Project Description

BST-234

April 9, 2018

1 Data

In a whole-genome sequencing study, 20000 study subjects have been sequenced. First 10000 subjects are cases and the remaining are controls. Data provided shows the marker status for 50 loci. marker status is binary with 1 indicating the existence of the marker in the subject.We want to test the association between these 50 loci and the affection status (case vs. control). No other covariates are provided.

You will use SKAT(Sequence Kernel Association Test) to conduct this analysis and assess the significance using either asymptotic theory or permutations (1000 replicates).

2 Deliverables

- Software implementation of a numerically efficient and stable algorithm to do the analysis (in Python). Consider parallelization as well.
  Name your implementation as `[Team_number]_BST234_FinalProject.py`

- 10 minutes presentation (7 minutes for presentation and 3 minutes Q & A)

- Report (max: 4 pages).
  Name your report as `[Team_number]_BST234_FinalProject.pdf`

3 Deadline

- Code for your analysis is due on Friday, April 27th 11:59 pm

- In-class presentation will be on May 2nd. Please upload the slides to Canvas by 9 am

- Report is due on May 2nd 9 am