Variable selection using summary statistics

Matti Pirinen[1]

[1] *University of Helsinki, Finland, Matti.pirinen@helsinki.fi*

With increasing capabilities to measure a massive number of variables, efficient variable selection methods are needed to improve our understanding of the underlying data generating processes. This is evident, for example, in human genomics, where genomic regions showing association to a disease may contain thousands of highly correlated variants, while we expect that only a small number of them are truly involved in the disease process.

I discuss ideas that have made variable selection practical in human genomics and demonstrate them through our experiences with the FINEMAP algorithm.

(1) Compressing data to light-weight summaries to avoid logistics and privacy concerns related to complete data sharing and to minimize the computational overhead.

(2) Efficient implementation of sparsity assumptions.

(3) Efficient search algorithms.

(4) Use of public reference databases to complement the available summary statistics.