bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Title: A machine-learning heuristic to improve gene score prediction of polygenic traits

Short title: Machine-learning boosted gene scores

Guillaume Pare[1,2,3*], Shihong Mao[1], Wei Q. Deng[4]

1 Population Health Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Canada, 2 Population Genomics Program, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada, 3 Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada, 4 Department of Statistical Sciences, University of Toronto, Toronto, Canada

*Corresponding author: pareg@mcmaster.ca

Abstract

The advent of precision medicine is largely dependent on the availability of accurate and highly predictive gene scores. While progress has been made identifying genetic determinants of polygenic traits, the phenotypic variance explained by gene scores derived from genome-wide associations remains modest. Machine-learning techniques have proven very useful for solving a broad range of prediction problems, yet are not widely applied to complex traits prediction using gene scores. We propose a novel machine-learning heuristic (MLH) to improve the predictive performance of gene scores. It is based on two innovative features. We first apply gradient boosted regression trees

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

models to leverage a large number of SNPs and optimize the weights of individual SNPs included in the gene scores. We show a calibration set sample size of ~200 individuals is sufficient for optimal performance. We then correct for linkage disequilibrium (LD) between SNPs using a novel procedure, enabling retention of all SNPs in the gene score irrespective of LD. Our novel heuristic yielded a prediction R^2 of 0.237, 0.082 for height and BMI using GIANT summary association statistics in the UKBiobank study (N=130K; 1.98M SNPs), explaining 46.6% and 32.6% of the overall polygenic variance, respectively. Corresponding area under the ROC was 0.602 for diabetes in the UKBiobank using DIAGRAM association statistics. MLH outperformed other gene score heuristics for height and BMI and was equivalent to LDpred for diabetes. Results were independently validated in participants of the HRS (N=8,292) study. Our report demonstrates the potential of machine-learning methods for polygenic trait prediction. Our method has wide-ranging applications, from predicting medically important traits to creating stronger instrumental variables for Mendelian randomization studies.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Main Text

Despite moderate to high narrow-sense heritability estimates for most polygenic traits, known genetic associations only explain a relatively small proportion of polygenic traits variance. It has been proposed that weak, yet undetected, associations underlie polygenic trait heritability[1]. Consistent with this hypothesis, polygenic scores including both strongly and weakly associated variants produce vastly superior prediction R^2 than the ones including only genome-wide significant variants. The most popular heuristic is based on linkage disequilibrium (LD) pruning of SNPs, prioritizing the most significant associations up to an empirically determined *p*-value threshold and pruning the remaining SNPs based on LD[2]. This "pruning and thresholding" (P+T) approach has the advantage of being simple and computationally efficient, but discards some information because of LD pruning. To remediate this issue, a novel method was recently proposed that uses LD information from an external reference panel to infer the mean causal effect size using a Bayesian approach (LDpred)[3]. While the latter method has been shown to improve prediction R^2, we hypothesized that a further gain in prediction R^2 could be made by tuning the weights of SNPs included in the gene score using machine-learning algorithms.

Machine learning encompasses a class of methods widely used to solve complex prediction problems. It has proven particularly useful when prediction is dependent on the integration of a large number of predictor variables, including higher-order interactions, and when sizeable training datasets are available for model fitting. Our novel

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

heuristic leverages the large number of SNPs in genome-wide studies to calibrate the weights of SNPs contributing to the gene score. This is done by partitioning the genome into non-overlapping, complementary parts. Our method involves two steps (Figure 1) and uses the univariate regression coefficients from external meta-analysis summary association statistics as a starting point (see Appendix for detailed Methods). First, these external univariate regression coefficients are updated with respect to a target population by the boosted regression trees models. Second, the updated weights are corrected for LD to produce the final gene score.

Boosted regression trees are powerful and versatile methods for continuous outcome prediction[4] and thus ideal for updating the SNP weights in a gene score. Tree-based models partition the predictor space according to simple rules to identify regions having the most homogeneous responses to predictors and fitting the mean response for observations in that region. Boosting is used to efficiently combine a large number of relatively simple tree models adaptively, to optimize predictive performance. Our approach uses boosted regression trees to adjust summary association statistics regression coefficients in order to maximize the prediction R^2 in a target population. Regression coefficients from large meta-analyses are implicitly assumed to provide the best initial estimates and regression trees "tune" them based on the regression coefficients observed in the target population. To avoid over-fitting, SNPs are divided into five distinct contiguous sets of SNPs (thus circumventing potential LD spillover) and weights of SNPs in each set is calculated using the prediction models trained on the remaining four sets. For example, the first set comprises SNPs from chromosomes 1, 2 and part of 3 such that

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

SNPs from the remaining part of chromosome 3 and chromosomes 4 to 22 would be used to derive prediction models for SNPs in that first set. The observed regression coefficient of any single SNP in the target population is thus never used directly or indirectly to derive its own gene score weight.

It is advantageous to correct the derived weights for LD when including multiple SNPs in a gene score, unless SNPs are first LD pruned. The novel correction we propose is based on the sum of pairwise LD r^2 of each SNP over neighboring SNPs. Gene score weights of each SNP is divided by the corresponding sum of r^2. To illustrate with a simple example, if five SNPs are in perfect LD (r^2=1) with each other but in linkage equilibrium with all other SNPs (r^2=0), then the gene score weights of these five SNPs would be divided by five. As all five SNPs are included in the gene score and the effect of all five SNPs summed, the corrected weight contributions are equivalent to including a single SNP without correction. This also explains why it is necessary to apply the LD correction only after adjusting SNP weights with boosted regression trees as otherwise important information on strength of association of individual SNPs would be lost. LD is summed over SNPs included in the gene score, such that our correction is specific to the set of SNPs included in a given gene score. When the genetic effects are strictly additive (i.e. no haplotype or interaction effect), the resulting gene score provides an unbiased estimate of the underlying genetic variance although at a tradeoff of increased gene score variance as compared to the “true” unobserved genetic model (see Appendix).

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We applied our machine-learning heuristic (MLH) to the prediction of height on a calibration set of 10,000 participants and an independent validation set of 130,215 from the UKBiobank (UKB) using Genetic Investigation of Anthropometric Traits (GIANT) consortium summary association statistics[3, 5] on 1.98M SNPs (Figure 1). Since the UKB is not part of the GIANT consortium, the reference and target populations can be assumed to be independent. As recently proposed[6], principal components were added to the model and included in the prediction R^2. Prediction R^2 of the gene score derived using our heuristic was 0.237, corresponding to 46.6% of total polygenic genetic variance estimated in UKB using variance component models[7] (i.e. 0.509). This compared advantageously to the optimal prediction R^2 obtained with P+T (0.217; 177K SNPs), LDpred (0.202) or an unadjusted gene score (0.163) ($p<10^{-100}$ for all pairwise comparisons with MLH; Figures 2 and 3).

We also tested the performance of MLH for prediction of body mass index (BMI) and diabetes in the UKB. The resulting gene score for BMI had a prediction R^2 of 0.082, outperforming the prediction R^2 of unadjusted gene score (0.069), P+T (0.069) and LDpred (0.076) ($p<0.006$ for all pairwise comparisons with MLH; Figure 2). Our heuristic accounted for 32.6% of the total polygenic variance, which was estimated at 0.251 for BMI in the UKB. The MLH gene score for diabetes had area under the receiver operator curve (AUROC) of 0.602, which was not statistically different from LDpred (0.613; $p=0.06$ for comparison) and compared favorably to unadjusted gene score (0.583) and P+T (0.576) ($p<10^{-5}$ for comparisons with MLH; Figures 2 and 3).

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Calibration, or the ability of a gene score to accurately predict real observations, is as important as predictiveness when gene scores are used to infer unobserved traits. To evaluate calibration, we calculated the average absolute difference between predicted trait and actual trait for height and BMI in the validation set. For all methods, gene scores were first calibrated in the training set through use of a simple regression coefficient (along with principal components regression coefficients). The average absolute difference was smallest for MLH for both height (0.701 SD) and BMI (0.744 SD) as compared to other gene score methods ($p<10^{-32}$ for all pairwise comparisons with MLH). We tested for calibration of diabetes gene scores using the Hosmer-Lemeshow test, dividing the UKB validation set by deciles of predicted trait (Figure 4). There was no evidence of mismatch between predicted and observed event rates for any of the gene scores ($p>0.05$).

The set of participants used for calibration of MLH can theoretically also be the test set since the regression coefficient of each SNP in the target population is not used to tune its own gene score weight. However, doing so presents practical challenges in the situation where one wants to predict a trait unobserved in the target population, in which case a smaller training sample size is advantageous. We therefore explored the effect of the size of the calibration set on gene score performance by sub-sampling an increasing proportion of our calibration set for MLH tuning. We determined that a calibration set as small as 200 was adequate to provide high prediction R^2 for height and BMI (Figure 5). For diabetes, we selected an increasing number of case-control pairs. 100 pairs were sufficient for adequate performance.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

For any given SNP, the regression coefficient observed in UKB was not used to determine its own gene score weight. Nonetheless, regression coefficients of other SNPs in UKB were used, raising the issue of transferability to other populations. We therefore tested gene scores derived from the UKB in Health Retirement Study (HRS) participants of European descent (*N*=8,292). Only directly genotyped SNPs were used for this analysis and 683K SNPs overlapped with both the UKB and consortia associations. For each method, optimal gene score derived in the UKB calibration set was tested in HRS without any further fitting or adjustment. Consistent with UKBiobank results, our machine-learning heuristic outperformed other methods for height and BMI, and was close second to LDpred for diabetes (Figures 2 and 3).

Our proposed machine-learning heuristic led to significant improvements in prediction R^2 as compared to existing methods. Furthermore, we showed that MLH gene scores are well calibrated, requiring only a very small calibration set sample size (N~200) to achieve maximal performance. This latter characteristic makes our method advantageous for prediction of unobserved traits and stems from the fact our heuristic leverages the large number of genetic variants reported in genome-wide association studies (GWAS) to train boosted regression trees models through genome partitioning. Regression trees can capture nonlinear effects and higher-order interactions while the boosting algorithm combines individually weak predictors to produce a strong classifier that enables a better prediction of genetic effects.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A few limitations are worth mentioning. First, our method is based on the premise that SNPs contribute additively to genetic variance. While empirical evidence suggests this holds true in most cases, our method is not expected to perform well in genomic regions where strong genetic interactions are present (e.g. HLA), and alternative methods such as LDpred might be better suited[3]. Second, there is a possibility that gene scores derived using our method are inherently population-specific. However, with the exception of unadjusted gene scores, all methods require the determination of parameters in the target population and ours is no different. Furthermore, if the genetic architecture varies between populations, then no gene score will perform universally well and it will be beneficial to tailor gene scores to each population. The observation that our heuristic performed as well in HRS as compared to other methods suggests this might not be the case. Third, our correction for LD yielded advantageous results yet is expected to lead to some loss of information when truly associated SNPs are in partial LD (see Appendix). Nonetheless, our correction for LD also has several benefits such as simplicity, use of summary association statistics and intrinsic robustness to minor misspecification of LD or association strength.

In summary, we propose a novel heuristic based on machine-learning concepts to improve the prediction of polygenic traits using gene scores. Our results show that for the classic polygenic traits height and BMI, 46.6% and 32.6% of the estimated polygenic genetic variance can be captured by boosted regression trees gene scores. These results demonstrate the potential of machine-learning methods to harness the considerable amount of information from large genetic meta-analyses. This is made possible through

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

partitioning of the genome, enabling training of regression trees over large number of observations. Indeed, a small training sample size (~200) was sufficient to greatly improve gene scores. As with other prediction problems involving machine-learning techniques, incremental improvements are to be expected with increases in sample size, use of additional predictor variables and availability of more precise summary association statistics.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Appendices

Gradient boosted regression trees

Boosted regression trees are powerful and versatile methods that combine otherwise weak classifiers to produce a strong learner for continuous outcome prediction[4]. They are thus ideal for prediction of SNP gene score weights ($\widehat{w}_{\text{pred}}$), where each fitted $\widehat{w}_{\text{pred}}$ gives the contribution of individual SNPs to the final gene score. The dependent variable used in boosted regression trees is constructed following:

$$w^*_{\text{pred}} = (\beta_{\text{obs}} - \beta_{\text{ext}})\text{sign}(\beta_{\text{ext}})$$

In other words, w^*_{pred} is derived to reflect the amount of deviation towards the null hypothesis of no association in the target population (β_{obs}) with respect to the externally derived summary association statistics estimates (β_{ext}). When $w^*_{\text{pred}} = 0$ then $\beta_{\text{obs}} = \beta_{\text{ext}}$, implicitly assuming regression coefficients from large meta-analyses provide the best initial estimates. While some information is lost because of this construct, the resulting estimates are more robust and the overall performance improved. Boosted regression trees can be expressed as

$$\text{E}\left(w^*_{\text{pred}} \middle| X_1, X_2, \ldots X_k\right) \sim \alpha + f(X)$$

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

where f is a regression function of trees with input variables $X = (X_1, X_2, \dots X_k)$. The gradient boost algorithm aims to minimize the expected square error loss with respect to f iteratively on weighted versions of the training data. While multiple SNP annotations could be included as inputs (i.e. $X_1, X_2, \dots X_k$), we only included the absolute value of the SNP regression coefficient from the external consortium to reflect the strength of association, irrespective of the direction of effect. Importantly, SNPs are divided into 5 distinct sets of contiguous SNPs (to avoid LD spillover) and fitted $\widehat{w}^*_{\text{pred}}$ which are used in calculation of the actual gene scores derived using the regression trees models trained on the remaining 4 sets. The observed regression coefficient (β_{obs}) of an individual SNP is thus never used directly or indirectly to derive its own gene score weight. Furthermore, the SNP annotations used in the regression trees model are independent of the population in which the gene score is applied as the UKBiobank and HRS were not part of GIANT[8, 9] or DIAGRAM[10]. The weights used in gene scores ($\widehat{w}_{\text{pred}}$) are given by the corresponding inverse transformation:

$$\widehat{w}_{\text{pred}} = \left(\beta_{\text{ext}} - \widehat{w}^*_{\text{pred}}\right)\text{sign}(\beta_{\text{ext}})$$

Gradient boosted regression trees models were fitted using the "GBM" R package (version 2.1.1) under a Gaussian distribution and squared error loss function. 2,000 trees were fitted with an interaction depth of 5, shrinkage parameter of 0.001 and bag fraction of 0.5. All other parameters were otherwise set to their default values.

Adjustment of regression coefficients for LD

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We propose a simple method to correct summary association regression coefficients for LD in such a way that all SNPs can be included in a gene score, irrespective of LD. Genotypes for *n* individuals at *m* SNPs are given by a matrix

$$\boldsymbol{X}_{n\times m} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \dots\ \ \boldsymbol{x}_n]^T$$

with each column vector $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ representing the coded genotypes for an individual. Without loss of generality, we assume each column of $\boldsymbol{X}$ (i.e. genotypes for a single SNP) to be standardized to have mean 0 and variance 1. For a standardized quantitative trait $\boldsymbol{y}$ with mean 0 and variance 1, the underlying linear model can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \textbf{(eq.1)}$$

$\boldsymbol{\beta}$ is a vector of true genetic effects that are fixed across individuals but random across SNPs, with mean $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{I}$ such that the total expected genetic variance

$$R^2_{\text{true}} = \mathrm{E}(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X})\sigma^2 = m\sigma^2$$

and $\boldsymbol{\varepsilon}$ the error term with mean $\mathbf{0}$ and covariance $(1 - m\sigma^2)\boldsymbol{I}$ so that the covariance of $\boldsymbol{y}$ is $\boldsymbol{I}$. Let $r^2_{d,k}$ denote the pairwise linkage disequilibrium (r^2) between the dth and k^{th}

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

SNPs. The LD adjustment (η_d) for the dth SNP is defined by the sum of r^2 between the dth SNP and the 100 SNPs upstream and downstream:

$$\eta_d = \sum_{k=d-100}^{k=d+100} r_{d,k}^2 \quad \textbf{(eq.2)}$$

with a distance of 100 SNPs assumed sufficient to ensure linkage equilibrium (other values might be used). Including only SNPs that are part of the gene score in the calculation of η_d, the LD-corrected regression coefficients are given by:

$$\tilde{b}_d = \frac{b_d{}^*}{\eta_d} \quad \textbf{(eq.3)}$$

where $b_d{}^*$ is the regression coefficient commonly reported in GWAS meta-analysis (assumed to have been standardized for allele frequency). Given $\boldsymbol{x_i}$, the genotypes of m SNPs for the ith individual, the gene score $g(\boldsymbol{x_i})$ is:

$$g(\boldsymbol{x_i}) = \boldsymbol{x_i^T}\tilde{\boldsymbol{b}} = \sum_{\boldsymbol{d}} \mathrm{x}_{i,d} \frac{b_d{}^*}{\eta_d} = \boldsymbol{x_i^T C b}^* \quad \textbf{(eq.4)}$$

where $\boldsymbol{C}$ is an $m \times m$ diagonal matrix with entries $(\frac{1}{\eta_1}, \frac{1}{\eta_2}, \ldots, \frac{1}{\eta_m})$. The prediction R^2 of the gene score in the target population is expressed as:

$$\mathrm{R}^2 = \frac{\mathrm{Cov}(g(X), \boldsymbol{y})^2}{\mathrm{Var}(g(X))\mathrm{Var}(\boldsymbol{y})} \quad \textbf{(eq.5)}$$

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The expected value can be approximated by:

$$\mathrm{E}[R^2] = \mathrm{E}\left[\frac{\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})^2}{\mathrm{Var}(g(\boldsymbol{X}))\mathrm{Var}(\boldsymbol{y})}\right] = \mathrm{E}\left[\frac{\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})^2}{\mathrm{Var}(g(\boldsymbol{X}))}\right] \sim \frac{\mathrm{E}\left[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})^2\right]}{\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))]} \sim \frac{\mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})]^2}{\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))]} \quad \textbf{(eq.6)}$$

and leading to

$$\mathrm{E}[R^2] \sim \frac{\mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})]^2}{\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))]} = \frac{(R^2_{\mathrm{true}})^2}{\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))]} < R^2_{\mathrm{true}} \quad \textbf{(eq.7)}$$

by deriving the following relations: (1) $\mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})] = R^2_{\mathrm{true}}$, implying the covariance between the gene score and the trait is an unbiased estimator of the true genetic variance; and (2) $\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))] > R^2_{\mathrm{true}}$ and thus $E[R^2] < R^2_{\mathrm{true}}$, implying the expected prediction R^2 must be bounded above by the true genetic variance. We demonstrate the validity of these two relations in the following sections and further verify with simulations (Supplementary Figure 1).

<u>(1) An Unbiased Estimator of the True Genetic Variance</u>

The sample covariance of the gene score with the observed $\boldsymbol{y}$ in the target sample is given by:

$$\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)y_i$$

$$= \frac{1}{n}(\boldsymbol{XCb}^*)^T(\boldsymbol{X\beta} + \boldsymbol{e})$$

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$= \frac{1}{nN}(\boldsymbol{XC}(\boldsymbol{X}^{*T}\boldsymbol{y}^{*}))^{T}(\boldsymbol{X\beta}+\boldsymbol{e})$$

$$= \frac{1}{nN}(\boldsymbol{XC}(\boldsymbol{X}^{*T}\boldsymbol{y}^{*}))^{T}(\boldsymbol{X\beta}+\boldsymbol{e})$$

$$= \frac{1}{nN}(\boldsymbol{\beta}^{T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{X\beta}+\boldsymbol{e}^{*T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{e}+\boldsymbol{\beta}^{T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{e}+\boldsymbol{e}^{*T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{X\beta})$$

(eq.8)

where $\boldsymbol{e}^{*}$ and $\boldsymbol{e}$ are the residual error in the unobserved population used to derive summary association statistics and the target population, respectively. The reported $\boldsymbol{b}^{*}$ in GWAS meta-analysis are constructed to estimate the univariate regression coefficients from the otherwise unobserved genotype matrix $\boldsymbol{X}^{*}{}_{\boldsymbol{N}\times\boldsymbol{m}}$ and quantitative trait $\boldsymbol{y}^{*}$:

$$\boldsymbol{b}^{*} \sim \frac{\boldsymbol{X}^{*T}\boldsymbol{y}}{N} = \frac{\boldsymbol{X}^{*T}(\boldsymbol{X}^{*}\boldsymbol{\beta}+\boldsymbol{e}^{*})}{N} = \frac{\boldsymbol{X}^{*T}\boldsymbol{X}^{*}}{N}\boldsymbol{\beta} + \frac{\boldsymbol{X}^{*T}\boldsymbol{e}^{*}}{N} \quad \textbf{(eq.9)}$$

Assume the target population is independent of the meta-analysis, i.e. $\boldsymbol{e}^{*}$ and $\boldsymbol{e}$ are independent, we have the expected value of the quadratic forms in **(eq.8)**:

$$\mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})]$$

$$= \frac{1}{nN}\mathrm{E}\left[(\boldsymbol{\beta}^{T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{X\beta}+\boldsymbol{e}^{*T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{e}+\boldsymbol{\beta}^{T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{e}+\boldsymbol{e}^{*T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{X\beta})\right]$$

$$= \frac{1}{nN}\mathrm{E}[\boldsymbol{\beta}^{T}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{C}\boldsymbol{X}^{T}\boldsymbol{X\beta}]$$

$$= \mathrm{tr}\left(\boldsymbol{\beta}^{T}\frac{\boldsymbol{X}^{*T}\boldsymbol{X}^{*}}{N}\boldsymbol{C}\frac{\boldsymbol{X}^{T}\boldsymbol{X}}{\mathrm{n}}\boldsymbol{\beta}\right)$$

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$= \sigma^2 \mathrm{tr}\left(\frac{\boldsymbol{X}^{*T}\boldsymbol{X}^{*}}{N}\boldsymbol{C}\frac{\boldsymbol{X}^{T}\boldsymbol{X}}{\mathrm{n}}\right)$$

$$= \sigma^2 m = R^2_{\mathrm{true}} \quad \textbf{(eq.10)}$$

This equality holds for all positive definite matrices of the form $\frac{\boldsymbol{X}^{*T}\boldsymbol{X}^{*}}{N} C \frac{\boldsymbol{X}^{T}\boldsymbol{X}}{\mathrm{n}}$, assuming the LD structure in the two populations is identical. We have thus shown that $\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})$ is an unbiased estimator of the true genetic variance.

<u>(2) Variance of the Gene Score</u>

The denominator in **(eq.7)**, $\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))]$, can be shown to be greater than R^2_{true}:

$$R^2_{\mathrm{true}} = \mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})] = \mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{X\beta})]$$

$$\mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{X\beta})] \leq \mathrm{E}[\sqrt{\mathrm{Var}(g(\boldsymbol{X}))\mathrm{Var}(\boldsymbol{X\beta})}] = \mathrm{E}[\sqrt{\mathrm{Var}(g(\boldsymbol{X}))R^2_{\mathrm{true}}}]$$

And thus:

$$R^2_{\mathrm{true}} \leq \mathrm{E}[\sqrt{\mathrm{Var}(g(\boldsymbol{X}))R^2_{\mathrm{true}}}] \text{ and } \sqrt{R^2_{\mathrm{true}}} \leq \mathrm{E}[\sqrt{\mathrm{Var}(g(\boldsymbol{X}))}]$$

Giving:

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))] \geq \mathrm{E}[\sqrt{\mathrm{Var}(g(\boldsymbol{X}))}]^2 \geq R^2_{\mathrm{true}} \quad \textbf{(eq.11)}$$

From the above inequality, we can conclude that $\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))]$ is biased and will always be greater or equal than the true genetic variance.

<u>A note on prediction R^2 of gene score as compared to true genetic variance</u>

The LD correction proposed thus offers a tradeoff between bias and variance, whereby genetic variance estimates are unbiased as $\mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})] = R^2_{\mathrm{true}}$ but $\mathrm{E}[\mathrm{Var}(g(\boldsymbol{X}))] \geq R^2_{\mathrm{true}}$ implying that $R^2 \leq R^2_{\mathrm{true}}$. It can be shown that $R^2 = R^2_{\mathrm{true}}$ in simple cases where pairwise r^2 LD is either 0 or 1 and summary association statistics are derived from an asymptotically large sample. However, in more common scenarios with partial LD $R^2 < R^2_{\mathrm{true}}$ reflecting the loss of information when, for example, two SNPs are in partial LD and have true genetic effects with opposite directions. To assess the importance of this effect in plausible situations we performed simulations. 5,000 individuals were simulated for 450 contiguous SNPs using phased haplotypes from the 1000 Genomes Project. The genetic effect of each SNP was randomly selected from a normal distribution according to a pre-defined, unobserved, true regional genetic variance that assumed genome-wide heritability varying from 0 to 0.5. For each genetic variance set-point, 1,000 simulations were completed and a gene score incorporating LD correction derived. The average (±SD) gene score prediction R^2, gene score variance and covariance between gene score and true (unobserved) genetic effect calculated (Supplementary Figure 1). Based on these simulations, we confirmed that (1) LD-corrected gene scores are unbiased

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

estimators of true genetic variance (i.e. $\mathrm{E}[\mathrm{Cov}(g(\boldsymbol{X}),\boldsymbol{y})] = R^2_{\mathrm{true}}$), and (2) variance of gene score is indeed higher than true genetic variance. We further estimated the loss of information at ~12%, or in other words gene score prediction R^2 was on average ~88% of true genetic effect.

UKBiobank Study

The UKBiobank[13] (UKB) is a large population-based study from the United Kingdom. 152,249 participants were genotyped using either the UK BiLEVE or the UK Biobank Affymetrix Axiom arrays. 140,215 participants were of European (British and Irish) Caucasian ancestry and included in the present analysis. Genotypes were imputed using the UK10K reference panel using IMPUTE2, resulting in ~72M SNPs. Height and BMI was adjusted for age and sex in all analyses; and to further mitigate the effect of outliers, values outside the 1st and 99th percentile range were removed. All analyses were adjusted for the first 15 genetic principal components unless stated otherwise. The UKB was not part of the GIANT meta-analysis of height and BMI[11, 12], or of the DIAGRAM consortium for diabetes[10]. There are 6,746 individuals with prevalent diabetes in the subset of the UKB included in the current report. We randomly selected 6,746 individuals without diabetes as paired controls on a 1:1 basis. Next, we randomly sampled 1,000 case-controls pairs as the calibration set, with the remaining 5,746 pairs constituting the validation dataset.

Health Retirement Study

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We downloaded publically available genome-wide data that are part of the Health Retirement Study (HRS; dbGaP Study Accession: phs000428.v1.p1) generated using the Human Omni2.5-Quad BeadChip (Illumina). HRS quality control criteria were used for filtering of both genotype and phenotype data, namely: (1) SNPs and individuals with missingness higher than 2% were excluded, (2) related individuals were excluded, (3) only participants with self-reported European ancestry and genetically confirmed by principal component analysis were included, (4) individuals for whom the reported sex does not match their genetic sex were excluded, (5) SNPs with Hardy-Weinberg equilibrium $p < 1x10^{-6}$ were excluded, (6) SNPs with minor allele frequency lower than 0.02 were removed. The final dataset included 8,292 European participants genotyped for 688,398 SNPs. Height and BMI was adjusted for age and sex in all analyses; and to further mitigate the effect of outliers, values outside the 1st and 99th percentile range were removed. There were 1,815 individuals with diabetes and 6,477 controls. All analyses were adjusted for the first 20 genetic principal components unless stated otherwise. HRS was not part of the GIANT meta-analysis of height and BMI[11, 12], or of the DIAGRAM consortium for diabetes[10].

Pruning and thresholding gene scores, LDpred and other alternative methods

Pruning and thresholding (P+T) polygenic scores were derived using the "clump" function of PLINK[14] with an LD r^2 threshold of 0.2 and testing *p*-value thresholds in a continuous manner from the most to the least significant association. LDpred adjusts

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

GWAS summary statistics for the effects of linkage disequilibrium, providing re-weighted effect estimates that are then used in gene scores[3]. LDpred was run as recommended by authors, including both the data synchronization and LDpred steps. LDpred requires specification of the fraction of SNPs assumed to be causal. For each model, we tested causal fractions of 1 (infinitesimal), 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001 as recommended. Results are presented using the causal fraction giving the best results only. A heritability estimate is also required by the algorithm and is estimated from summary association statistics by LDpred. As a sensitivity analysis, we additionally used heritability estimates given by the variance component models in the UKB. Results were consistent and only the default option is shown. Polygenic genetic variance (i.e. narrow sense heritability) was estimated for height and BMI in the UKB using variance components, as implemented in GCTA[7]. All LD measures or related estimates used throughout the manuscript were derived from UKB calibration set genotypes.

Conflicts of Interests

The authors are listed as inventors on patent disclosures owned by McMaster University and related to trait prediction using genetic data.

Author Contributions:

G. P. designed the experiment; G.P and W.Q.D. wrote the manuscript; S. M. analyzed the data and prepared tables and figures; All authors reviewed the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplemental Data

Supplemental data include 1 figure.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

REFERENCES

1. Yang J, *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**, 565--569 (2010).

2. International Schizophrenia C, *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

3. Vilhjalmsson BJ, *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576-592 (2015).

4. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning : data mining, inference, and prediction*, 2nd edn. Springer (2009).

5. Speliotes EK, *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics* **42**, 937-948 (2010).

6. Chen CY, Han J, Hunter DJ, Kraft P, Price AL. Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. *Genet Epidemiol* **39**, 427-438 (2015).

7. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).

8. Wood AR, *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics* **46**, 1173-1186 (2014).

9. Locke AE, *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

10. Morris AP, *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981-990 (2012).

11. Berndt SI, *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature genetics* **45**, 501-512 (2013).

12. Lango Allen H, *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838 (2010).

13. Sudlow C, *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

14. Purcell S, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure Legends

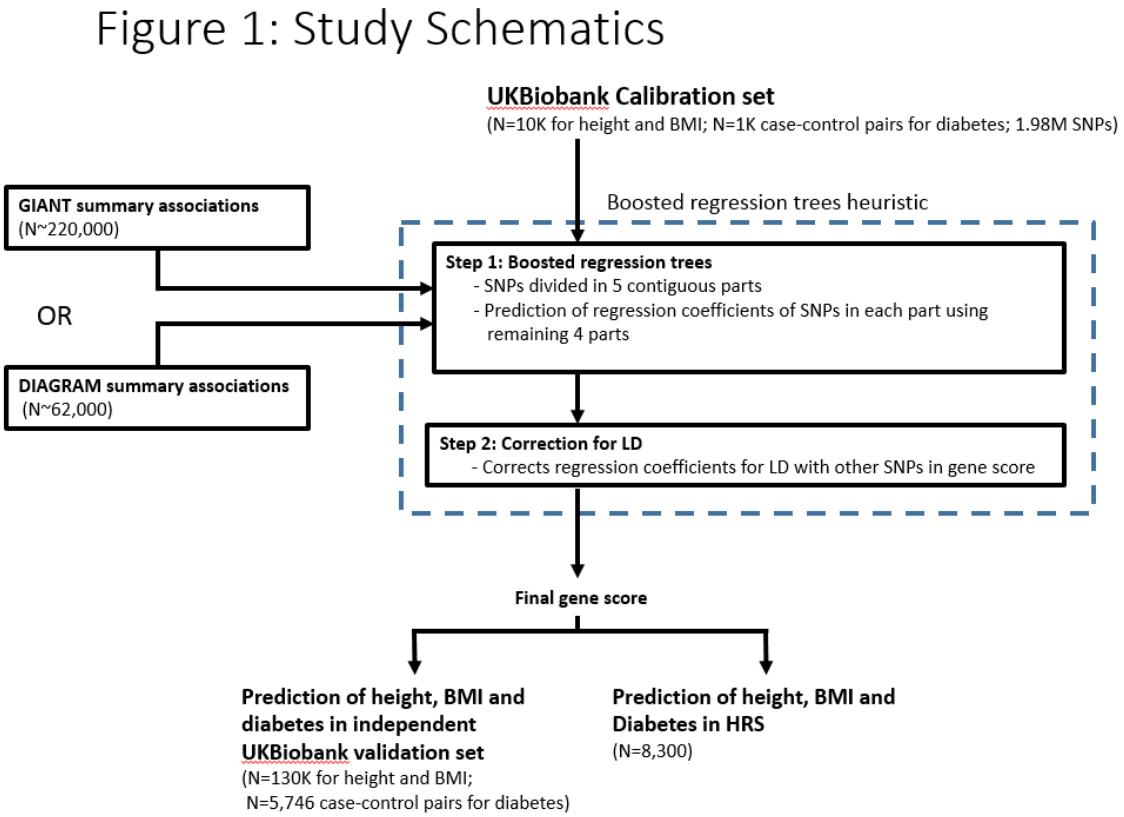

Figure 1: An overview of Machine-learning Heuristic (MLH) for gene scores and study design

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

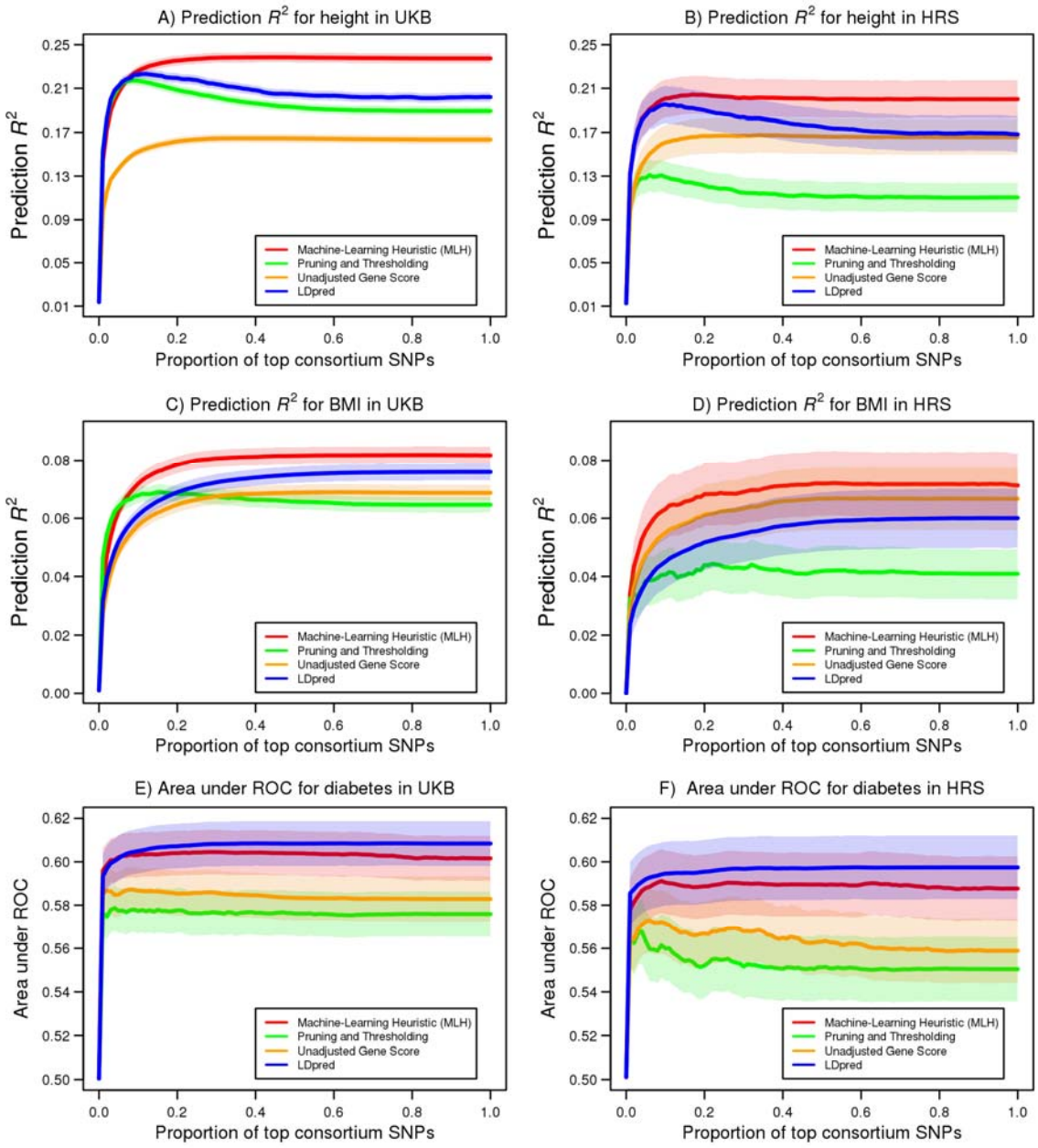

Figure 2: Discrimination of height, BMI and diabetes gene scores

Gene scores prediction R^2 as a function of the proportion of SNPs included for height (Panel A) and BMI (Panel C) in the UKB validation set (N=130,215), with 95% confidence intervals. A total of 1.98M SNPs were considered and SNPs were ordered from the most to the least significant according to GIANT summary association statistics.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

LPpred requires determining the causal fraction of SNPs and only the best scores are illustrated, setting the causal fraction at 0.3 and 0.01 for height and BMI, respectively. Prediction R^2 of UKB gene scores in HRS (N=8,292) is similarly illustrated for height (Panel B) and BMI (Panel D). UKB gene scores were tested in HRS without any further fitting or adjustment. The area under the ROC (AUROC) is illustrated for diabetes in the UKB validation set (Panel E) and HRS (Panel F), with 95% confidence intervals. The LDpred causal fraction was 0.003 for diabetes, as determined in the UKB calibration set.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

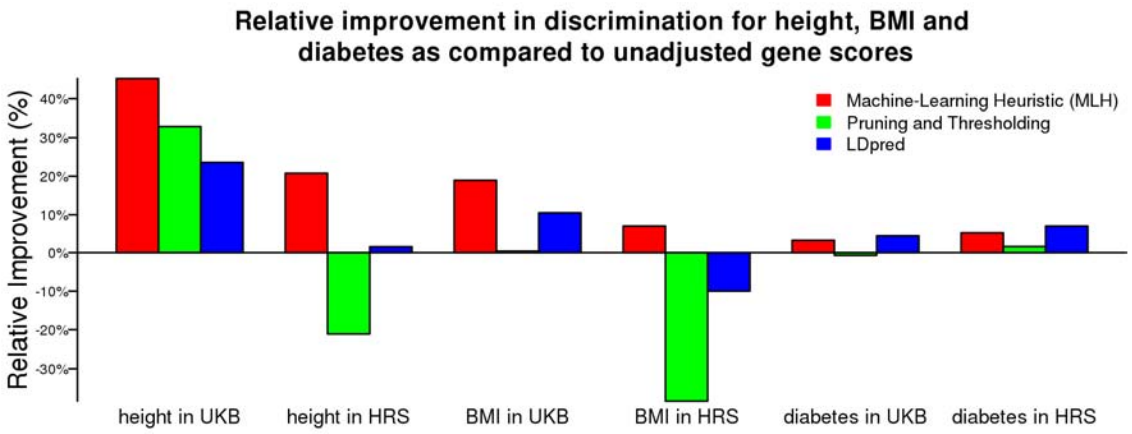

Figure 3: The relative improvement in discrimination for height, BMI and diabetes as compared to unadjusted gene scores

The relative improvement in prediction R^2 of gene scores as compared to the unadjusted gene score is illustrated for height and BMI in the UKB validation set and HRS. For diabetes, the relative improvement in AUROC is illustrated. In all cases, the optimal gene score was derived from the UKB calibration set and tested without any further fitting or adjustment.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

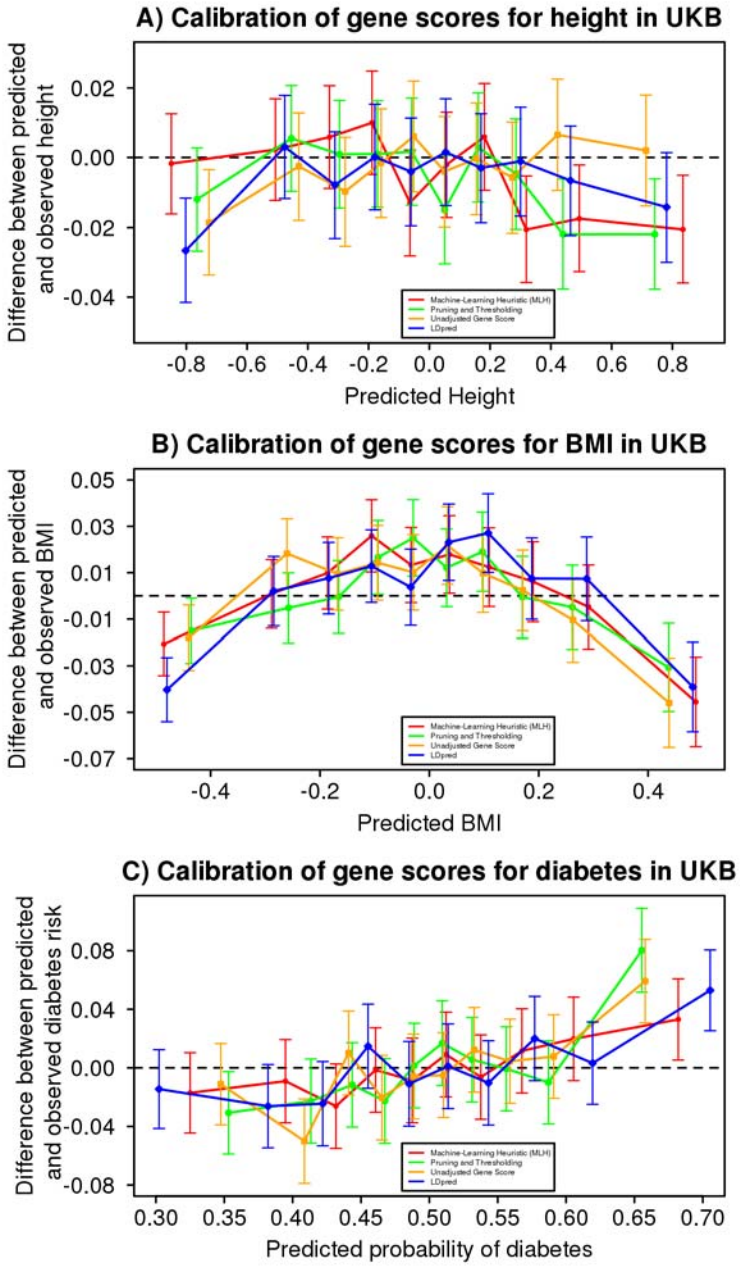

Figure 4: Calibration of height, BMI and diabetes gene scores.

For each trait and gene score method, the UKB validation set was divided into deciles of gene score. For each decile, the difference between the mean observed and predicted trait (95% confidence interval) is illustrated as function of the mean predicted trait for that

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

gene score decile. The trait is expressed per SD unit for height (Panel A) and BMI (Panel B). A similar analysis was performed for diabetes, whereby the difference between the observed probability of diabetes and predicted probability is illustrated as function of the predicted probability for each gene score decile.

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

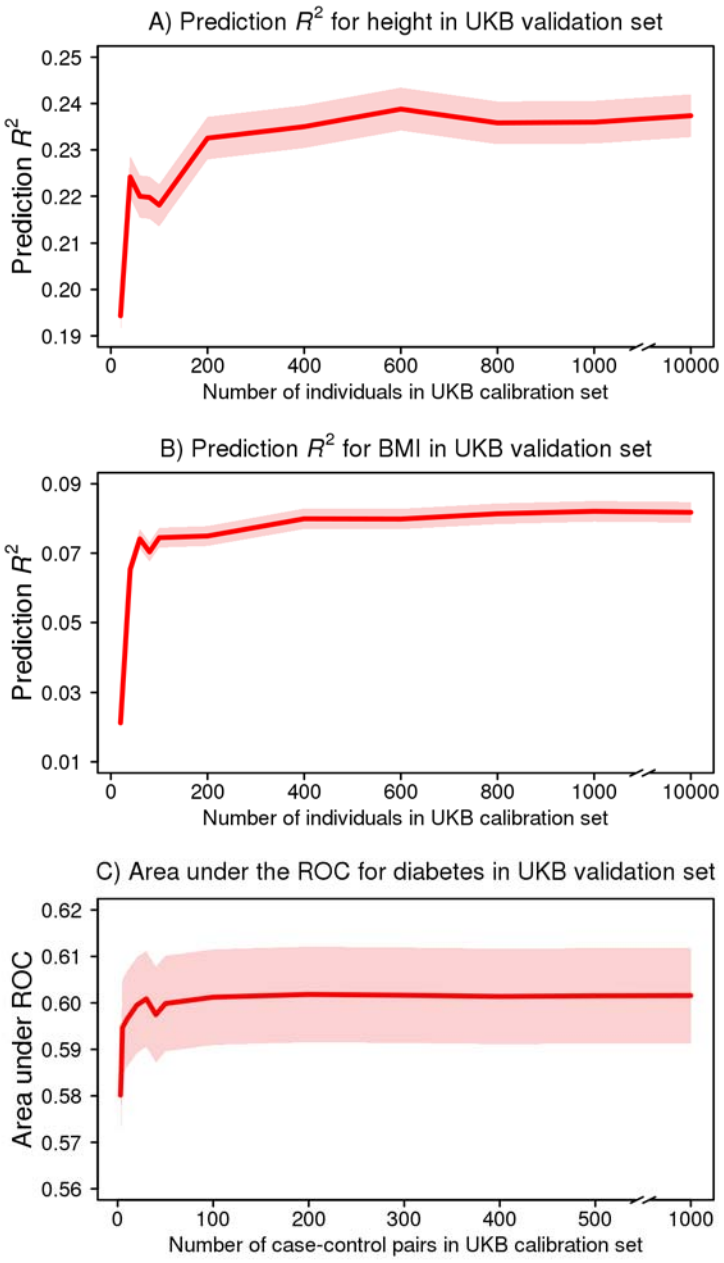

Figure 5: MLH gene score discrimination as function of calibration set sample size

bioRxiv preprint doi: https://doi.org/10.1101/107409; this version posted February 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The size of the UKB calibration set was varied from 20 to 10,000 for height and BMI, and from 3 to 1,000 case-control pairs for diabetes. For each calibration sample size, discrimination of the corresponding gene score was calculated in the independent UKB validation set (N=130,215 for height and BMI; N=5,746 case-control pairs for diabetes).