Workflow

An overview of the scripts

There are 5 major steps in the workflow. The scripts start with a number and this number indicates which step they belong to. The necessary functions for running each script are located in a file with the same name of the script, but ends with the keyword "_Utillity_Functions".

At each script, there are more details and comments explaining how things are working. We here briefly go over the scripts of each step and explain what their general role is. There is a diagram at the end of this document which summarizes the steps which might be helpful.

Step1. Simulate population and sample data

At this step, we simulate the population and sample data. This includes simulation of individuals in the population, cSNVs, variants, genotype matrix, case and control sample individuals and more. The scripts for this step are as follows:

1. “1_SimulateData.R”
   In this script, we can simulate population and sample data. User can pass parameters including causal region coordinates, number of equifrequent cSNVs, number of indiviudals carrying one copy of cSNV and more. User can also sample from this population by defining the desired number of case and control individuals.

2. “1_SimulateData_Utillity_Functions.R”
   All the necessary functions for running “1_SimulateData.R” are located in this script.

3. “1_simulation_in_msprime.py”
   This is a Python script that is being called during the simulation. This script holds the necessary Python code for simulating the ancestory of samples using a combination of Wright-Fisher and CwR model (This is a hybrid simulation, for more details please see: https://msprime.readthedocs.io/en/stable/tutorial.html#hybrid-simulations). We run this script during the simulation in R to get the tree sequence object.

This step saves two different R data files: 1) sample_data.RData which is a list holding all the necessary information about the sample, 2) pop_data.RData which is a list holding the information about the simulated population. We need to pass these objects to the later scripts when we are doing the analysis.

Step 2. Reconstructing the partitions and calculating the distance matrices

At this step, we reconstruct the partitions and calculate the distance matrices. We load “sample_data.RData” from step1. The scripts for this step are as follows:

(i) "2_part_dist.R"
This script is responsible for reconstructing the partitions and calculating the distance matrices.

(ii) "2_part_dist_Utillity_Functions.R"
All the necessary functions for "2_part_dist.R" are located in this script.

The output of this step is two objects: 1) part.RData which is a list of all reconstructed partitions, 2) dists.RData which is a list of all calculated distance matrices.

Step 3. Non-IBD methods

All the scripts at this step are responsible for running the non-IBD methods including Fishers exact test and SKAT-O. The scripts for this step are as follows:

(i) "3_FishersExactTest.R"
This script loads "sample_data.RData" from step1 and runs the Fishers exact test.

(ii) "3_SKATO.R"
This script loads "sample_data.RData" from step1 and runs SKATO test.

(iii) "3_NonIBD_Utillity_Functions.R"
All the necessary functions for this step are located in "3_NonIBD_Utillity_Functions.R".

Step 4. IBD methods: dCorN and dCorIT

All scripts at this step are responsible for running two of the IBD based methods: dCorN and dCorIT. To run IBD based methods at this step, we load "sample_data.RData" and "dists.RData" generated at step 1 and step 2, respectively.

(i) "4_IBD_dCorN.R"
This script loads the distance matrices and calculate the Naive dCor profile (dCorN profile) across the genome.

(ii) "4_IBD_dCorIT.R"
This script loads the distance matrices and calculate the dCor profile based on the true cSNV carrier status of case sequences/haplotypes (dCorIT profile) across the genome.

(iii) "4_IBD_Utilitty_Functions.R"
All the necessary functions for this step are located at this.

Step 5. IBD methods: dCorIG

At this step, we calculate the GNN statistic for each sequence. We reclassify the case sequences into cSNV carrier and non-carrier using some quantile of GNN statistic in the control group. After reclassifying the sequences, we calculate the dCorIG profile across the region.

(i) "5_IBD_dCorIG.R"
This script loads the sample data, reconstructed partitions, and distance matrices to calculate GNN and dCorIG profile.

(ii) "5_IBD_Utilitty_Functions.R"
All the neccessary functions for this step are located in this script.

Diagram of the workflow

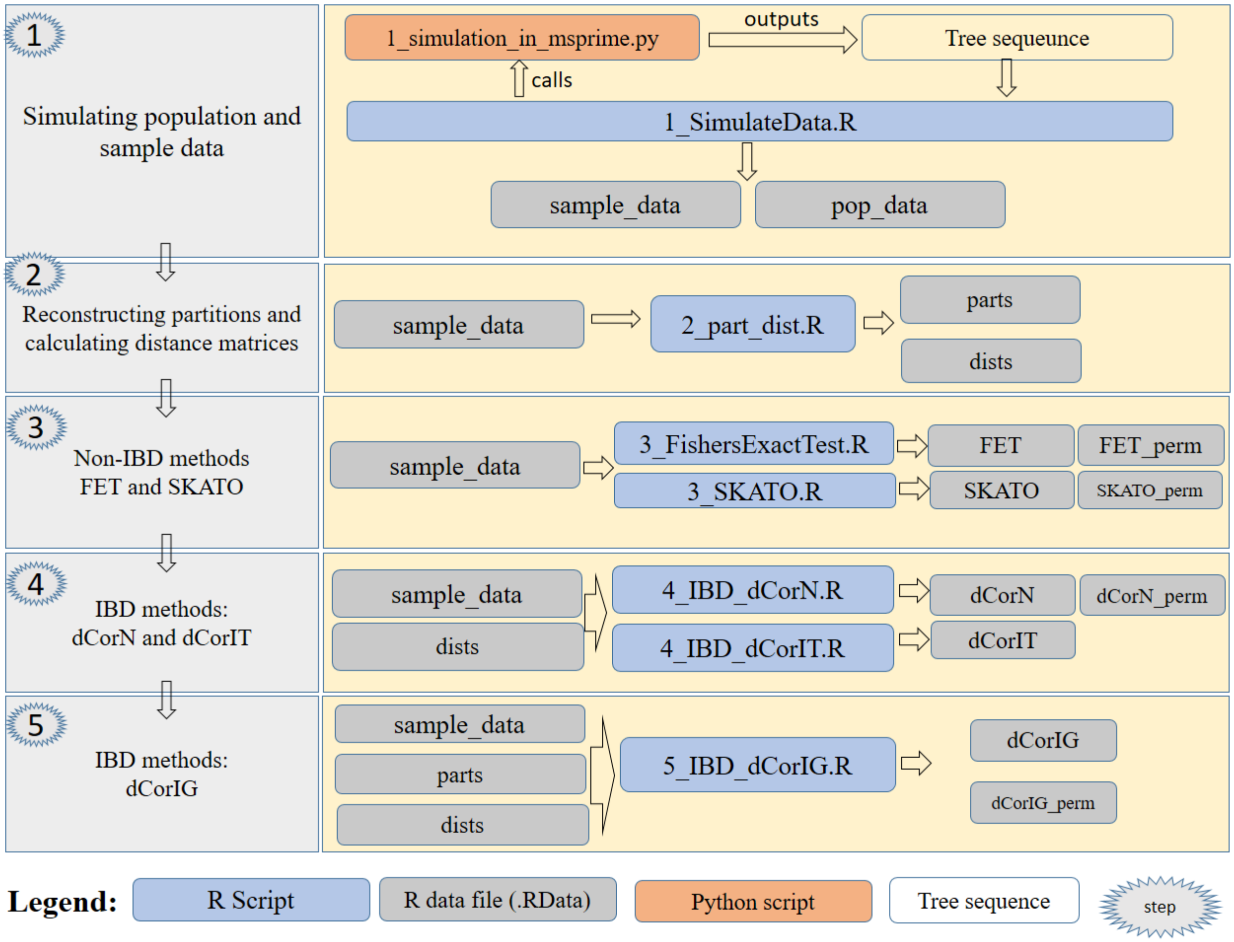

Figure 1: Diagram of the workflow