The Future of Genetic Prediction [Preliminary Draft]

Patrick Turley, Dan Benjamin, David Cesarini, Christopher Chabris, David Laibson

May 30, 2014

Introduction

The predictive power of polygenic scores still falls far below their potential as estimated in heritability research (Wray et al 2013). This has largely been due to relatively small sample sizes and due to simple methods that don't allow researchers to use correlated SNPs in prediction (Chatterjee et al 2013, Daetwyler et al 2008, Daetwyler et al 2010, Dudbridge 2013, Wray et al 2013). Advances in genetics research and computational power have made more advanced methods of calculating scores increasingly feasible, and the improvement of these methods over traditional methods has been explored through simulations (de los Campos et al 2013, Goddard 2008, Goddard et al 2009, Meuwissen et al 2001, Vattikuti et al 2014, Yang et al 2012). Analytic forms for the predictive accuracy in various settings would be useful to researchers who are weighing the benefits of implementing more computationally intensive methods and who are considering the improvements to predictive power by increasing their discovery sample size.

We have derived an analytic form for the predictive accuracy when the researcher is using a set of SNPs with a flexible LD structure and when score weights are estimated in a Bayesian setting with a normally- distributed prior. This formula is flexible enough that it could be also applied to a setting with gene-by-gene or gene-by-environment interactions. In a simple example, we show that using Bayesian weights can lead to strong improvement in predictive power over traditional OLS-based weights and that there is reason to believe that as sample sizes continue to grow, we are entering an era when the predictive power of polygenic scores will rise drastically.