Abstract

Identifying genetic drivers associated with plasma proteome abundance is of critical importance to understand how DNA shapes disease mechanisms. Recent studies have used improved prediction performance of deep learning over linear regression, to identify proteins whose abundance is underlain by DNA regions with complex interactions and non-linear e ects. Our work shows that these performance improvements – and, therefore, previous inference on the complexity of genetic e ects – are significantly altered when scaling linear regression to more correlated data and, crucially, when employing a non-linear, one-hot encoding. More precisely, we propose an Approximate Message Passing (AMP) approach in two stages: the first identifies regions associated with protein abundance on a smaller set of DNA variants; the second generates a predictor within selected regions from the one-hot encoding of a set of 8 million variants. Our linear regression approach performs on par with deep learning, and a further improvement is achieved by applying an o -the-shelf non-linear method (XGBoost) to DNA variants selected in the second stage. Our findings demonstrate that scalable linear regression methods are highly e ective at identifying associated regions within high-dimensional DNA data. One-hot encoding can then more appropriately represent interactions and e ects of these regions as compared to smaller scale analyses.