Performing many simulations on the cluster in a smart way

Thijs Janzen

The problem

```
speciation_rate <- 1
extinction_rate <- 0.1
num_species <- do_complicated_simulation(speciation_rate,
                                         extinction_rate)

cat(speciation_rate, extinction_rate, num_species,
    file = "output.txt", append = TRUE)
```

issues:

- multiple jobs writing to same file

- program in bash to explore parameter values

The solution (one of many)

```
#!/bin/bash

#### SLURM settings ####
#SBATCH --partition=gelifes
#SBATCH --array=1-100

module load R

sim_num=$SLURM_ARRAY_TASK_ID # array number

Rscript --vanilla perform_analysis.R ${sim_num}
```

what is in perform_analysis.R?

```r
args <- commandArgs(trailingOnly = TRUE)
sim_number <- as.numeric(args[[1]])

param_grid <- expand.grid(spec_rate = seq(0, 1, by = 0.01),
                          ext_rate  = seq(0, 1, by = 0.01))

speciation_rate <- param_grid$spec_rate[sim_number]
extinction_rate <- param_grid$ext_rate[sim_number]

num_species <- do_complicated_simulation(speciation_rate,
                                         extinction_rate)

file_name <- paste0("results_", sim_number, ".txt")

cat(speciation_rate, extinction_rate, num_species,
    file = file_name, append = TRUE)
```

Collecting results

```
f <- list.files(pattern = "results_*")
found_results <- c()
for (x in f) {
  temp_results <- read.table(x)
  found_results <- rbind(found_results, temp_results)
}

colnames(found_results) <- c("speciation_rate",
                             "extinction_rate",
                             "number_of_species")
found_results <- tibble::as_tibble(found_results)
ggplot(found_results, aes(x = speciation_rate, y = number_of_species)) +
  geom_point()
```

Take-home message

* Use array style job submission
* Pass array number to R via command line argument
* Accept argument with commandArgs
* Organize parameter values via expand.grid
* Collect results with list.files