Simulation study overview

2023-05-14

- The simulation study is implemented in several RMarkdown (Rmd) and R scripts.
- A flowchart describing the relationships between the Rmd/R scripts and data files they read and write is shown below in Figure 1
- Briefly, above the horizontal line, the top-left and top-right of the flowchart depict independent workflows to get (a) a pool of pedigrees from which to sample studies (top-left) and (b) exome sequences and cRVs for the population (top-right). The data files generated by these two workflows feed into scripts depicted below the horizontal line to (1) simulate studies under the alternative hypothesis (`simalt.R` on left), for estimating the power and ranking ability of the different statistical methods and (2) simulate studies under the null hypothesis (`simnull.R` on right), for estimating the type-1 error rate of the different methods.
- Some .R and .Rmd scripts must be run on the Compute Canada cluster, some must be run on your PC and some can be run on either computer:
  - `simrvped.R`, `simalt.R` and `simnull.R` must be run on the Compute Canada cluster.
  - `simoverview.Rmd` (this script), `simaltSummary.Rmd` and `simnullSummary.Rmd` must be run (i.e. "knitted") on your PC. The summary scripts require that the outputs of `simalt.R` (`pvalresi.csv` and `rankresi.csv` for `i`=1,...,200), and `simnull.R` (`pvalnullresi.csv` for `i`=1,...,200) be copied from the Compute Canada cluster to your PC.
  - `checkpeds.R` and `getseqscrvs.R` can be run on either the Compute Canada cluster or your PC. If you decide to run `checkpeds.R` on your PC, you will need to port the output files of `simrvped.R`, `ascertained_pedi.txt` for `i`=1,...,150, from the cluster to your PC. If you decide to run `getseqscrvs.R` on your PC, you will need the file `Chromwide.Rdata` (see: https://zenodo.org/record/6499208#.ZGJenezMJJU) to be downloaded to your PC.
  - Any R script that *can* be run on the Compute Canada cluster has an associated SLURM script to run it there.
- Further details on the purpose of each script are given in the following sections.

1. Prepare for simulations

- Simulations involve conditional gene-dropping of sequences down pedigrees ascertained according to study criteria. To prepare for the simulations we need (a) a pool of pedigrees that meet the ascertainment criteria and (b) a population of exome sequences to sample for pedigree founders.

a. Get a pool of 55 pedigrees

- We first get a pool of 55 pedigrees with multiple disease-affected relatives.
- The pedigrees are ascertained according to criteria given in Section 4.4.2 of Christina's thesis (pages 47 and 48). Briefly, pedigrees must have 4-6 affected members, at least one affected member with HL that carries a causal rare variant (cRV), and at least one affected member with NHL. The limit of 6 affected members is imposed to keep computations manageable and most simulated pedigrees conform (a total of 141 out of 150 pedigrees).
- Some of these criteria can be enforced by the call to Christina's simulation function `sim_RVped()`, while others need to be checked afterwards:
  (i) Simulate 150 pedigrees with the `sim_RVped()` to have at least 4 affected members and have at least one HL among the affected members.

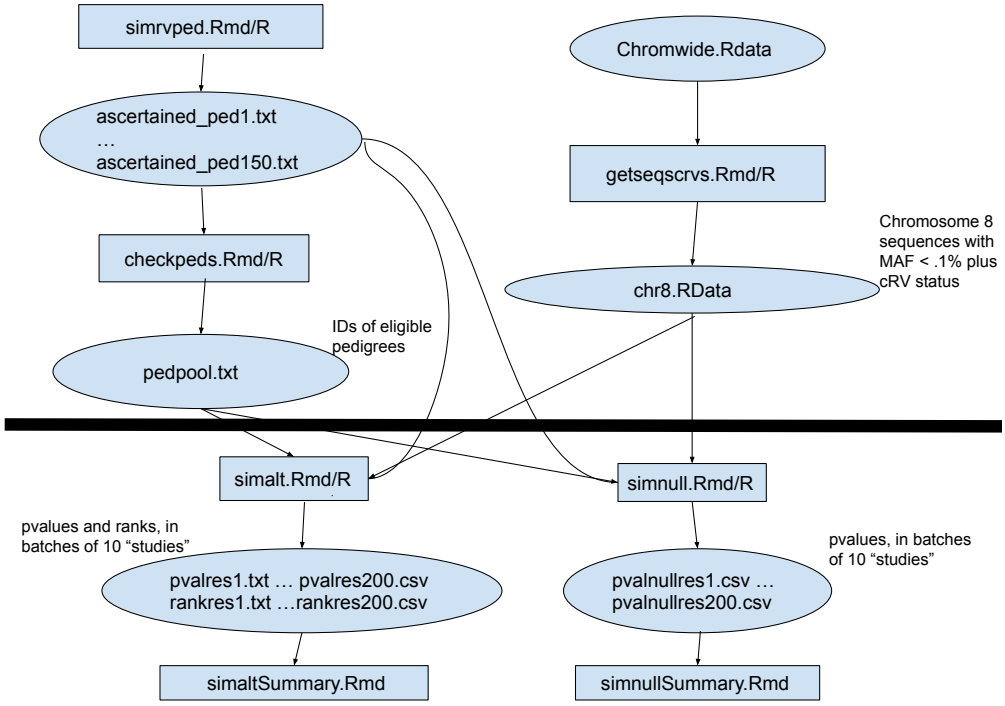

Figure 1: Flowchart of R and RMarkdown scripts and data files. Scripts are enclosed in rectangles. Data files are enclosed in ovals.

(ii) Keep only those pedigrees that have no more than 6 affected members, at least one HL case who carries a cRV and at least on NHL case. (This resulted in 101 pedigrees.)

- The first 55 pedigrees that meet all ascertainment criteria are used as our pool of pedigrees for the simulation study.
- Computational note: Simulating the 150 pedigrees with `sim_RVped()` in step (i) is *very* computationally-demanding and impractical to run off-cluster. By contrast, checking the validity of the pedigrees simulated by `sim_RVped()` is not too computationally-demanding and can be done either on your PC or the cluster.
- The scripts to simulate pedigrees are as follows.
  (i) The slurm script `simrvped.sh` calls the R script `simrvped.R` to simulate 150 pedigrees with `sim_RVped()`. The R commands in `simrvped.R` are documented in the RMarkdown document `simrvped.Rmd`. The R script generates 150 plain-text files `ascertained_ped1.txt` … `ascertained_ped150.txt` that each contain a tabular description of a simulated pedigree.
  (ii) The slurm script `checkpeds.sh` calls the R script `checkpeds.R` to select pedigrees with no more than 6 affected members, at least one HL case who carries a cRV and at least one NHL case. The R commands in `checkpeds.R` are documented in the RMarkdown document `checkpeds.Rmd`. The script reads the pedigrees output by `simrvped.R` and returns a plain-text file `pedpool.txt` that contains the IDs of the 55 eligible pedigrees in our pool.
  – See Figure 1 for a graphical view of the relationship between the .Rmd/R scripts and data files.

b. Get exome sequences and cRVs for the population

- To get the population of 107752*2=215504 exome sequences to sample for the pedigree founders, we only need Nirodha's R data file `Chromwide.Rdata` in the Zenodo repo (https://zenodo.org/record/6369360). `Chromwide.Rdata`'s data object, `out`, contains elements for each chromosome (except sex chromosomes) and provides information on the whole-exome SNVs with population minor allele frequency <0.01

(i.e. derived allele frequency <0.01 or >.99). Christina selects chromosome 8 and filters the SNVs to have minor allele frequency <0.001 (i.e. derived allele frequency <0.001 or >.999).

- The conditional gene drop through the pedigrees requires information on the population cRVs to be bundled with the sequence data object.
  - We sample 10 cRVs with population minor-allele count ≤ 10 from the TNFRSR10B and TNFRSF11B genes in the apoptosis pathway of Christina's thesis. TNFRSR10B and TNFRSF11B are the two causal genes that Christina uses in the simulation study of her thesis.
  - The cumulative probability of our selected cRVs is 0.00016.
- Computational note: Sampling cRVs and filtering variants in the chromosome 8 sequences is not too computationally-demanding and can be done on either your PC or the cluster.
- The slurm script `getseqscrvs.sh` calls `getseqscrvs.R` to get the population of (filtered) exome sequences to sample pedigree founders, including the information on the population cRVs.
  - The R commands in `getseqscrvs.R` are documented in the RMarkdown document `getseqscrvs.Rmd`.
  - `getseqscrvs.R` reads Nirodha's exome-wide data from the file `Chromwide.Rdata` and returns the chromosome 8 sequences with MAF < .1% and cRV status in the file `chr8.RData`.
- See Figure 1 for a graphical view of the relationship between the .Rmd/R scripts and data files.

2. Simulations under the alternative hypothesis

- Separate scripts (a) perform the simulation study and (b) summarize it to estimate the power and ranking abilities of the methods.

a. Perform the simulation study

- We simulate 2000 studies of three pedigrees under the alternative hypothesis that a cRV is segregating in at least one family in the study.
- For each study we:
  a. sample 3 pedigrees,
  b. for each sampled pedigree, use the `sim_RVstudy()` function from the `SimRVSequences` package to (i) sample a cRV from the pool of cRVs in the population, (ii) sample founder sequences and (iii) do *conditional* gene drop (conditional on cRV status) of these sequences through the pedigree.
  c. call `cd_new()` to get the lookup tables of test stats and p-values for the 5 methods. `cd_new()` is documented in the RMarkdown file `cd_new.Rmd`.
  d. For each cRV in the study, find its global configuration and:
     (i) look up the p-values for this global configuration in the p-value lookup table and write them to the p-value output file
     (ii) look up the statistic values for this global configuration in the statistics lookup table, rank them against other configurations in the observed sequence data and write the rankings to the rankings output file.
- Computational note: The simulation is *very* computationally-demanding and is done on the cluster as an "array" job of 200 jobs, each of which involves a batch of 10 simulation replicates, for a total of 2000 reps. The p-values from batch `i` are written to a comma-separated-values (CSV) file `pvalresi.csv` and the statistic ranks from batch `i` are written to a CSV file `rankresi.csv`.
- The slurm script `simalt.sh` calls the R script `simalt.R` to perform the simulation study under the alternative hypothesis.
  - The R commands in `simalt.R` are documented in the RMarkdown document `simalt.Rmd`.
  - `simalt.R` reads pedigree files from the pool of 55 eligible pedigrees output by `simrvped.R` and `checkpeds.R`, and reads founder sequences and cRV status information output by `getseqscrvs.R`.
  - `simalt.R` writes files of p-values and statistic ranks for the five methods.

b. Summarize the simulation study

- We read in the simulation results from the output of `simalt.R` and prepare summaries that can be included in our paper.
- In particular, the p-value results are used to make four LaTeX tables of power estimates and their SEs that we can use to compare the power of the five methods. The ranking results are used to make a graphical summary that we can use to compare the ranking ability of the five methods.
- We also look at some exploratory summaries of the simulation results to give a feel for the data and look for possible errors in the simulation code.
- Computational note: The summary RMarkdown file `simaltSummary.Rmd` is knitted on your PC. You must first copy the `simalt.R` output files from the cluster to your PC.
- The "output" of `simaltSummary.Rmd` is the four LaTeX tables of power estimates that are embedded in the RMarkdown file. For publication, the graph is saved as an encapsulated postscript file named `rankres.eps`.

3. Simulations under null hypothesis

- Separate scripts (a) perform the simulation study and (b) summarize it to estimate the type-1 error rates of the methods.

a. Perform the simulation study

- We simulate 2000 studies of three pedigrees under the null hypothesis of no association between the RV and trait.
- The key difference between simulations under the null and alternative hypotheses is that under the null hypothesis we do *unconditional* gene drop.
- For each study we:
  a. Sample 3 pedigrees.
  b. For each sampled pedigree, use the `sim_RVstudy()` function from the `SimRVSequences` package to *unconditionally* gene drop the chromosome 8 founder sequences through the pedigree. We gene-drop repeatedly through the study pedigrees until one of the RVs from the pool of candidate cRVs is observed in the affected individuals.
     – Details on the rationale for the repeated gene dropping are given in Section 2.2 of the file `simnull.pdf`
     (rendered from the RMarkdown document `simnull.Rmd`).
  c. Call `cd_new()` to get the lookup tables of test statistics and p-values for the five methods. `cd_new()` is documented in the RMarkdown file `cd_new.Rmd`.
  d. For each polymorphic cRV in the study, find its global configuration, look up the p-values for this global configuration in the p-value lookup table and write them to the p-value output file.
- Computational note: The simulation is *very* computationally-demanding and is done on the cluster as an "array" job of 200 jobs, each of which involves a batch of 10 study replicates, for a total of 2000 studies. The p-values from batch `i` are written to a CSV file `pvalnullresi.csv`.
- The slurm script `simnull.sh` calls the R script `simnull.R` to perform the simulation study under the null hypothesis.
  – The R commands in `simnull.R` are documented in the RMarkdown document `simnull.Rmd`.
  – `simnull.R` reads pedigree files from the pool of 55 eligible pedigrees returned by `simrvped.R` and `checkpeds.R`, and reads founder sequences and cRV status information returned by `getseqscrvs.R`.
  – `simnull.R` writes files of p-values and statistic ranks for the five methods.

b. Summarize the simulation study

- We read in the results of the simulations from the output of `simnull.R` and prepare a summary that can be included in our manuscript.

- Specifically, the p-value results for each study are used to make a LaTeX table of estimated type-1 error rates and their SEs that we can use to verify that the five methods control the type-1 error rate at the nominal 5% level.
- Computational note: The summary RMarkdown file `simnullSummary.Rmd` should be knitted on your PC. Before knitting, you must first copy the `simnull.R` output files `pvalnullres`i`.csv` for i=`1:200` from the cluster to your PC.
- The "output" of `simnullSummary.Rmd` is the LaTeX table of estimated type-1 error rates that is embedded in the RMarkdown file.