Appendix: Technical details of the DEPTH procedure

The version of the DEPTH algorithm used in this paper is conceptually similar to a previously published version (1), but utilises a different model fitting approach. Rather than using linear models to quantify association, this version uses non-parametric classification trees. These models naturally incorporate marginal and joint interaction effects, as well as the potential for interactions between terms, which automatically takes into account linkage disequilibrium/correlation between SNPs.

Further, this version uses the minimum message length principle to quantify evidence of association (2). The minimum message length principle is a powerful, general purpose approach to inductive inference based on information theory that is particularly appropriate for complex, non-parametric machine learning models such as decision trees and mixtures models.

The DEPTH algorithm used in this paper is as follows:

1) Compute a goodness-of-fit measure (the message length) for the phenotypes without reference to any SNPs; denote this quantity by I_0. This is done by forming a decision tree for the phenotypes using no predictors, and acts as a null (no association) model.
2) Divide the SNPs into n windows, $W_1, \ldots, W_n$. A window is formed for each SNP in our dataset, and includes all SNPs within a pre-specified genetic distance (100 Kb) of the reference SNP.
3) For each window, compute an observed measure of association. This is done by training a decision tree on the phenotypes and the SNPs within the window. The decision tree is estimated by finding the tree structure that minimises the message length (3); denote the message length for the best tree for window W_j by I_j.

4) The empirical “null” distribution of the message length for window W_j is then found by permuting the phenotype labels m times, and training m trees on these permutated phenotype vectors. The resulting message lengths can be used to approximate the distribution of the message length under the no-assocation model for window W_j.

Once the algorithm was run, there are n decision trees, one for each window, along with their associated message lengths. Using the well-known relationships between minimum message length and Bayesian inference (2), the posterior-odds (PO_j) for each window j in favour of the association (tree) model can be computed using

$$\mathrm{PO}_j = \exp(\delta_j),$$

where $\delta_j = I_0 - I_j$ is the difference in message lengths between the null model and the fitted decision tree model for window j, a positive value of $\delta_j > 0$ indicating a preference for the association model over the no-association model in that particular window.

References

1. Makalic E, Schmidt DF, Hopper JL. DEPTH: A Novel Algorithm for Feature Ranking with Application to Genome-Wide Association Studies. In: Cranefield S, Abhaya N, editors. AI 2013: Advances in Artificial Intelligence. Cham, Switzerland: Springer International Publishing; 2013. p. 80-5.

2. Wallace, C. S. Statistical and Inductive Inference by Minimum Message Length Springer, 2005

3. Wallace, C. S. & Patrick, J. D. Coding Decision Trees. Machine Learning. 1993; 11: 7-22