# Day 2 – 02 Simple Heatmap (Exercises)

Seminar practice worksheet

This worksheet mirrors the guided heatmap build but uses the three-genome dataset (`second_day_part2/data/dataset2_subs`
Work through the prompts, typing your own code where `# TODO` markers appear. The goal is to recreate the full pipeline: load data, explore it briefly, build the `ComplexHeatmap`, and save the PDF.

> **Tip:** Ensure packages from `00_prepare.Rmd` are installed, and regenerate the dataset via `data/00_prepare_dataset.Rmd` if needed. A worked key lives in `scripts/02_simple_heatmap_exercises_solution.R` check it only after trying on your own.

## 1. Load packages and define paths

```
# TODO: library(ComplexHeatmap); library(circlize)
# TODO: subset_path <- file.path('..','data','dataset2_subset.csv')
#       long_path <- file.path('..','data','dataset2_subset_long.csv')
#       pdf_path <- file.path('..','pdf','dataset2_heatmap.pdf')
```

*Question:* Why do we still need both the wide and long versions?

## 2. Load/inspect the data

```
# TODO: read the CSVs into wide_df and long_df (stringsAsFactors = FALSE)
# TODO: print their dimensions and call head() on each
```

## 3. Choose a treatment group subset

Decide which `treatment_group` you want to display (e.g., Control vs Ciprofloxacin). Filter the columns of the wide matrix so only samples from that group remain. Hint: use the long table to map `mouse_id` + `day` to sample IDs like `1683-0` before subsetting the wide table.

```
# TODO: target_group <- 'Control'
# TODO: build sample_meta <- unique(long_df[, c('mouse_id','day','treatment_group')])
# TODO: sample_meta$sample_id <- paste(...)
# TODO: keep_samples <- sample_meta$sample_id[sample_meta$treatment_group == target_group]
# TODO: subset the wide_df columns with keep_samples
```

## 4. Quick summaries

- Count SNPs per genome using `table(wide_df$Genome)`.
- Build `with(long_df, table(mouse_id, day))` for coverage.
- Use `tapply(long_df$value, long_df$Genome, ...)` to show min/median/max.

```
# TODO: add commands described above
```

## 5. Build the heatmap matrix

```
# TODO: sample_cols <- setdiff(names(wide_df), c('Genome','snp_id','Position'))
# TODO: heatmap_matrix <- as.matrix(wide_df[, sample_cols]); mode(heatmap_matrix) <- 'numeric'
# TODO: rownames <- paste(wide_df$Genome, wide_df$snp_id, sep = ' | ')
# TODO: create sample_meta with mouse_id/day parsed from column names
# TODO: order columns by mouse/day and reorder heatmap_matrix accordingly
```

*Hint:* reuse the ordering logic from the guided notebook (`order()` on `mouse_id`, `day`, `sample_id`).

## 6. Colors and annotations

```
# TODO: compute min/mid/max for the matrix (na.rm = TRUE)
# TODO: color_fun <- circlize::colorRamp2(...)
# TODO: build mouse/day annotation via HeatmapAnnotation()
```

Challenge yourself to switch the palette (e.g., use `RColorBrewer::brewer.pal`).

## 7. Draw and export the heatmap

```
# TODO: construct Heatmap(...) object with top_annotation, column_split, etc.
# TODO: draw() it in the notebook
# TODO: save to pdf_path (dir.create + pdf + draw + dev.off())
```

## 8. Reflection prompts

1. Do you notice any new patterns when Turicimonas is included?
2. How would you highlight just the Turicimonas rows (hint: `row_split`)?
3. What additional annotation (e.g., day as a gradient) could help the reader?

Document your answers below.

```
# TODO: jot down observations or extra code experiments
```

Once you have a working script, compare with the solution notebook and proceed back to the main workflow.