Using gene genealogies to localize rare variants associated with complex traits in diploid populations

Charith B. Karunarathna

Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, Canada.

BIRS Workshop, August 5-10, 2018

SFU

Purpose of the study

- To compare the performance of selected association methods to localize trait-influencing causal variants within a 2-Mbp candidate genomic region.
- Our work extends that of Burkett et al. 2014, which investigated the ability to detect causal variant.
- First, we present a case study of one of 200 simulated datasets for insight into the methods.
- Then, using the 200 simulated datasets, we score which method localizes best, overall.

Data simulation

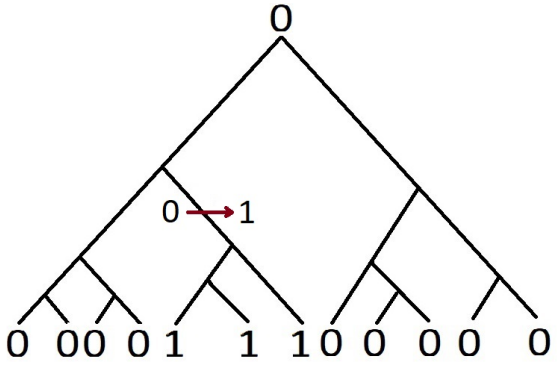
- Simulate data for 3000 haplotypes in a 2-Mbp genomic region using 'fastsimcoal2' (Excoffier et al. 2013).
  - Keep the ancestral trees connecting haplotypes.
- Randomly pair the 3000 haplotypes into 1500 diploid individuals.
- Assign disease status to the 1500 individuals based on randomly sampled risk SNVs (rSNVs) from the mid region (950kbp-1050kbp).
- Sample 50 diseased individuals (cases) and 50 non-diseased individuals (controls).

Association Methods

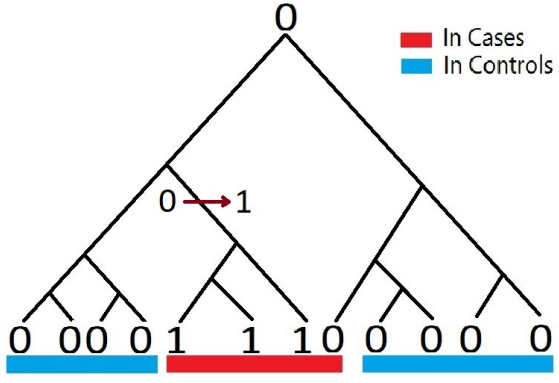
1. Single-variant method
   - Fisher's exact test
2. Pooled-variant methods
   - Variable Threshold test (VT) (Price et al. 2010)
   - C-alpha test (Neale et al. 2011)
3. Joint-modeling methods
   - CAVIARBF (Chen et al. 2015)
   - Elastic-Net (Zou & Hastie. 2005)
4. Tree-based methods
   - Blossoc (Mailund et al. 2006)
   - Naive Mantel (Burkett et al. 2014), and informed Mantel (Karunarathna & Graham. 2018)

Genealogical Trees

- Genealogical tree represents the ancestry of a genetic variant.



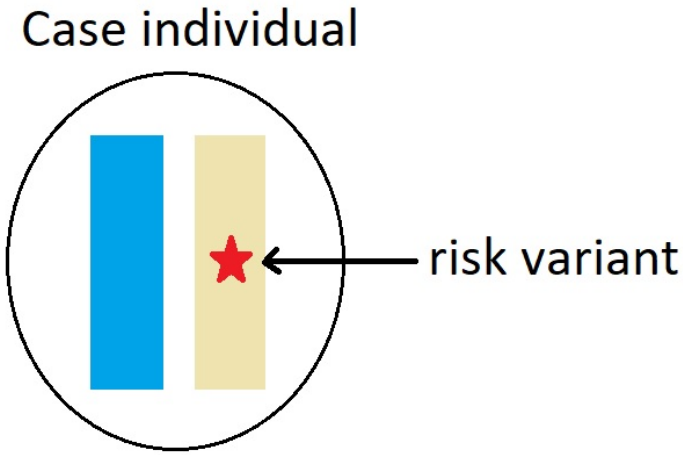
- Case alleles tend to cluster together on a tree for a variant that influences disease susceptibility.



- Tree-based methods evaluate the clustering of the disease status on the trees.

Tree-based methods

- Two methods to assess clustering of disease status in genealogical trees, Blossoc and Mantel test.
- Blossoc uses reconstructed unknown trees.
- Mantel test uses the known trees.
- Two versions of Mantel test as the bench mark for comparison: naive Mantel, and informed Mantel.



Single-variant method: Fisher's exact test

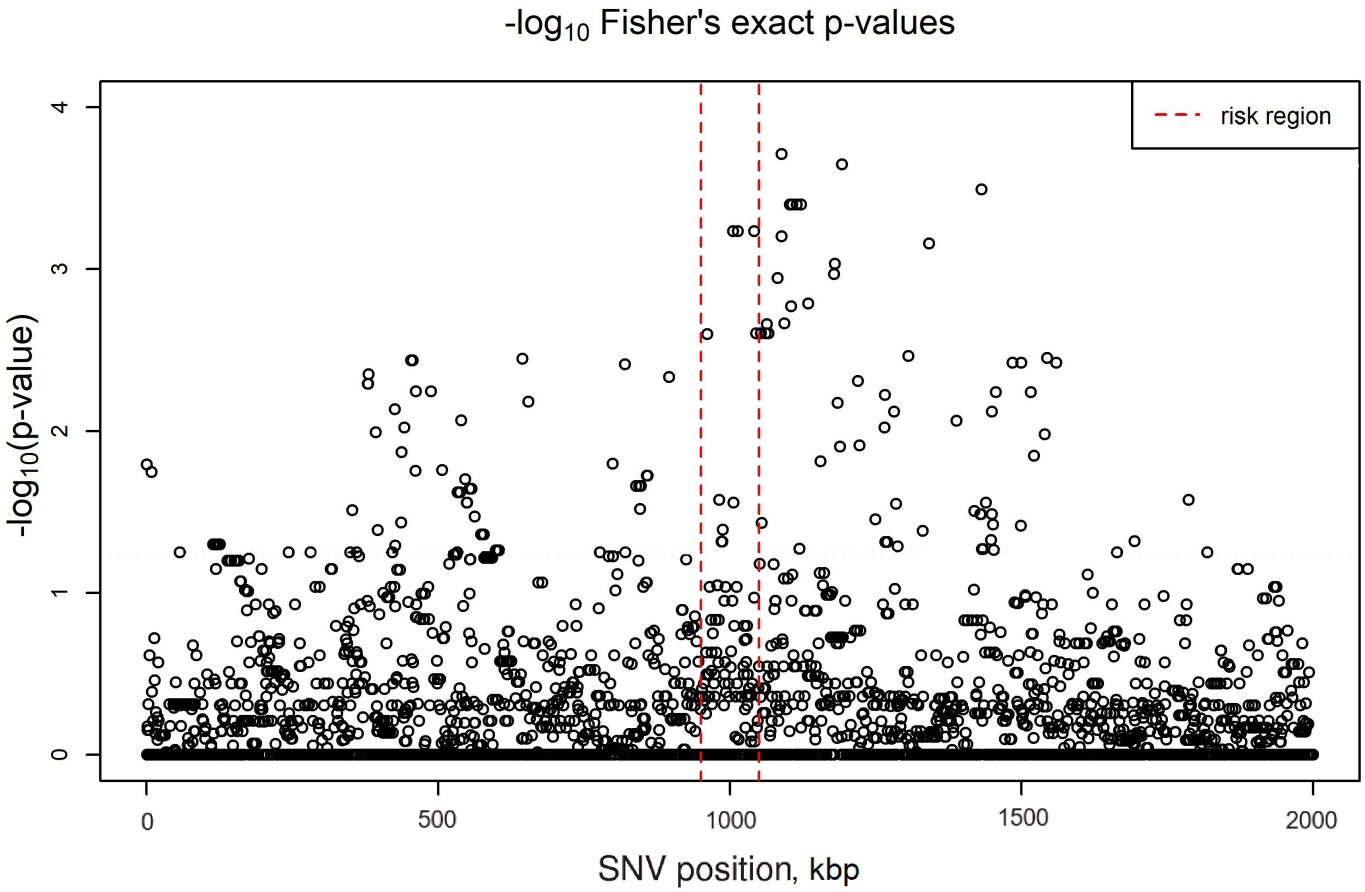


Figure 1: Manhattan plot for Fisher's exact test in the 100 case and 100 control haplotypes.

Pooled-variant methods

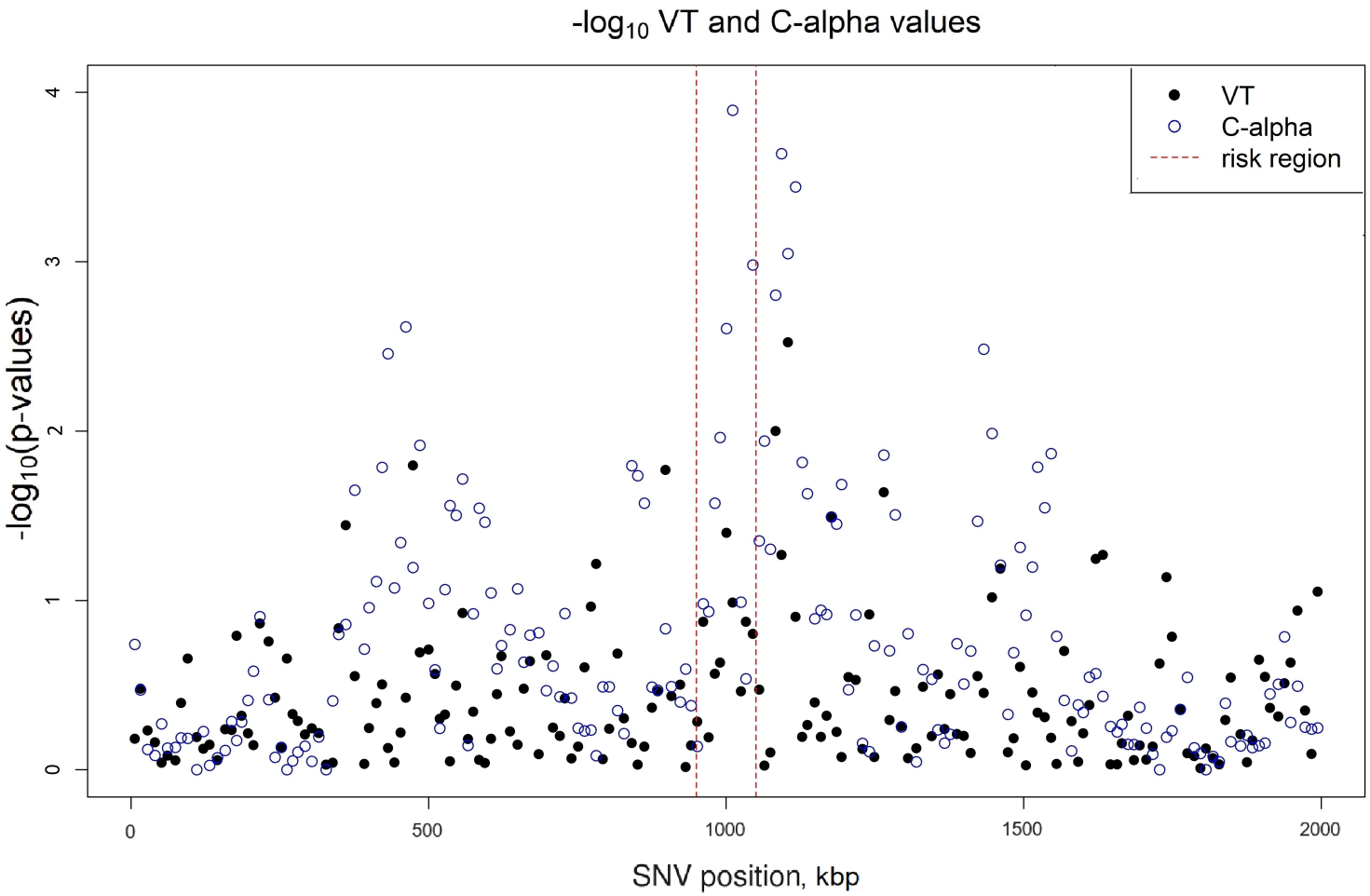


Figure 2: The negative $\log_{10}$ of p-values are shown on the vertical axes, obtained by applying VT and C-alpha tests across the simulated region using sliding windows of 20 SNVs overlapping by 5 SNVs.

Joint-modeling methods

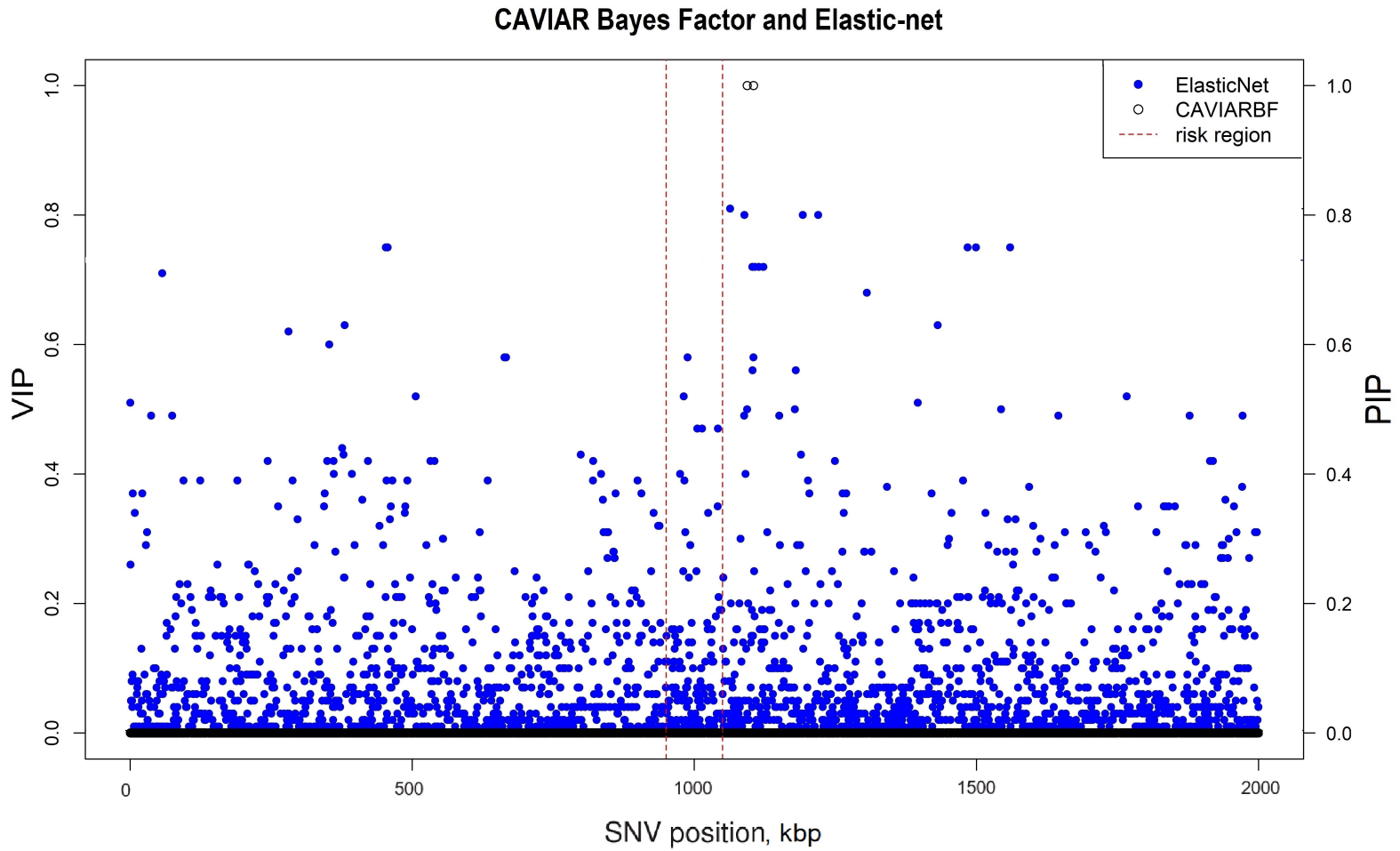


Figure 3: Variable-inclusion probabilities (VIPs) for SNVs computed from Elastic-net (left axis), and posterior inclusion probabilities (right axis), computed from CAVIARBF, across the simulated region.

Tree-based methods: Blossoc

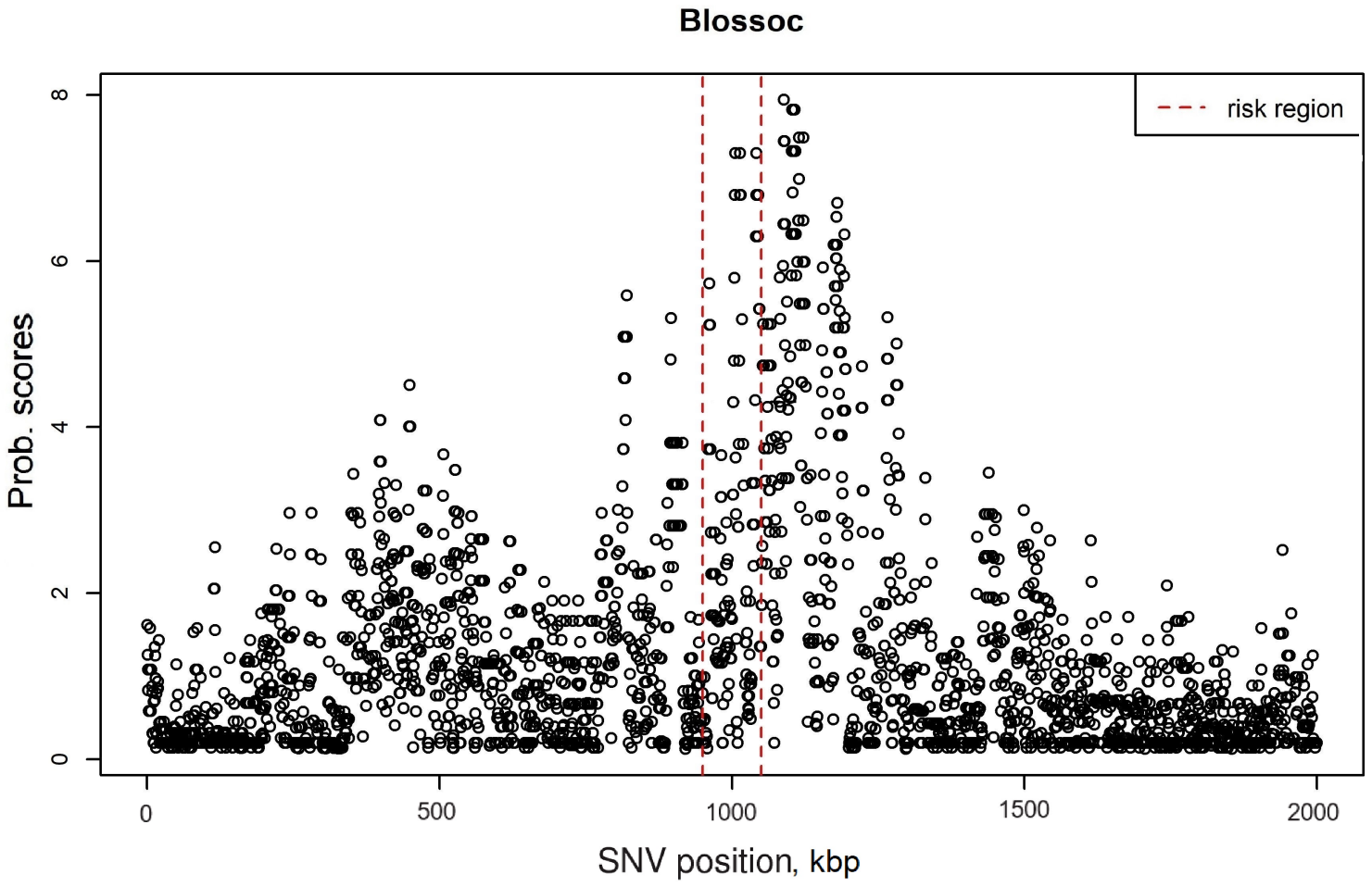


Figure 4: Plot showing the output from Blossoc.

Tree-based methods: Naive Mantel

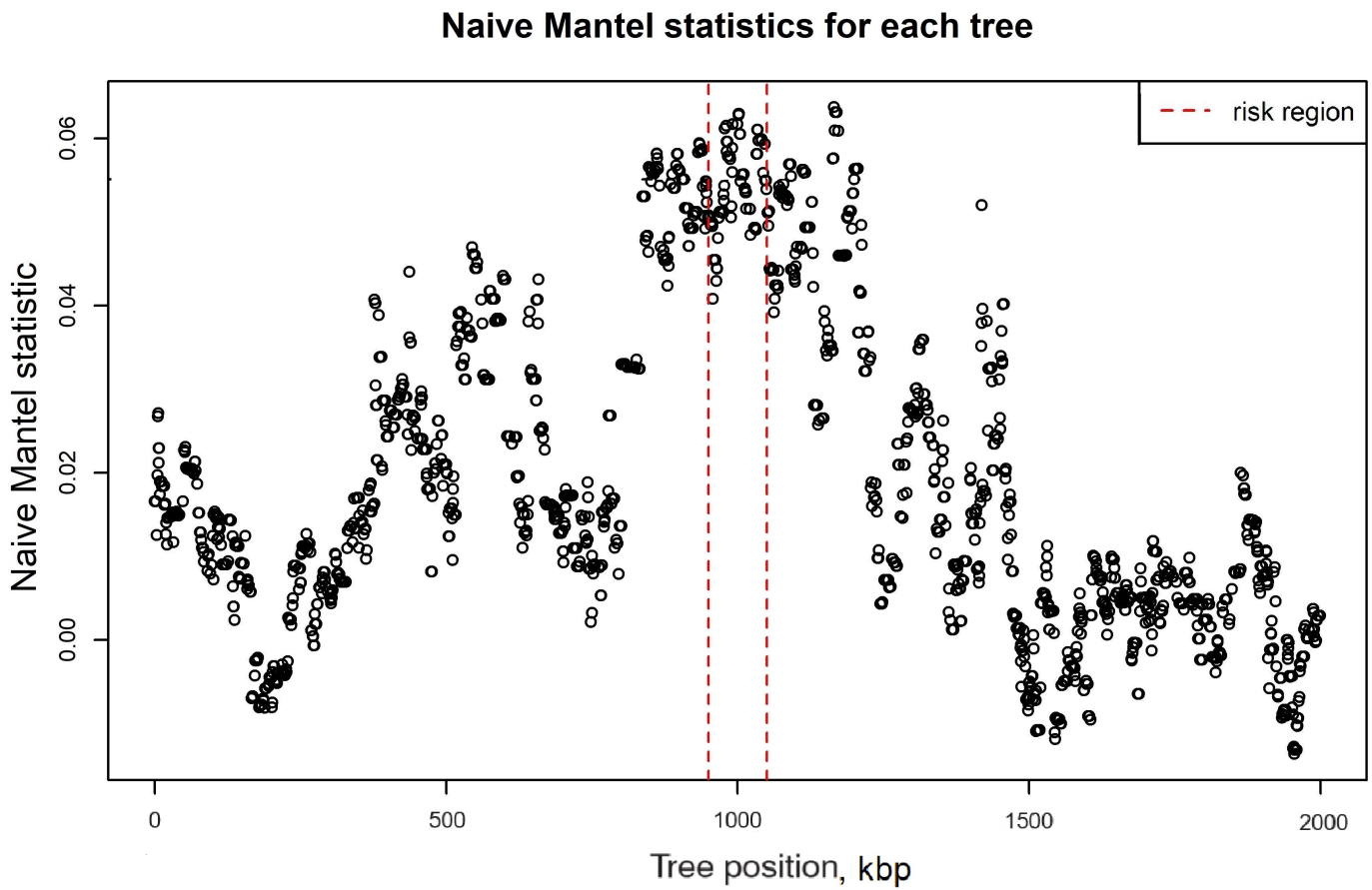


Figure 5: Naive-Mantel statistics for each tree position across the simulated region.

Tree-based methods: Informed Mantel

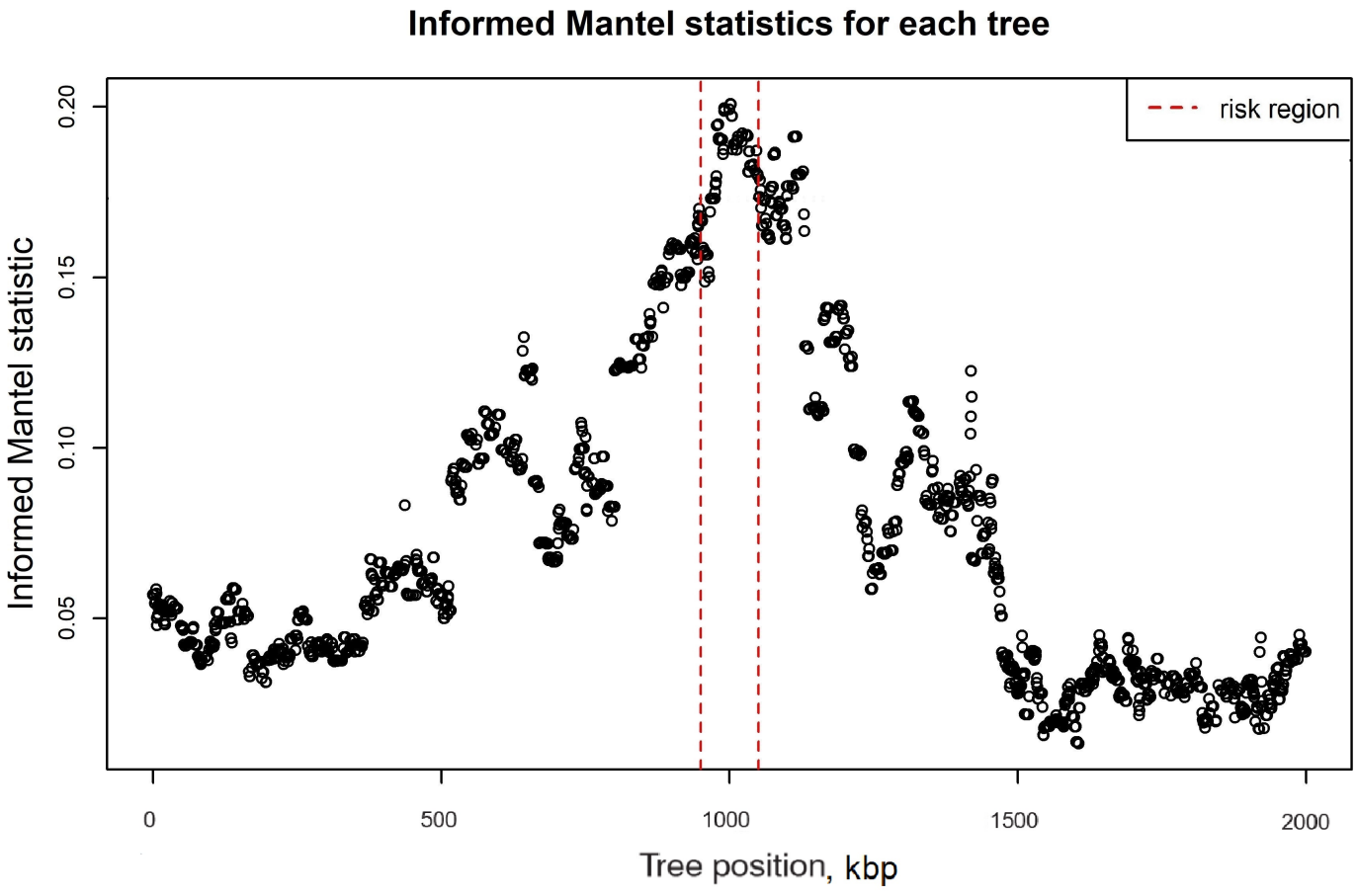


Figure 6: Informed-Mantel statistics for each tree position in the genomic region.

Signal Localization

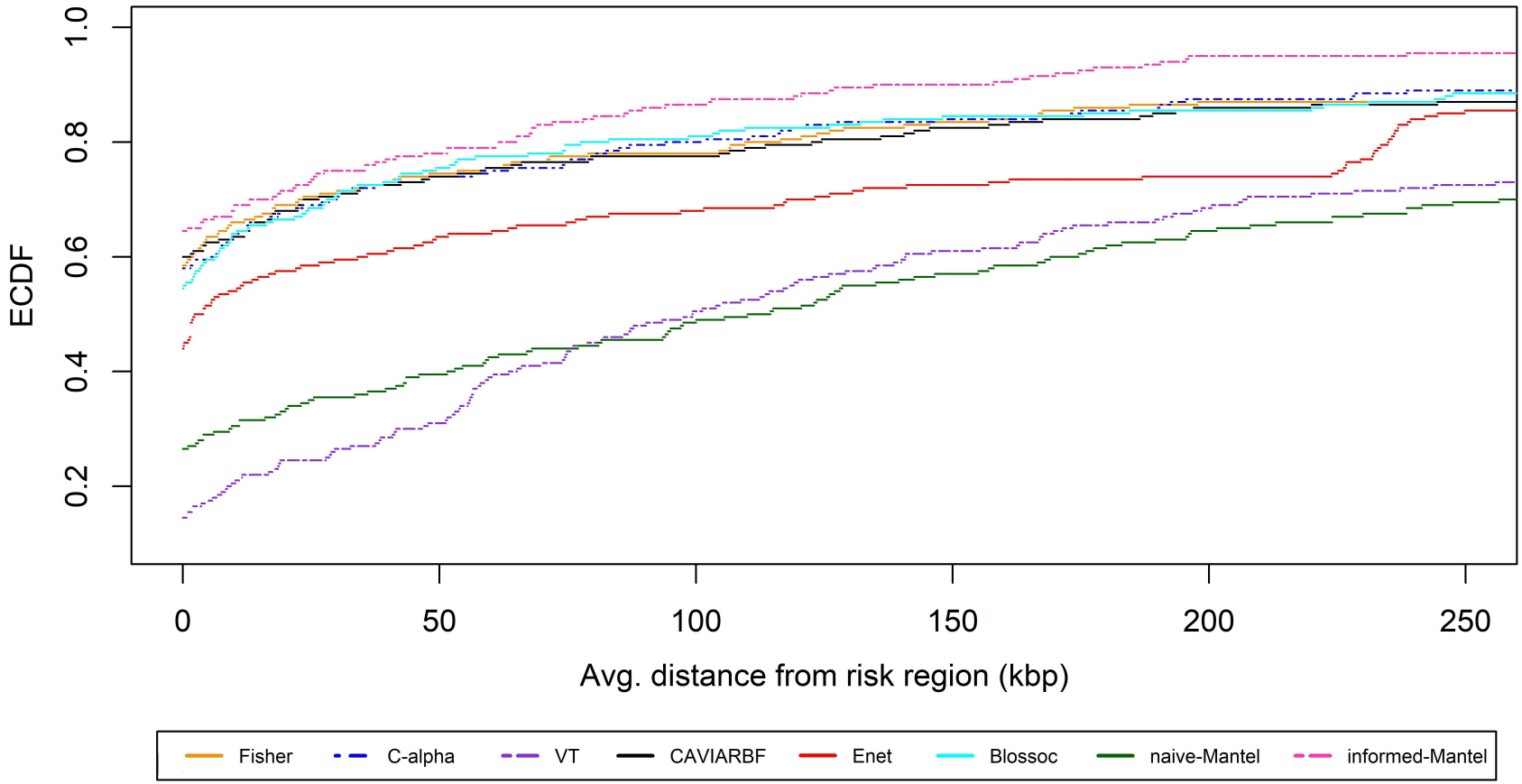


Figure 7: Empirical Cumulative Distribution Functions (ECDFs) of average distances of the peak association signals from the risk region for the 200 datasets.

Discussion

- Localization results on the example dataset.
  - C-alpha test and informed Mantel test are the only methods that successfully localize the association signal.
  - However, the peak signal from all the other methods is close to the disease risk region.
- Localization results from the simulation study.
  - Naive Mantel test performed very poorly relative to the other methods.
  - Not surprisingly, the informed Mantel test outperformed all the other methods.
  - Blossoc, CAVIARBF, C-alpha, and Fishers exact test performed better in localizing the signal.

Reference

Karunarathna, C. B., & Graham, J. (2018). Using gene genealogies to localize rare variants associated with complex traits in diploid populations. Human heredity, 83(1), 30-39.

Acknowledgements

- Thanks to my supervisor, Dr.Jinko Graham, and to Dr.Kelly Burkett for helpful discussions.

SFU