Simulate pool of pedigrees for simulation study

2023-05-11

Overview

- **NOTE** This .Rmd document is intended to be run on your PC and describes commands to simulate pedigrees in the R script `simrvped.R`.
- Simulating pedigrees is a compute-intensive task that should be performed on the Compute Canada cluster.
- Therefore, the first code chunk in this RMarkdown document sets `eval=FALSE` for the entire document, to avoid having the R commands run when you knit on your PC.
- At the end of the document there is a `purl()` command that you must manually execute on your PC (i.e., cut-and-paste into the R console) to generate the R script `simrvped.R`.
- After generating `simrvped.R` on your PC, port it to the Compute Canada cluster along with the SLURM script `simrvped.sh`.
- Then submit the SLURM script to the cluster with the command `sbatch simrvped.sh`.

Task-specific workflow

- We first load the `SimRVPedigree` package and a dataset of subtype-specific hazard rates for Hodgkin's and non-Hodgkin's lymphoma.

```
library(SimRVPedigree)
data(SubtypeHazards)
head(SubtypeHazards)

my_hazards <- hazard(SubtypeHazards,
                     subtype_ID = c("HL", "NHL"))
```

- We will run an "array job" on the cluster, taking guidance from Nirodha's SLURM scripts.
- The R script `simRVped.R` will be called by the SLURM script a specified number of times (e.g. 150), and each call will have its own "job ID" (e.g. 1:150).
- Each call of the R script can access its job ID through a Unix environment variable called `SLURM_ARRAY_TASK_ID`.
- The relevant code in the R script is:

```
dID = Sys.getenv("SLURM_ARRAY_TASK_ID")
seed = as.numeric(dID)
# Set a seed value to assure the reproducibility.
if(!is.na(seed)) {
  set.seed(seed)
} else {
  warning("No task ID, setting seed to 1")
  set.seed(1)
}
```

- Nirodha wrote an R function `generatePeds()` to call the `sim_RVped()` function and write the output to a file whose name includes the job ID.

- The arguments to `sim_RVped()` are mostly as in Christina's thesis, Appendix C.6, page 91, with the exception of the carrier probability. For our simulations of chromosome 8 the cumulative probability of the causal rare variants (cRVs) is 0.00016, and the `carrier_prob` should be twice this, or about 0.00032.
- The simulation program returns a full pedigree including all family members and also an ascertained version with only those family members recalled by the proband. We keep only the ascertained pedigree, writing it to a file in the
  `Outputfiles` directory. The ascertained-pedigree files for job ID `i` are called `full_pedi.txt`.

```
generatePeds = function(dataID){
  # Simulate pedigree ascertained for at least two individuals
  # affected by either Hodgkin's lymphoma or non-Hodgkin's lymphoma.
  out <- sim_RVped(hazard_rates = my_hazards,
                       GRR = c(35, 1),
                       RVfounder = TRUE,
                       FamID = 1,
                       founder_byears = c(1825, 1850),
                       ascertain_span = c(2000, 2010),
                       num_affected = 4,
                       stop_year = 2018,
                       carrier_prob = 0.00032, # 2x cum prob of cRVs
                       recall_probs = c(1, 1, 1, .75,.125,.125, 0),
                       first_diagnosis = 1980,
                       sub_criteria = list("HL",1)) # ascertain only if at least one HL

   write.table(out$ascertained_ped, file = paste0("Outputfiles/ascertained_ped",dataID,".txt"))
}
# Run the function.
generatePeds(dID)
```

R script for the cluster.

- To create the R script `simrvped.R` for the Compute Canada cluster, cut-and-paste the following into the R command line on your PC:

```
knitr::purl(input="simrvped.Rmd",output="simrvped.R")
```

- This command will return a file `simrvped.R` that includes every code chunk in this file `simrvped.Rmd`, including a code chunk at the very top of the file that sets `knitr` options.
- Delete the code chunk at the top of `simrvped.R` that sets the `knitr` options and the code chunk at the bottom that contains the `knitr::purl()` command.
- Then port `simrvped.Rmd` over to the Compute Canada cluster.