Analyze CRISPRi growth competition data for growth at high CO2 gas feeds

Ute Hoffmann (Science For Life Laboratory (KTH), Stockholm, Sweden)

januari 18, 2024

Contents

1 Aim of the analysis

Basic visualization of CRISPRi data for cultivation with lysine. Data analysis was performed using nf-core-crispriscreen pipeline (https://github.com/MPUSP/nf-core-crispriscreen).

2 Analysis

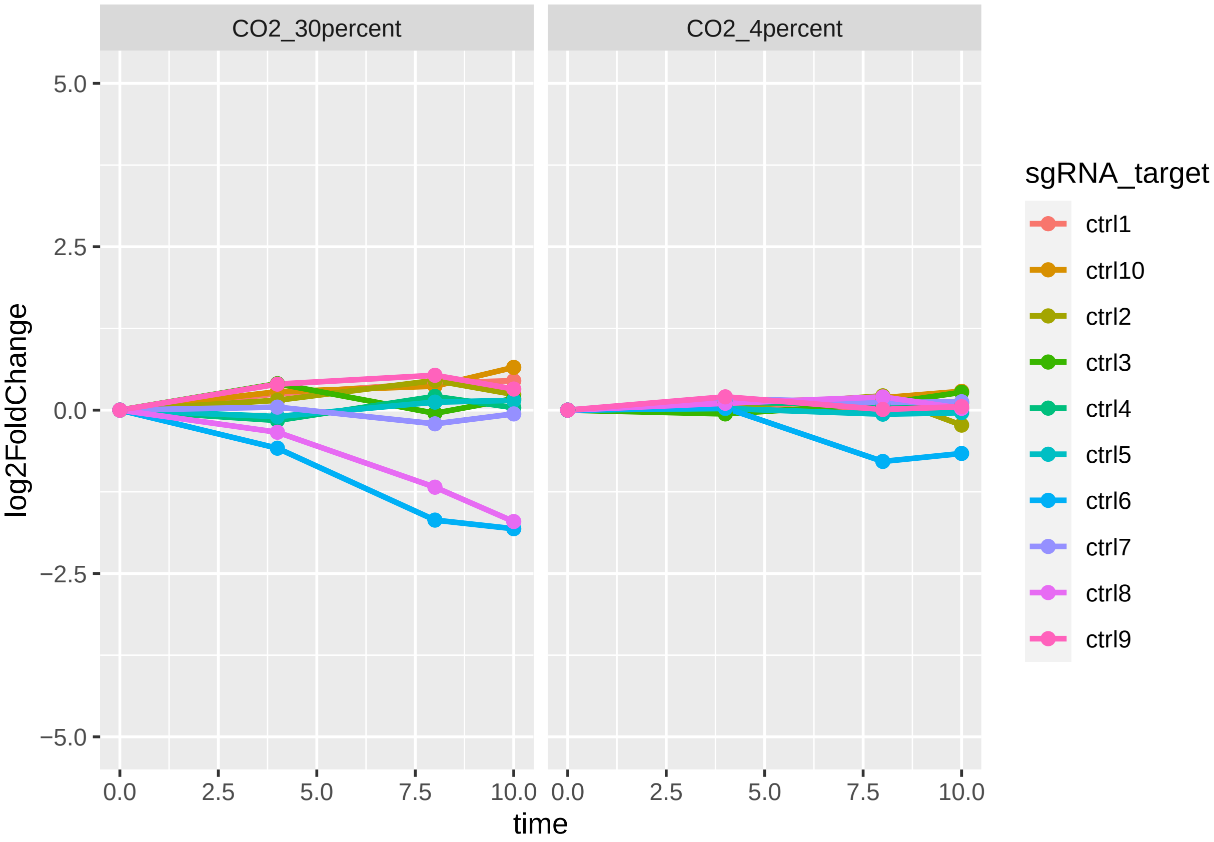
In a first step, the results given by the Nextflow pipeline are loaded.

```
load("../results/fitness/result.Rdata")
```

2.1 Diagnostic plot to check if control sgRNAs look ok

Several control sgRNAs are included in the CRISPRi library. These control sgRNAs do not target any specific gene and serve as a control.

```
plot_controls_sgRNAs <- DESeq_result_table %>% filter(grepl("ctrl", sgRNA_target)) %>%
  ggplot(aes(x = time, y = log2FoldChange, color = sgRNA_target)) +
  geom_line(linewidth = 1) + geom_point(size = 2) + ylim(-5, 5) + facet_wrap(~ condition, ncol = 4)
print(plot_controls_sgRNAs)
```



```
ggsave("../R_results/plot_control_sgRNAs.pdf", plot=plot_controls_sgRNAs, width=12, height=12, units="cm
```

2.2 Add annotation to results tables

In the following, annotation is added to the results table provided by the Nextflow pipeline. Mapping of the sgRNA targets to slr-locus tags is given in this file, downloaded on 24/02/23: https://github.com/m-jahn/R-notebook-crispri-lib/blob/master/sgRNA_library_V2/data/input/mapping_trivial_names.tsv The appended annotation is based on Uniprot and Cyanobase, partially edited manually. The table used for annotation was created beginning of 2021. Therefore, it does not include several genes which were only recently characterized. For a detailed description of all the columns given in the results tables, consult https://mpusp.github.io/nf-core-crispriscreen/output or https://www.biorxiv.org/content/10.1101/2023.02.13.528328v1.full.pdf+htmls

```
mapping_gene_locus <- read_tsv("../input/2023-02-24_mapping_trivial_names.tsv", show_col_types=FALSE)
names(mapping_gene_locus) <- c("sgRNA_target", "locus")
DESeq_result_table <- DESeq_result_table %>% left_join(mapping_gene_locus)
```

```
annotation <- read_tsv("../input/annotation_locusTags_stand13012021.csv", show_col_types = FALSE)
annotation_2 <- annotation[,c(1,2,3)]
names(annotation_2) <- c("locus", "Gene name","Product")
DESeq_result_table <- DESeq_result_table %>% left_join(annotation_2)
```

```
write_tsv(DESeq_result_table, file="../R_results/annotated_DESeq_result_table.tsv")
df_reduced_info <- unique(subset(DESeq_result_table, DESeq_result_table$time==8 | DESeq_result_table$tim
write_tsv(df_reduced_info, file="../R_results/Reduced_annotated_DESeq_result_table.tsv")
```

```
df_red_wide <- pivot_wider(df_reduced_info, names_from=condition, values_from=c(wmean_fitness, sd_fitnes
write_tsv(df_red_wide, file="../R_results/Wide_DESeq_result_table.tsv")
```

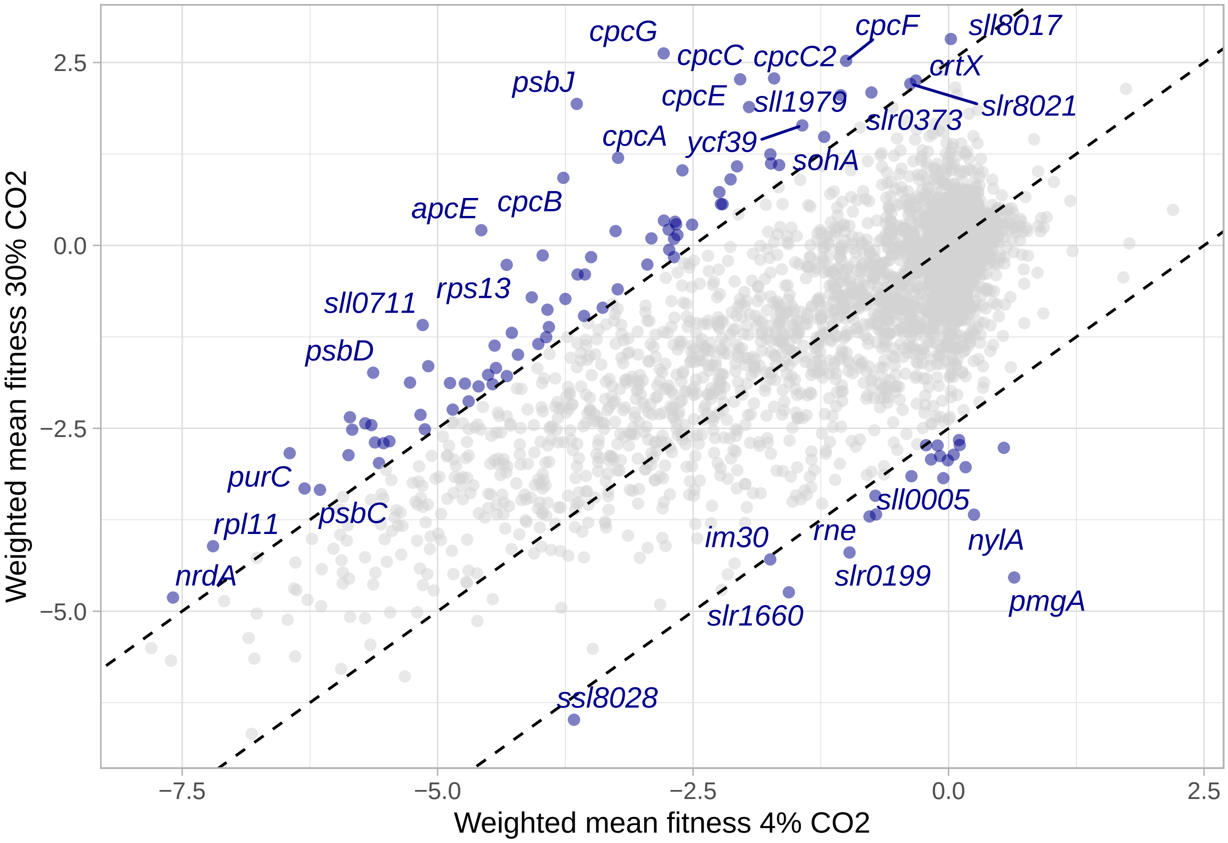
2.3 Visualization

The weighted mean fitness value combines the values of the different sgRNAs targeting the same gene. Fitness-fitness plots were created to identify genes which behave differently at different gas conditions. This was performed separately for ncRNAs and protein-coding genes.

2.3.1 Protein-coding genes

```
df_reduced <- unique(subset(DESeq_result_table, DESeq_result_table$time==8)[,c(2,4,20)])
df_red_ncRNAs <- subset(df_reduced, grepl("nc_", df_reduced$sgRNA_target))
df_red_no_ncRNAs <- subset(df_reduced, !grepl("nc_", df_reduced$sgRNA_target))
df_red_wide <- pivot_wider(df_red_no_ncRNAs,names_from="condition", values_from=c("wmean_fitness"))
```

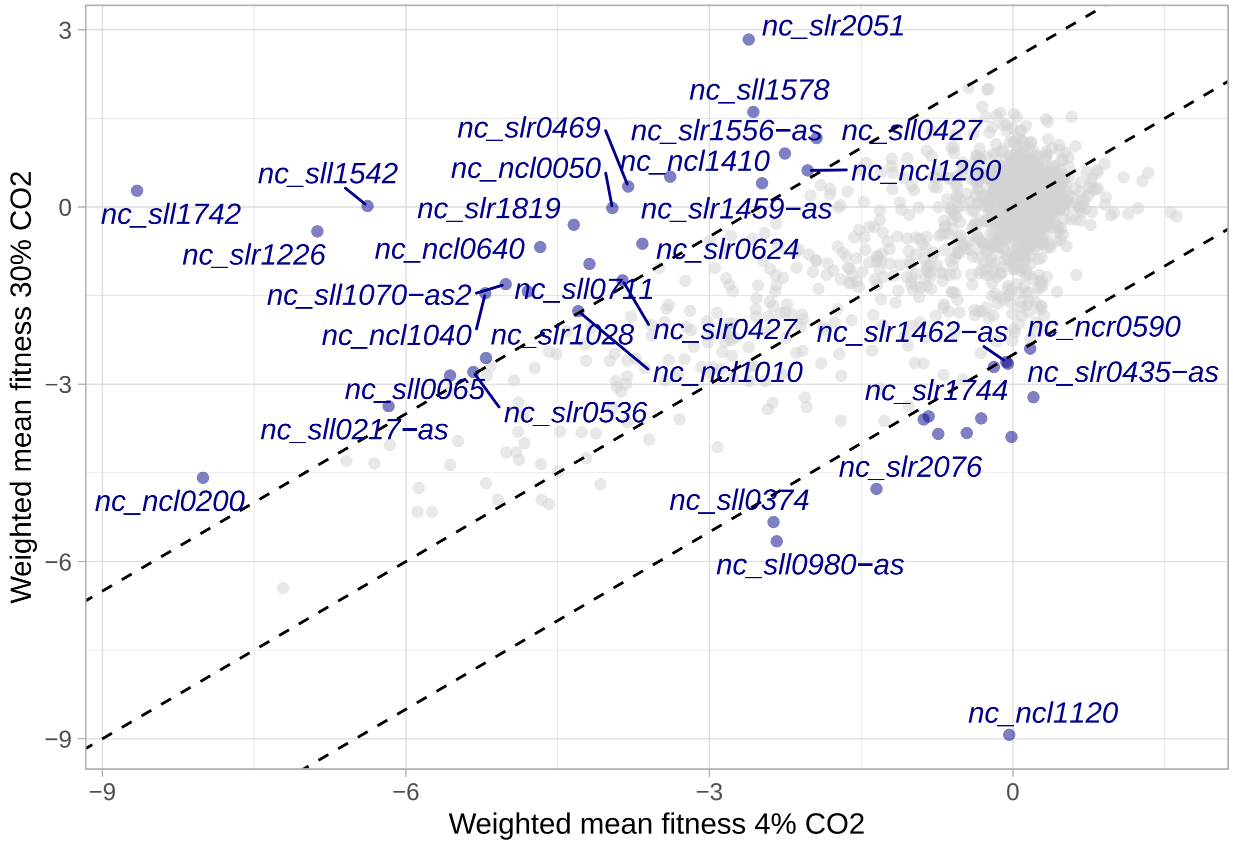
```
plot_fitness_fitness <- function(df_input, y_axis, y_axis_label, x_axis="CO2_4percent", x_axis_label="We
  df_input$diff <- "NO"
  df_input$diff[(df_input[[x_axis]] - df_input[[y_axis]] > 2.5) | (df_input[[x_axis]] - df_input[[y_axi
  # prepare labels for plot
  df_input$delabel <- NA
  df_input$delabel[df_input$diff !="NO"] <- df_input$sgRNA_target[df_input$diff != "NO"]
  mycolors <- c("darkblue",  "#d3d3d3b2")
  names(mycolors) <- c("YES", "NO")
  p <- ggplot(data=df_input, aes(x=eval(parse(text=x_axis)), y=eval(parse(text=y_axis)), label=delabel,
    theme_light() + labs(y=y_axis_label, x=x_axis_label) + theme(legend.position = "none") + geom_ablin
  ggsave(filename = filename_save, plot=p, width=12, height=12, units="cm")
return(p)
}

plot_fitness_fitness(df_red_wide, "CO2_30percent", y_axis_label="Weighted mean fitness 30% CO2", filenam
```



2.3.2 ncRNAs

These include antisense RNAs, but also other ncRNAs.

```
df_red_wide_ncRNA <- pivot_wider(df_red_ncRNAs,names_from="condition", values_from=c("wmean_fitness"))
plot_fitness_fitness(df_red_wide_ncRNA, "CO2_30percent", y_axis_label="Weighted mean fitness 30% CO2", 
```



2.4 GSEA

Functional enrichment analyses and gene set enrichment analyses help to check if a certain pathway or specific group of genes is especially affected by a treatment. Here, gene set enrichment analyses were performed for either Gene Ontology terms or KEGG pathways. To perform a gene set enrichment analysis, genes are sorted according to some measure, e.g. the log2FC after a certain time or the calculated fitness. Here, we used the weighted fitness of several sgRNAs as measure. The mapping of locus tags to Gene Ontology terms was downloaded from UniProt on the 18th Jan. 2024. There is the possibility to somehow weigh the adjusted p value in these calculation, e.g. by multiplying the weighted mean with the adjusted p value. Here, only the first few rows of each table is given. Full tables with all found terms/pathways are available.

In a first step, GSEAs were calculated for the two different CRISPRi libraries separately. The depletion of essential pathways related to “Ribosomes” or “photosynthesis” is a first good quality measure for a CRISPRi screen.

2.4.1 4% CO2

```
DESeq_result_table_4percent <- unique(subset(DESeq_result_table, DESeq_result_table$condition=="CO2_4pe
geneList <- DESeq_result_table_4percent$wmean_fitness
names(geneList) <- DESeq_result_table_4percent$locus
geneList = sort(geneList, decreasing = TRUE)
set.seed(513)
go_gsea_object <- GSEA(geneList, TERM2GENE = term_to_gene, TERM2NAME=term_to_name, seed=TRUE)
print(head(go_gsea_object)[,columns_to_show])
```

```
##                                        Description setSize enrichmentScore       NES
## GO:0003735 structural constituent of ribosome      56      -0.8538492 -1.519966
## GO:0006412                        translation      63      -0.8300411 -1.478153
## GO:0005737                          cytoplasm     304      -0.7027158 -1.280898
## GO:0005524                        ATP binding     302      -0.6828270 -1.244631
## GO:0005829                            cytosol     222      -0.7005036 -1.272217
## GO:0019843                       rRNA binding      36      -0.8460258 -1.489785
##                p.adjust       qvalue
## GO:0003735 4.939268e-09 3.523444e-09
## GO:0006412 4.939268e-09 3.523444e-09
## GO:0005737 4.939268e-09 3.523444e-09
## GO:0005524 1.640107e-08 1.169976e-08
## GO:0005829 1.905868e-08 1.359558e-08
## GO:0019843 7.048019e-06 5.027730e-06
```

```
write.csv(go_gsea_object, "../R_results/GSEA_output/GO_GSEA_CRISPRi_4percent.csv")

set.seed(914)
kegg_gsea_object <- gseKEGG(geneList, organism="syn", minGSSize=10, pvalueCutoff = 0.05, seed=TRUE)
print(head(kegg_gsea_object)[,columns_to_show_KEGG])
```

```
##                ID
## syn03010 syn03010
## syn01110 syn01110
## syn01230 syn01230
## syn01120 syn01120
## syn01232 syn01232
## syn01240 syn01240
##                                                                         Description
## syn03010                                     Ribosome - Synechocystis sp. PCC 6803
## syn01110        Biosynthesis of secondary metabolites - Synechocystis sp. PCC 6803
## syn01230                  Biosynthesis of amino acids - Synechocystis sp. PCC 6803
## syn01120 Microbial metabolism in diverse environments - Synechocystis sp. PCC 6803
## syn01232                        Nucleotide metabolism - Synechocystis sp. PCC 6803
## syn01240                    Biosynthesis of cofactors - Synechocystis sp. PCC 6803
##          setSize enrichmentScore     p.adjust
## syn03010      54      -0.8631470 3.300000e-09
## syn01110     287      -0.7502957 3.300000e-09
## syn01230      93      -0.7804653 4.116630e-09
## syn01120     129      -0.7280589 4.347612e-07
## syn01232      27      -0.8532515 3.768218e-05
## syn01240     138      -0.7006127 3.768218e-05
```

```
write.csv(kegg_gsea_object, "../R_results/GSEA_output/KEGG_GSEA_CRISPRi_4percent.csv")
```

```
DESeq_result_table_30percent <- subset(DESeq_result_table, DESeq_result_table$condition=="CO2_4percent"
geneList <- DESeq_result_table_30percent$fitness
names(geneList) <- DESeq_result_table_30percent$locus
geneList = sort(geneList, decreasing = TRUE)
go_gsea_object <- GSEA(geneList, TERM2GENE = term_to_gene, TERM2NAME=term_to_name, seed=TRUE)
print(head(go_gsea_object)[,1:10])
```

```
##                    ID                               Description setSize
## GO:0016787 GO:0016787                        hydrolase activity      26
## GO:0003700 GO:0003700 DNA-binding transcription factor activity      29
## GO:0051607 GO:0051607                                GO:0051607      15
## GO:0000160 GO:0000160   phosphorelay signal transduction system      30
## GO:0003824 GO:0003824                        catalytic activity      21
## GO:0022857 GO:0022857        transmembrane transporter activity      19
##            enrichmentScore      NES       pvalue     p.adjust       qvalue rank
## GO:0016787       0.7478879 2.420940 1.744271e-07 1.831484e-05 1.248531e-05 4642
## GO:0003700       0.7010921 2.158597 1.599747e-06 8.398671e-05 5.725410e-05 4241
## GO:0051607       0.8092227 2.254130 6.046637e-06 2.116323e-04 1.442706e-04 4255
## GO:0000160       0.6739676 2.053891 1.063025e-05 2.790439e-04 1.902254e-04 5090
## GO:0003824       0.6961460 2.218995 2.324083e-04 4.880575e-03 3.327109e-03 4144
## GO:0022857       0.6800678 2.089564 3.797985e-04 6.646473e-03 4.530929e-03 4185
##                               leading_edge
## GO:0016787  tags=12%, list=22%, signal=9%
## GO:0003700 tags=13%, list=20%, signal=11%
## GO:0051607  tags=11%, list=20%, signal=9%
## GO:0000160 tags=20%, list=24%, signal=15%
## GO:0003824  tags=11%, list=19%, signal=9%
## GO:0022857 tags=14%, list=19%, signal=11%
```

2.4.2 30% CO2

```
DESeq_result_table_30percent <- unique(subset(DESeq_result_table, DESeq_result_table$condition=="CO2_30p
geneList <- DESeq_result_table_30percent$wmean_fitness
names(geneList) <- DESeq_result_table_30percent$locus
geneList = sort(geneList, decreasing = TRUE)
set.seed(513)
go_gsea_object <- GSEA(geneList, TERM2GENE = term_to_gene, TERM2NAME=term_to_name, seed=TRUE)
print(head(go_gsea_object)[,columns_to_show])
```

```
##                                     Description setSize enrichmentScore       NES
## GO:0005737                            cytoplasm     304      -0.6264218 -1.460623
## GO:0005524                          ATP binding     302      -0.5940986 -1.385151
## GO:0006412                          translation      63      -0.7342477 -1.649999
## GO:0005829                              cytosol     222      -0.6023640 -1.397581
## GO:0003735 structural constituent of ribosome      56      -0.7463098 -1.674496
## GO:0008360                regulation of cell shape      18      -0.8770575 -1.818224
##                p.adjust       qvalue
## GO:0005737 1.210000e-08 8.315789e-09
## GO:0005524 2.418278e-08 1.661975e-08
## GO:0006412 4.519931e-07 3.106347e-07
## GO:0005829 4.519931e-07 3.106347e-07
## GO:0003735 4.815568e-07 3.309525e-07
## GO:0008360 7.914863e-06 5.439532e-06
```

```
write.csv(go_gsea_object, "../R_results/GSEA_output/GO_GSEA_CRISPRi_30percent.csv")

set.seed(914)
kegg_gsea_object <- gseKEGG(geneList, organism="syn", minGSSize=10, pvalueCutoff = 0.05, seed=TRUE)
print(head(kegg_gsea_object)[,columns_to_show_KEGG])
```

```
##                ID
## syn01240 syn01240
## syn01110 syn01110
## syn01230 syn01230
## syn01120 syn01120
## syn03010 syn03010
## syn01210 syn01210
##                                                                         Description
## syn01240                    Biosynthesis of cofactors - Synechocystis sp. PCC 6803
## syn01110        Biosynthesis of secondary metabolites - Synechocystis sp. PCC 6803
## syn01230                  Biosynthesis of amino acids - Synechocystis sp. PCC 6803
## syn01120 Microbial metabolism in diverse environments - Synechocystis sp. PCC 6803
## syn03010                                     Ribosome - Synechocystis sp. PCC 6803
## syn01210              2-Oxocarboxylic acid metabolism - Synechocystis sp. PCC 6803
##          setSize enrichmentScore     p.adjust
## syn01240     138      -0.6721127 3.250000e-09
## syn01110     287      -0.6598737 3.250000e-09
## syn01230      93      -0.7123944 4.054257e-09
## syn01120     129      -0.6479570 1.605638e-07
## syn03010      54      -0.7504348 1.677401e-07
## syn01210      26      -0.8254895 2.372095e-06
```

```
write.csv(kegg_gsea_object, "../R_results/GSEA_output/KEGG_GSEA_CRISPRi_30percent.csv")
```

2.4.3 Difference between 4% and 30%

In a next step, we tried to check which GO terms or KEGG pathways show a divergent enrichment or depletion in the two libraries. For this, weighted fitness means belonging to the two conditions were subtracted from each other. These differences were used as input for the GSEA.

```
df_difference <- unique(subset(DESeq_result_table, DESeq_result_table$time==8 & !is.na(DESeq_result_tabl
df_difference_wide <- pivot_wider(df_difference, names_from=condition, values_from=wmean_fitness)
df_difference_wide$difference <- df_difference_wide$CO2_4percent - df_difference_wide$CO2_30percent

geneList <- df_difference_wide$difference
names(geneList) <- df_difference_wide$locus
geneList = sort(geneList, decreasing = TRUE)
set.seed(513)
go_gsea_object <- GSEA(geneList, TERM2GENE = term_to_gene, TERM2NAME=term_to_name, seed=TRUE)
print(head(go_gsea_object)[,columns_to_show])
```

```
##                                           Description setSize enrichmentScore
## GO:0006412                                translation      63      -0.7849345
## GO:0003735         structural constituent of ribosome      56      -0.7926398
## GO:0031676 plasma membrane-derived thylakoid membrane     114      -0.6789756
## GO:0030089                              phycobilisome      24      -0.8446239
## GO:0015979                             photosynthesis      82      -0.6267800
## GO:0022625          cytosolic large ribosomal subunit      20      -0.8085764
##                  NES     p.adjust       qvalue
## GO:0006412 -2.373268 4.000000e-09 3.298246e-09
## GO:0003735 -2.338241 4.000000e-09 3.298246e-09
## GO:0031676 -2.208104 4.000000e-09 3.298246e-09
## GO:0030089 -2.154821 3.587417e-07 2.958046e-07
## GO:0015979 -1.962030 4.378488e-06 3.610332e-06
## GO:0022625 -2.017168 8.170327e-05 6.736936e-05
```

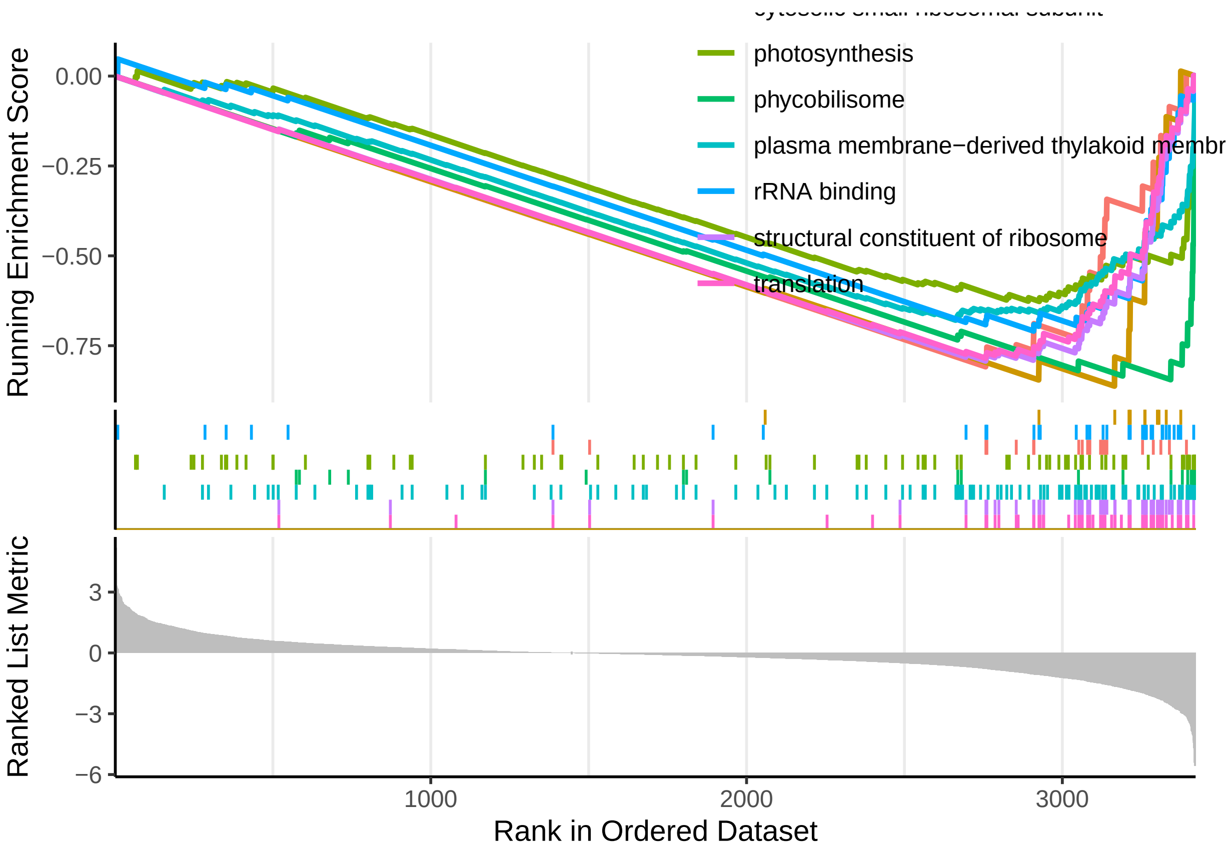
```
write.csv(go_gsea_object, "../R_results/GSEA_output/GO_GSEA_difference.csv")

set.seed(914)
kegg_gsea_object <- gseKEGG(geneList, organism="syn", minGSSize=10, pvalueCutoff = 0.05, seed=TRUE)
print(head(kegg_gsea_object)[,columns_to_show_KEGG])
```

```
##                ID
## syn03010 syn03010
## syn01110 syn01110
## syn00196 syn00196
## syn01230 syn01230
## syn00195 syn00195
## syn00230 syn00230
##                                                                  Description
## syn03010                              Ribosome - Synechocystis sp. PCC 6803
## syn01110 Biosynthesis of secondary metabolites - Synechocystis sp. PCC 6803
## syn00196     Photosynthesis - antenna proteins - Synechocystis sp. PCC 6803
## syn01230           Biosynthesis of amino acids - Synechocystis sp. PCC 6803
## syn00195                        Photosynthesis - Synechocystis sp. PCC 6803
## syn00230                     Purine metabolism - Synechocystis sp. PCC 6803
##          setSize enrichmentScore     p.adjust
## syn03010      54      -0.8042460 3.300000e-09
## syn01110     287      -0.5433327 3.300000e-09
## syn00196      15      -0.9230557 1.345852e-08
## syn01230      93      -0.6525954 1.815761e-08
## syn00195      63      -0.6436630 2.007166e-05
## syn00230      45      -0.6188541 1.890343e-03
```

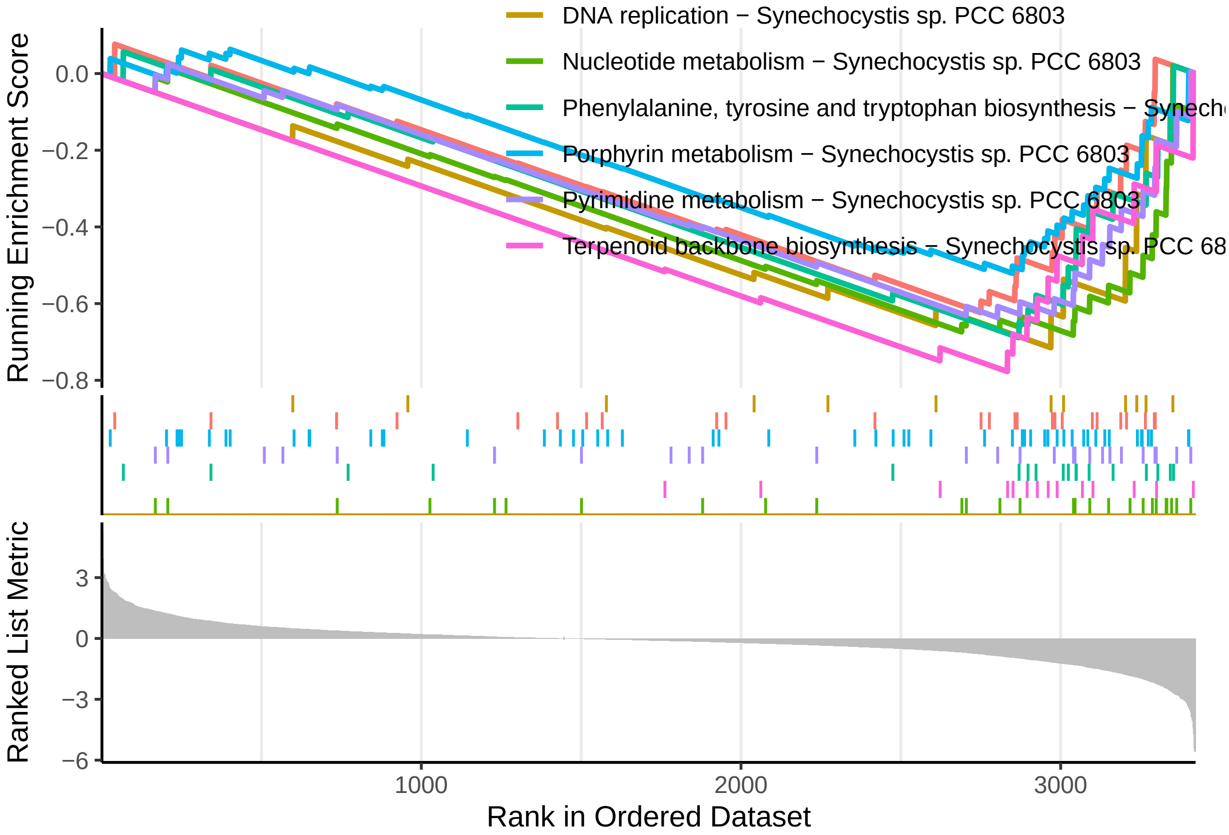
```
write.csv(kegg_gsea_object, "../R_results/GSEA_output/KEGG_GSEA_difference.csv")
```

Results for the comparison of the two libraries were visualized. These plots are separated in three panels. The lowest panel ("Ranked List Metric") shows the metric according to which the genes were sorted. In this case, this was the weighted fitness mean associated with the different genes. The upper panel shows the running enrichment score of terms/pathways which were enriched/depleted in a statistically significant manner. The middle panel shows where the genes associated with these terms are located within the ranked list of genes in the same color code as used in the upper panel.

```
p <- gseaplot2(go_gsea_object, geneSetID =1:8)
p
```



```
ggsave("../R_results/GSEA_output/KEGG_GSEA_differences_part1.pdf", plot=p, width=20, height=25, units="
p <- gseaplot2(kegg_gsea_object, geneSetID =9:15)
p
```



```
ggsave("../R_results/GSEA_output/KEGG_GSEA_differences_part2.pdf", plot=p, width=20, height=25, units="
```

Session info

```
## R version 4.3.2 (2023-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.3 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so;  LAPACK version 3.10.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=sv_SE.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=sv_SE.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=sv_SE.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=sv_SE.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Stockholm
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
```

```
##  [1] enrichplot_1.20.3     clusterProfiler_4.8.3 magrittr_2.0.3
##  [4] forcats_0.5.2         stringr_1.5.0         dplyr_1.0.10
##  [7] purrr_1.0.2           readr_2.1.4           tidyr_1.3.0
## [10] tibble_3.2.1          tidyverse_1.3.1       ggpubr_0.6.0
## [13] ggrepel_0.9.4         ggplot2_3.4.4         knitr_1.45
##
## loaded via a namespace (and not attached):
##   [1] RColorBrewer_1.1-3      rstudioapi_0.14         jsonlite_1.8.7
##   [4] farver_2.1.1            rmarkdown_2.25          ragg_1.2.6
##   [7] fs_1.6.3                zlibbioc_1.46.0         vctrs_0.6.4
##  [10] memoise_2.0.1           RCurl_1.98-1.13         ggtree_3.8.2
##  [13] rstatix_0.7.2           htmltools_0.5.7         haven_2.5.1
##  [16] broom_1.0.1             cellranger_1.1.0        gridGraphics_0.5-1
##  [19] plyr_1.8.9              lubridate_1.8.0         cachem_1.0.8
##  [22] igraph_1.5.1            lifecycle_1.0.4         pkgconfig_2.0.3
##  [25] gson_0.1.0              Matrix_1.6-3            R6_2.5.1
##  [28] fastmap_1.1.1           GenomeInfoDbData_1.2.10 digest_0.6.33
##  [31] aplot_0.2.2             colorspace_2.1-0        patchwork_1.1.3
##  [34] AnnotationDbi_1.62.2    S4Vectors_0.38.2        textshaping_0.3.7
##  [37] RSQLite_2.3.3           labeling_0.4.2          fansi_1.0.5
##  [40] httr_1.4.4              polyclip_1.10-6         abind_1.4-5
##  [43] compiler_4.3.2          bit64_4.0.5             withr_2.5.0
##  [46] downloader_0.4          backports_1.4.1         BiocParallel_1.34.2
##  [49] carData_3.0-5           viridis_0.6.4           DBI_1.1.3
##  [52] highr_0.10              ggforce_0.4.1           ggsignif_0.6.4
##  [55] MASS_7.3-60             HDO.db_0.99.1           tools_4.3.2
##  [58] scatterpie_0.2.1        ape_5.7-1               glue_1.6.2
##  [61] nlme_3.1-163            GOSemSim_2.26.1         shadowtext_0.1.2
##  [64] grid_4.3.2              reshape2_1.4.4          fgsea_1.26.0
##  [67] generics_0.1.3          gtable_0.3.4            tzdb_0.4.0
##  [70] data.table_1.14.8       hms_1.1.3               tidygraph_1.2.3
##  [73] xml2_1.3.5              car_3.1-2               utf8_1.2.4
##  [76] XVector_0.40.0          BiocGenerics_0.46.0     pillar_1.9.0
##  [79] vroom_1.6.4             yulab.utils_0.1.0       splines_4.3.2
##  [82] tweenr_2.0.2            treeio_1.24.3           lattice_0.22-5
##  [85] bit_4.0.5               tidyselect_1.2.0        GO.db_3.17.0
##  [88] Biostrings_2.68.1       gridExtra_2.3           IRanges_2.34.1
##  [91] stats4_4.3.2            xfun_0.41               graphlayouts_1.0.2
##  [94] Biobase_2.60.0          stringi_1.7.12          lazyeval_0.2.2
##  [97] ggfun_0.1.3             yaml_2.3.7              evaluate_0.23
## [100] codetools_0.2-19        ggraph_2.1.0            qvalue_2.32.0
## [103] ggplotify_0.1.2         cli_3.6.1               systemfonts_1.0.4
## [106] munsell_0.5.0           modelr_0.1.9            Rcpp_1.0.11
## [109] GenomeInfoDb_1.36.4     readxl_1.4.3            dbplyr_2.2.1
## [112] png_0.1-8               parallel_4.3.2          assertthat_0.2.1
## [115] blob_1.2.3              DOSE_3.26.2             reprex_2.0.2
## [118] bitops_1.0-7            viridisLite_0.4.2       tidytree_0.4.5
## [121] scales_1.2.1            crayon_1.5.2            rlang_1.1.2
## [124] cowplot_1.1.2           fastmatch_1.1-4         KEGGREST_1.40.1
## [127] rvest_1.0.3
```