1219023	0.3416993	Inhibited	FALSE
Epstein-Barr virus infection	-0.3856	0.1230569	0.3416993	Inhibited	FALSE
Choline metabolism in cancer	-0.3105	0.1245446	0.3416993	Inhibited	FALSE
Apelin signaling pathway	0.2871	0.1330388	0.3603836	Activated	FALSE
Adherens junction	0.3023	0.1362206	0.3643900	Activated	FALSE
Alcoholism	0.2735	0.1457567	0.3850857	Activated	FALSE
Alcoholic liver disease	0.3097	0.1518312	0.3962424	Activated	FALSE
Central carbon metabolism in cancer	-0.2967	0.1583378	0.4037397	Inhibited	FALSE
Bacterial invasion of epithelial cells	0.2818	0.1584772	0.4037397	Activated	FALSE
Aldosterone synthesis and secretion	0.2538	0.1613727	0.4062796	Activated	FALSE
AGE-RAGE signaling pathway in diabetic complications	0.2664	0.1659032	0.4128290	Activated	FALSE
Hippo signaling pathway - multiple species	-0.2509	0.1702716	0.4188290	Inhibited	FALSE
NF-kappa B signaling pathway	0.3130	0.1759813	0.4279546	Activated	FALSE

Table 2. . Detection of stress-associated and proliferative DNA repair signatures, as reported by Zhang et al., 2022, in significantly perturbed pathways identified by sSNAPPY.

Marker Type	Gene Name	Gene-set Name
Stress-associated	GADD45B	p53 signaling pathway; Cellular senescence
Stress-associated	TNF	Pathogenic Escherichia coli infection, Salmonella infection, Toxoplasmosis, Yersinia infection, Human T-cell leukemia virus 1 infection, Antigen processing and presentation, Osteoclast differentiation, Human immunodeficiency virus 1 infection, Fluid shear stress and atherosclerosis, Chagas disease
Stress-associated	JUN	Pathogenic Escherichia coli infection, Salmonella infection, Yersinia infection, Human T-cell leukemia virus 1 infection, Pertussis, Relaxin signaling pathway, Osteoclast differentiation, Th17 cell differentiation, Rheumatoid arthritis, Human immunodeficiency virus 1 infection, B cell receptor signaling pathway, Fluid shear stress and atherosclerosis, Chagas disease
Stress-associated	FOS	Pathogenic Escherichia coli infection, Salmonella infection, Parathyroid hormone synthesis, secretion and action, Yersinia infection, Human T-cell leukemia virus 1 infection, Pertussis, Relaxin signaling pathway, Osteoclast differentiation, Th17 cell differentiation, Rheumatoid arthritis, Human immunodeficiency virus 1 infection, B cell receptor signaling pathway, Fluid shear stress and atherosclerosis, Chagas disease
Stress-associated	IL6	Pathogenic Escherichia coli infection, Salmonella infection, Yersinia infection, Human T-cell leukemia virus 1 infection, Th17 cell differentiation, Chagas disease
Stress-associated	CXCR4	Human immunodeficiency virus 1 infection, Leukocyte transendothelial migration
Proliferative DNA repair	GADD45B	p53 signaling pathway; Cellular senescence
Proliferative DNA repair	TNF	Pathogenic Escherichia coli infection, Salmonella infection, Toxoplasmosis, Yersinia infection, Human T-cell leukemia virus 1 infection, Antigen processing and presentation, Osteoclast differentiation, Human immunodeficiency virus 1 infection, Fluid shear stress and atherosclerosis, Chagas disease
Proliferative DNA repair	JUN	Pathogenic Escherichia coli infection, Salmonella infection, Yersinia infection, Human T-cell leukemia virus 1 infection, Pertussis, Relaxin signaling pathway, Osteoclast differentiation, Th17 cell differentiation, Rheumatoid arthritis, Human immunodeficiency virus 1 infection, B cell receptor signaling pathway, Fluid shear stress and atherosclerosis, Chagas disease
Proliferative DNA repair	FOS	Pathogenic Escherichia coli infection, Salmonella infection, Parathyroid hormone synthesis, secretion and action, Yersinia infection, Human T-cell leukemia virus 1 infection, Pertussis, Relaxin signaling pathway, Osteoclast differentiation, Th17 cell differentiation, Rheumatoid arthritis, Human immunodeficiency virus 1 infection, B cell receptor signaling pathway, Fluid shear stress and atherosclerosis, Chagas disease
Proliferative DNA repair	IL6	Pathogenic Escherichia coli infection, Salmonella infection, Yersinia infection, Human T-cell leukemia virus 1 infection, Th17 cell differentiation, Chagas disease
Proliferative DNA repair	CXCR4	Human immunodeficiency virus 1 infection, Leukocyte transendothelial migration