suppressPackageStartupMessages({

library(dplyr); library(stringr); library(tidyr)

library(readxl); library(clusterProfiler); library(enrichplot); library(ggplot2)

})

## 路径设置 -------------------------------------------------------------

kegg\_annotation\_path <- "/home/data/t210549/liuhuacheng/wrky注释/全基因组/kegg\_ko\_long.csv"

module\_gene\_path <- "/home/data/t210549/liuhuacheng/wrky注释/全基因组/性状相关模块/KEGG/Module\_cyan\_gene.xlsx"

out\_prefix <- "/home/data/t210549/liuhuacheng/wrky注释/全基因组/性状相关模块/KEGG/Module\_cyan\_KEGG\_KO"

## 1️⃣ 读取 KEGG 注释表（long 格式）

anno <- read.csv(kegg\_annotation\_path, header = TRUE, check.names = FALSE) %>%

rename(Gene\_ID = gene, KEGG\_ko = kegg\_ko) %>%

filter(!is.na(KEGG\_ko), KEGG\_ko != "-") %>%

mutate(

KEGG\_ko = str\_replace\_all(KEGG\_ko, "^ko:", ""),

KEGG\_ko = str\_trim(KEGG\_ko)

) %>%

filter(str\_detect(KEGG\_ko, "^K\\d+")) %>%

distinct(Gene\_ID, KEGG\_ko)

cat(sprintf("[INFO] Gene–KO 映射：%d 行，基因数=%d，KO数=%d\n",

nrow(anno), n\_distinct(anno$Gene\_ID), n\_distinct(anno$KEGG\_ko)))

## 2️⃣ 读取模块基因列表（替换 DEG）

module\_genes <- read\_excel(module\_gene\_path) %>%

pull(GENEID) %>% as.character() %>% unique() %>% na.omit()

cat(sprintf("[INFO] 模块基因数：%d 个\n", length(module\_genes)))

## 3️⃣ 映射 KEGG KO -----------------------------------------------------

deg\_ko <- anno %>% filter(Gene\_ID %in% module\_genes) %>% pull(KEGG\_ko) %>% unique()

bg\_ko <- anno$KEGG\_ko %>% unique()

cat(sprintf("[INFO] 模块基因映射到 KO：%d 个；背景 KO：%d 个\n", length(deg\_ko), length(bg\_ko)))

if (length(deg\_ko) < 10) stop("映射到 KO 的基因太少（<10），无法进行 KEGG 富集。")

## 4️⃣ 运行 KEGG 富集分析 ----------------------------------------------

ek <- enrichKEGG(

gene = deg\_ko,

organism = "ko",

keyType = "kegg",

universe = bg\_ko,

pvalueCutoff = 0.2,

qvalueCutoff = 0.5

)

res <- as.data.frame(ek)

if (nrow(res) == 0) stop("当前阈值下无显著通路，建议暂时把 pvalueCutoff 提到 0.2 观察趋势。")

# =====================================================

# 🚫 去除非植物通路（如人类疾病、动物代谢）

# =====================================================

# 🚫 去除动物/人类相关通路

remove\_keywords <- c(

"cancer", "atherosclerosis", "tuberculosis", "TGF", "Apelin",

"Influenza", "Alcoholism", "HIF"

)

# 使用不区分大小写的正则过滤

res <- res %>%

filter(!str\_detect(Description, regex(paste(remove\_keywords, collapse = "|"), ignore\_case = TRUE)))

cat(sprintf("[INFO] 已去除动物/人类通路后剩余 %d 条记录。\n", nrow(res)))

res <- res %>%

filter(!str\_detect(tolower(Description), paste(remove\_keywords, collapse = "|")))

if (nrow(res) == 0) stop("过滤后无植物相关通路，请放宽阈值或检查注释来源。")

# 同步更新富集对象

ek@result <- res

cat("[INFO] 去除动物/疾病通路后保留：", nrow(res), " 条。\n")

# 示例输出

cat("[INFO] 过滤后示例结果：\n")

print(head(res[, c("ID", "Description", "pvalue", "p.adjust", "GeneRatio")]))

## 5️⃣ 绘图与结果输出 ---------------------------------------------------

p1 <- dotplot(ek, showCategory = 20, label\_format = 60) +

ggtitle("KEGG Enrichment: Module Cyan")

ggsave(paste0(out\_prefix, "\_dotplot.pdf"), p1, width = 8, height = 9, dpi = 300)

ggsave(paste0(out\_prefix, "\_dotplot.png"), p1, width = 8, height = 9, dpi = 300)

write.csv(res, paste0(out\_prefix, "\_results.csv"), row.names = FALSE)

cat("[DONE] ✅ 富集结果已输出：\n",

" - ", paste0(out\_prefix, "\_results.csv"), "\n",

" - ", paste0(out\_prefix, "\_dotplot.pdf / .png"), "\n",

"通路名称（Description）已作为 y 轴标签。\n")