Accurate Genomic Prediction Of Human Height

Louis Lello[1], Steven G. Avery[1], Laurent Tellier[1,3,5], Ana I. Vazquez[2], Gustavo de los Campos[2,4], and Stephen D.H. Hsu[1,3]

[1]*Michigan State University*
Department of Physics and Astronomy
East Lansing, MI 48824

[2]*Michigan State University*
Department of Epidemiology and Biostatistics
East Lansing, MI 48824

[3]*Cognitive Genomics Laboratory, BGI*

[4]*Michigan State University*
Department of Statistics and Probability
East Lansing, MI 48824

[5]*University of Copenhagen*
Department of Biology, Functional Genetics
Copenhagen, DK

Abstract

We construct genomic predictors for heritable and extremely complex human quantitative traits (height, heel bone density, and educational attainment) using modern methods in high dimensional statistics (i.e., machine learning). Replication tests show that these predictors capture, respectively, ~40, 20, and 9 percent of total variance for the three traits. For example, predicted heights correlate ~0.65 with actual height; actual heights of most individuals in validation samples are within a few cm of the prediction. The variance captured for height is comparable to the estimated SNP heritability from GCTA (GREML) analysis, and seems to be close to its asymptotic value (i.e., as sample size goes to infinity), suggesting that we have captured most of the heritability for the SNPs used. Thus, our results resolve the common SNP portion of the "missing heritability" problem – i.e., the gap between prediction R-squared and SNP heritability. The ~20k activated SNPs in our height predictor reveal the genetic architecture of human height, at least for common SNPs. Our primary dataset is the UK Biobank cohort, comprised of almost 500k individual genotypes with multiple phenotypes. We also use other datasets and SNPs found in earlier GWAS for out-of-sample validation of our results.

1 Introduction

Recent estimates [1] suggest that common SNPs account for significant heritability of complex traits such as height, heel bone density, and educational attainment (EA). Large GWAS studies of these traits have identified many associated SNPs at genome-wide significance ($p < 5 \times 10^{-8}$) [2–6]. However, the total variance accounted for by these SNPs is still a small fraction of the trait heritability and of the proportion of variance that could be captured by regression on common SNPs as suggested by SNP heritability estimates [7].

The simplest hypothesis explaining this (so far) "missing heritability" is that previous studies have not had enough statistical power to identify most of the relevant SNPs, due to their small effect size, low minor-allele frequency (MAF), or both. In this letter, we provide evidence in support of this hypothesis by constructing genomic predictors capturing much of the estimated SNP heritability. We make use of a newly available large data set (the UK Biobank 500k genomes release) and new computational methods.

Association studies (GWAS) focus on reliable (high confidence) identification of associated SNPs. In contrast, genomic prediction based on whole genome regression methods [8] seeks to construct the most accurate predictor of phenotype, and tolerates possible inclusion of a small fraction of false-positive SNPs in the predictor set. The SNP heritability of the molecular markers used to build the predictor can be interpreted as an upper bound to the variance that could be captured by the predictor.

While identification of GWAS SNPs is accomplished by single SNP regression, construction of a best predictor is a global optimization problem in the high dimensional space of possible effect sizes of all SNPs. In this letter we use L_1-penalized regression (LASSO or Compressed Sensing) to obtain our predictors. This method is particularly effective in cases where only a small subset of variables have non-zero effect on the predicted quantity (i.e., the effects vector is sparse, or approximately sparse). In earlier work [9] it was shown that matrices of human genomes are good compressed sensors, and that they are in the universality class of Gaussian random matrices. The L_1 algorithm exhibits phase transition behavior as the sample size and penalization parameter are varied; this behavior can be used to optimize the penalization as a function of sample size. Technical details are provided in the Methods section below.

Beyond the theoretical considerations given above, the practical outcome of our work is to significantly improve accuracy in genomic prediction of complex phenotypes. Using these predictors, one can, for example, reliably identify outliers in the population based on DNA alone. The activated SNPs in the predictors (i.e., those that have been assigned non-zero effect size by the LASSO algorithm) are likely to be associated with the phenotype, although they may not reach genome-wide significance in ordinary regression analysis. While there may be some contamination of false-positives among these SNPs, one can nevertheless infer properties of the overall genetic architecture of the trait (e.g., distribution of effect sizes with MAF).

2 Data and Methods

Our main dataset is the July 2017 release of nearly 500k UK Biobank genotypes and associated phenotypes [10, 11]. (See Supplement for more detailed description of data, quality control, algorithms, and computations.)

We compute an estimator $\vec{\beta}^*$ for the vector of linear effects, $\vec{\beta} \in \mathbb{R}^p$, using L_1-penalized regression (LASSO) [12]. This corresponds to minimizing the objective function below (phenotypes $\vec{y}$ are age and gender adjusted; both $\vec{y}$ and genotype values X are standardized).

$$\vec{\beta}^* = \underset{\vec{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, O_\lambda(\vec{y}, X; \vec{\beta}), \qquad O_\lambda(\vec{y}, X; \vec{\beta}) = \frac{1}{2}\left\|\vec{y} - X\vec{\beta}\right\|^2 + n\lambda\|\vec{\beta}\|_1, \tag{1}$$

where λ is a penalty (hyper-)parameter and the L_1 norm is defined to be the sum of the absolute values of the coefficients

$$\|\vec{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

The resulting effects vector $\vec{\beta}^*$ defines a linear predictive model which captures a large portion of the heritable genetic variance.

In our procedure, a first screening based on standard single marker regression is performed on the training set to reduce the set of candidate SNPs from 645,589 SNPs that passed QC (Supplement) to the top $p = 50$k and 100k by statistical significance.

3 Results

Figure (1) displays results from a typical LASSO run for height. 5 non-overlapping sets of 5k individuals each were held back from LASSO training using the top 100k candidate SNPs. For each value of the L_1 penalization λ the resulting predictor $\vec{\beta}^*$ is applied to the genomes of the holdback sets and the correlation between predicted and actual height is computed. A phase transition (region of rapid variation in results) is expected and occurs at roughly $10 < -\ln(\lambda) < 12$. The penalization is reduced until the correlation is maximized. In Figure (1), the correlation is shown as a function of number of SNPs assigned non-zero effect sizes (i.e., activated) by LASSO. In the phase transition regime, where correlation rapidly increases, the number of activated SNPs grows rapidly from about zero to 7k. Each of the 5 colored curves in the figure corresponds to a training run on 453k individuals, with a different 5k held back (and slightly different training set) for each run. The phase transition is shown in terms of the penalization $-\ln(\lambda)$ in Figure (2).

Figure (3) shows the correlation between predicted and actual phenotypes in a validation set of 5000 individuals not used in the training optimization described in above - this is shown both for height and heel bone mineral density. The horizontal axis shows the number of individuals used in the training set and the error bars reflect 1 SD uncertainty estimated from five replications. The correlation obtained indicates convergence to an asymptotic value of somewhat less than 0.7 (corresponding to roughly 50 percent of total variance) for height, and perhaps 0.45 for heel bone mineral density. Figure (4) shows a scatterplot (each point is an individual) of predicted and actual height for 2000 individuals (roughly equal numbers of males and females) not used in the training. The actual heights of most individuals are within about 3 cm of the predicted value.

The corresponding result for Educational Attainment does not indicate any approach to a limiting value. Using all the data in the sample, we obtain maximum correlation of ~ 0.3, activating about 10k SNPs. Presumably, significantly more or higher quality data will be required to capture most of the SNP heritability of this trait.

The number of activated SNPs in the optimal predictors for height and bone density is roughly 20k. Increasing the number of candidate SNPs used from $p = 50$k to $p = 100$k increased the maximum correlation of the predictors somewhat, but did not change the number of activated SNPs significantly.

We computed the GCTA heritability for the top 50k SNPs used, using randomly selected sets of 20k individuals. For height, $h^2 = 0.5003 \pm 0.0209$ (95%) and heel bone density $h^2 = 0.4355 \pm 0.0226$ (95%). However, there has been debate in the literature over the statistical properties of

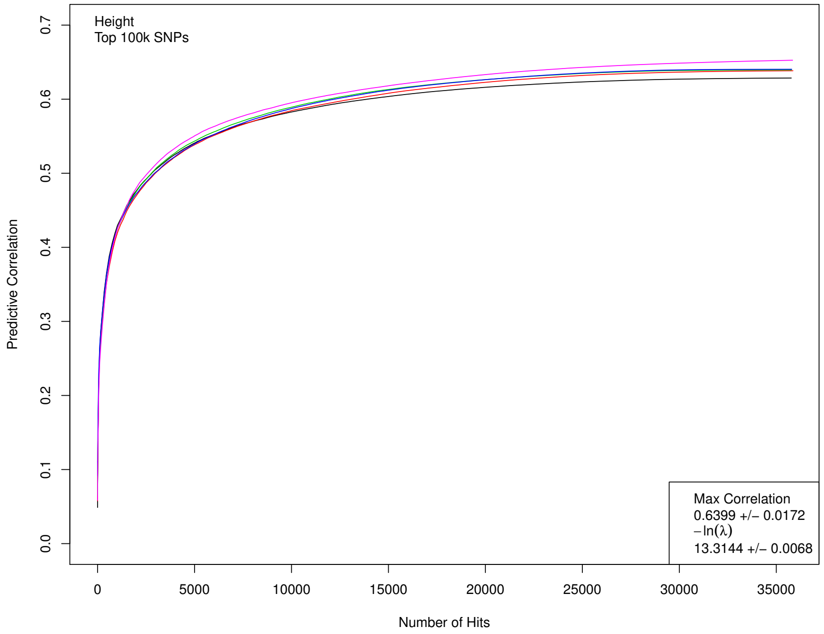

Figure 1: Correlation between actual and predicted heights as a function of the number of SNP hits activated in the predictor. While difficult to visually separate, each line represents the training of a predictor using 453k individuals. Correlation is computed on 5k individuals not used in training. The phase transition region (roughly, $10 < -\ln(\lambda) < 12$) corresponds to rapid growth in correlation on this graph, with number of hits growing from near 0 to over 5000.

GREML estimates of SNP heritability and it is not clear that standard estimation methods yield reasonably unbiased estimates even with large sample size [7, 13–16]. Therefore, we suggest that GCTA estimates of SNP heritability should only be used as a rough guide. Perhaps the only way to determine the heritability of a trait over a specific set of genomic variants is to build the best possible predictor [17] (i.e., with, in principle, unlimited sample size n) to determine how much variance can be accounted for.

For height we tested out-of-sample validity by building a predictor model using SNPs whose state is available for both UKBB individuals (via imputation) and on Atherosclerosis Risk in Communities Study (ARIC) [18] individuals (the latter is a US sample). This SNP set differs from the one used above, and is somewhat more restricted due to the different genotyping arrays used by UKBB and ARIC. Training was done on UKBB data and out-of-sample validity tested on ARIC data. A ~5% decrease in maximum correlation results from the restriction of SNPs and limitations of imputation: the correlation fell to ~0.58 (from 0.61) while testing within the UKBB. On ARIC participants the correlation drops further by ~7%, with a maximum correlation of ~0.54. Only this latter decrease in predictive power is really due to out-of-sample effects. It is plausible that if ARIC participants were genotyped on the same array as the UKBB training set there would only be a ~7% difference in predictor performance. An ARIC scatterplot analogous to Figure (4) is shown in the Supplement. Most ARIC individuals have actual height within 4 cm or less of predicted height.

We also checked (see Supplement) that familial relationships in UKBB do not have an important impact on our results. LASSO training was done both on the full set of data and on a smaller data set where all first degree cousin or stronger relations were removed (kinship > 0.10). After filtering

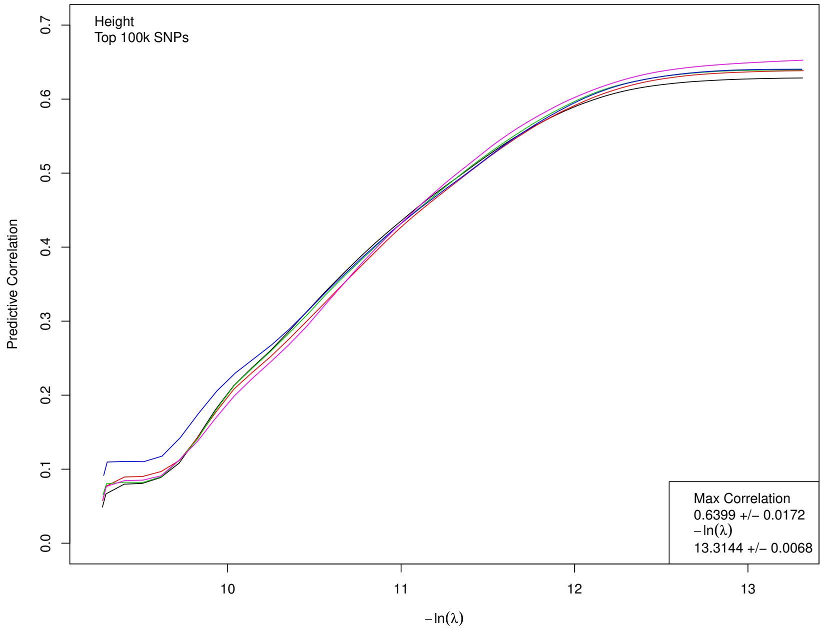

Figure 2: Correlation between actual and predicted heights as a function of L_1 penalization λ. Each line represents the training of a predictor using 453k individuals. Correlation is computed on 5k individuals not used in training.

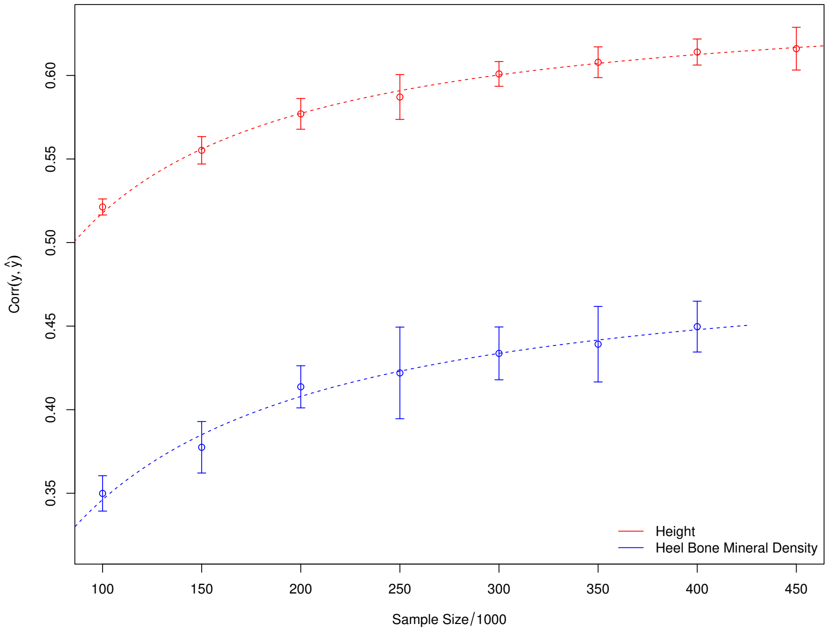

Figure 3: Correlation between predicted and actual height as number of individuals n in training set is varied. $p = 50\text{k}$ candidate SNPs used in optimization. Fit lines of the form Corr $\sim \frac{n}{n+b}$ are included to aid visualization.

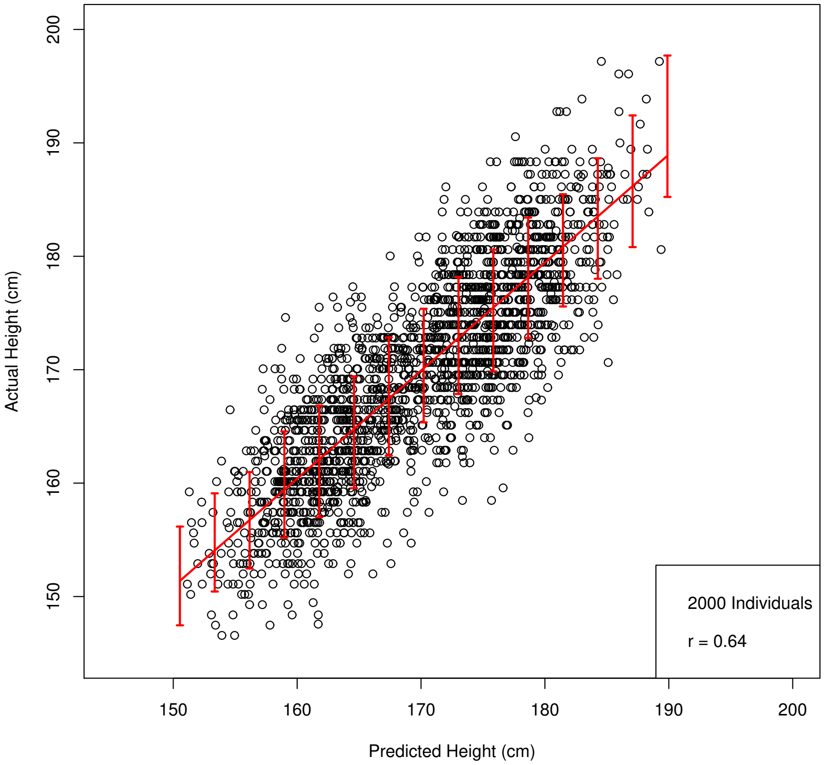

Figure 4: Actual height (cm) versus predicted height (cm) using 2000 randomly selected individuals held back from predictor optimization. (Roughly equal numbers of males and females; no corrections of actual height for age or gender). Error bars indicate ±1 SD range computed using larger validation set.

for kinship on the calls, this left 423,510 individuals for height and 382,727 individuals for heel bone density. This unrelated dataset was used for model training using random sets of 100k, 150k, ... , 400k individuals and there was no discernible difference in the results between using a training set drawn from the set of 423,510 kinship-filtered individuals and individuals from the unfiltered set.

The genetic architecture of a height model is displayed in Figure (5), which shows the effect size (minor allele) for each activated SNP. The horizontal axis represents the SNP position in the genome, if each chromosome (1-22) were laid end to end to form a continuous linear region. The specific height predictor from which these SNPs are taken was built from 50k candidate SNPs and achieves a correlation between actual and predicted height of ~0.61. The activated SNPs seem to be uniformly distributed across the genome.

There is significant overlap between regions of the genome near previously known SNPs and regions identified by our algorithm (Supplement). However, our activated SNPs are roughly uniformly distributed over the entire genome, and number in the many thousands for each trait. This means that many of our SNPs, including some of those that account for the most variance, are in regions not previously identified by earlier GWAS.

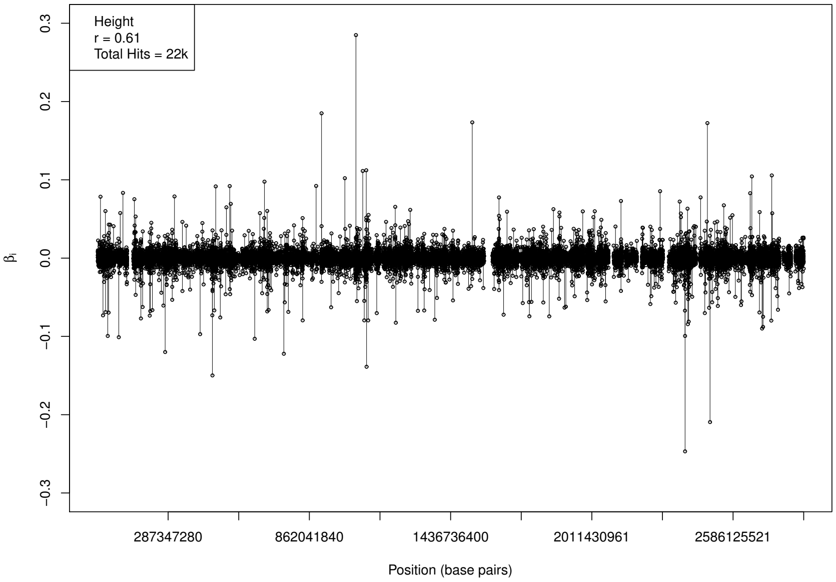

Figure 5: Effect size (minor allele) for each activated SNP in a predictor model. The horizontal axis represents the SNP position in the genome, if chromosomes (1-22) were connected end to end to form a continuous linear region. Activated SNPs are distributed roughly uniformly throughout the genome.

4 Discussion

Until recently most work with large genomic datasets has focused on finding *associations* between markers (e.g., SNPs) and phenotype [17]. In contrast, we focused on optimal *prediction* of phenotype from available data. We show that much of the expected heritability from common SNPs can be captured, even for complex traits affected by thousands of variants. Recent studies using data from the interim release of the UKBB reported prediction correlations of about 0.5 for human height using roughly 100K individuals in the training [19]. These studies forecast further improvement of prediction accuracy with increased sample size, which have been confirmed here.

We are optimistic that, given enough data and high quality phenotypes, results similar to those for height might be obtained for other quantitative traits, such as cognitive ability or specific disease risk. There are numerous disease conditions with heritability in the 0.5 range, such as Alzheimer's, Type I Diabetes, Obesity, Ovarian Cancer, Schizophrenia, etc [20]. Even if the heritable risk for these conditions is controlled by thousands of genetic variants, our work suggests that effective predictors might be obtainable (i.e., comparable to the height predictor in Figure (4)). This would allow identification of individuals at high risk from genotypes alone. The public health benefits are potentially enormous.

We can roughly estimate the amount of case-control data required to capture most of the variance in disease risk. For a quantitative trait (e.g., height) with $h^2 \sim 0.5$, our simulations [9] predict that the phase transition in LASSO performance occurs at $n \sim 30s$ where n is the number of individuals in the sample and s is the sparsity of the trait (i.e., number of variants with non-zero effect sizes). For case-control data, we find $n \sim 100s$ (where n means number of cases with equal number controls) is sufficient. Thus, using our methods, analysis of $\sim$ 100k cases together with a similar number of controls might allow good prediction of highly heritable disease risk, even if the genetic architecture is complex and depends on a thousand or more genetic variants.

Acknowledgments LL, SA, and SH acknowledge support from the Office of the Vice-President for Research at MSU. The authors are grateful for useful correspondence and discussion with Alexander Grueneberg and Hwasoon Kim. We also acknowledge support from the NIH Grants R01GM099992 and R01GM101219, and NSF Grant IOS-1444543, subaward UFDSP00010707.

A SUPPLEMENT: Methods

A.1 UKBB Dataset QC

In July 2017, the UK Biobank [10, 11] released a set of 488,377 genotyped individuals which were genotyped using two Affymetrix platforms—approximately 50,000 samples on the UK BiLEVE Axiom array and the remainder on the UK Biobank Axiom array. The initial genotype information was collected for 488,377 individuals for 805,426 SNPs and then subsequently imputed. Quality Control was done on the un-imputed data by removing SNPs which had missing call rates over 3%, removing individuals which had missing call rates over 10% and, so as not to deal with very rare variants, removing SNPs which had minor frequencies below 0.1%. The resulting genetic data contained 645,589 SNPs and 488,371 individuals. This set was then further filtered for self-reported caucasians for whom the necessary phenotype measurements were available: for height, the number of remaining individuals was 457,484; for heel bone mineral density there were 413,444 individuals; and for educational attainment, there were 455,637 individuals.

The imputed data set was generated using the set of 805,426 raw markers using the Haplotype Reference Consortium and UK10K haplotype resources. After imputation and initial QC, there were a total of 92,060,613 SNPs and 487,411 individuals. From this imputed data, further quality control was performed using Plink version 1.9 by excluding SNPs and samples which had missing call rates exceeding 3% and also removing snps with minor allele frequency below 0.1%. For out-of-sample validation of height, we extracted SNPs which survived the prior quality control measures, and are also present in a second dataset from the Atherosclerosis Risk in Communities Study (ARIC) [18]. This resulted in a total of 632,155 SNPs and 464,192 samples. All quality control steps, except those performed by the UK Biobank involving the imputation, were performed using version 1.9 of the Plink software [21].

A.2 Confounding variables: age, sex and family structure

All traits for self-identified Caucasians were adjusted on the basis of age and sex. The phenotypes for self-reported Caucasians were adjusted by z-scoring the phenotypes amongst all individuals of the same sex. To correct for the effects of societal changes, a univariate linear regression was performed on z-scored phenotypes using year-of-birth as the dependent variable. The adjusted phenotype was set equal to the residual of the z-scored phenotype and the regression line. Before making these corrections, it was shown that the mean phenotypic value was indeed increasing with year-of-birth—this was seen in all three phenotypes: height, heel bone mineral density and educational attainment.

Relatedness calculations were provided with the UKBB dataset in order to account for family structure and cryptic relatedness. There were 107,163 familial relationships identified amongst UKBB participants which were at the level of third cousins or higher and, due to the large number of relationships, filtering out these individuals results in a nontrivial decrease in the size of data available for model selection. To investigate the relevance of this issue, LASSO training was done both on the full set of data and on a smaller data set where all first degree cousin or stronger relations were removed (kinship > 0.10). After filtering for kinship on the calls, this left 423,510 individuals for height and 382,727 individuals for heel bone density. This unrelated dataset was used for model training using random sets of 100, 150, ... , 400 thousand individuals and there was no discernible difference in the results between using a training set drawn from the set of 423,510 kinship-filtered individuals and individuals from the unfiltered set. Therefore we do not believe that the familial relationships have an important impact on our results.

A.3 L$_1$-penalized regression

Consider the regression problem in generality. We have n observations of the phenotype, y_I, with $I = 1, \ldots, n$ as the vector $\vec{y}$. The genotype data is encoded in the $n \times p$ design matrix X_{Ij} with $j = 1, \ldots, p$. The X_{Ij} is the number of copies of the most frequent minor allele of the jth SNP for the Ith person, and thus takes values 0, 1, or 2. Missing values are mean-imputed.

We use a standard linear model for the dependence of y on the SNP data x_j. That is, we assume a relationship of the form

$$y_I = y_0 + \hat{\vec{\beta}} \cdot \vec{x}_I + e_I, \tag{2}$$

where the errors, e_I, are assumed to be (identically and independent) normally distributed with unknown variance σ_e. The errors, e_I, receive contributions from potential environmental effects, gene–gene nonlinear effects, and gene–environment nonlinear effects. For discussion of methods to recover nonlinear effects, see [22].

We compute an estimator $\vec{\beta}^*$ for the vector of linear effects, $\hat{\vec{\beta}} \in \mathbb{R}^p$, using L$_1$-penalized regression (LASSO) [12]. This corresponds to minimizing the objective function (after standardizing $\vec{y}$ and X)

$$\vec{\beta}^* = \underset{\vec{\beta} \in \mathbb{R}^p}{\operatorname{argmin}}\, O_\lambda(\vec{y}, X; \vec{\beta}), \qquad O_\lambda(\vec{y}, X; \vec{\beta}) = \frac{1}{2}\left\|\vec{y} - X\vec{\beta}\right\|^2 + n\lambda\|\vec{\beta}\|_1, \tag{3}$$

where λ is a penalty (hyper-)parameter and the L$_1$ norm is defined to be the sum of the absolute values of the coefficients

$$\|\vec{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

(We use $\|\cdot\|$ with no subscript to denote the standard L$_2$ norm.) The extra factor of n in the second term is a convention that factors out the explicit sample size scaling of n. The squaring in the first term is (implicitly) of the L$_2$ norm of the residual.

The first term is the standard ordinary least-squares (OLS) loss function. The effect of the second term is to regularize the regression problem by favoring sparse solutions with the nonzero coefficients shrunk toward 0. This seems appropriate for genomic problems, since we expect that for any given phenotype most SNPs have no effect. Biasing the nonzero coefficients toward 0 reduces variance and improves the expected fit for small sample size.

Even for $n \ll p$, LASSO can obtain an accurate $\vec{\beta}$ under the right conditions: the effects vector must be sparse and the heritability of the trait must be sufficiently high (equivalently, the amount of noise variance is bounded). For fixed σ_e^2 and sparse effects vector, there is a critical sample size n^* (depending on σ_e and the sparsity of the trait) above which one expects to get good recovery of $\vec{\beta}$ in terms of the L$_2$ error. A phase transition at $n \sim n^*$ has been demonstrated numerically for real and simulated genomic data in [9].

For our specific calculations we follow the following cross-validation procedure:

1. Break the data into training sets, and smaller test and validation sets.

2. Perform a standard GWAS on the training sets, and rank the SNPs by p-value.

3. To ease the computational burden, restrict the calculation to a fixed number of lowest p-value SNPs on each training set. Replace any missing SNP values by the SNP-mean for the training data.

4. Perform LASSO on the standardized training data, scanning a range of values for the penalty λ that passes through the phase transition region of rapid variation in results.

5. Choose the λ that has the maximum correlation on the test set, which was held back from training.

6. Finally, evaluate performance of optimal predictor β^* on validation sets.

A.4 Coordinate Descent

Most algorithms for minimizing the objective function (3) use (some variation of) coordinate descent [23, 24].[1] The basic form of the algorithm is as follows. Proceeding from an initial guess $\vec{\beta}_0$, we cycle through the p "coordinates" sequentially, minimizing O with respect to each β_j (holding others fixed). To that end, note that

$$\frac{\partial O}{\partial \beta_j} = n\left[\langle x_j^2\rangle\beta_j + \sum_{k\neq j}\langle x_j x_k\rangle\beta_k - \langle x_j y\rangle + \lambda\,\mathrm{sgn}(\beta_j)\right] = 0. \tag{4}$$

Thus, the updated coefficient should satisfy

$$\beta_j^* = \frac{1}{\langle x_j^2\rangle}\left[\langle x_j y\rangle - \sum_{k\neq j}\langle x_j x_k\rangle\beta_k - \lambda\,\mathrm{sgn}(\beta_j^*)\right]. \tag{5}$$

To solve for β_j^*, one should determine the $\lambda = 0$ solution. If it is positive (negative), then guess that $\mathrm{sgn}(\beta_j^*)$ should be positive (negative) and subtract (add) the λ term. If the sign flips, then the solution is spurious, and the optimal solution is at $\beta_j^* = 0$. (To see this note that for $\beta_j^* = 0^+$ the derivative is positive, and for $\beta_j^* = 0^-$ the derivative is negative.)

Introduce the "soft thresholding function"

$$S(z,\gamma) = \mathrm{sgn}(z)\max(|z| - \gamma, 0). \tag{6}$$

Then, the update for the jth component of $\vec{\beta}$ is

$$\beta_j^* = \frac{1}{\langle x_j^2\rangle} S\left(\langle x_j y\rangle - \sum_{k\neq j}\langle x_j x_k\rangle\beta_k,\ \lambda\right). \tag{7}$$

The basic Coordinate descent algorithm is as shown in Alg. 1.

A.5 Out-of-sample Validation

Model (i.e., predictor) construction was performed by implementing LASSO on the UK Biobank data. In order to validate models and check against overtraining, a second dataset is needed in order to test the results. We 1) withheld a small subset of UKBB individuals from the initial training for in-sample validation, and 2) applied the model to individuals from a completely different dataset (ARIC) for out-of-sample validation. In-sample validation was done by withholding a predetermined number of randomly selected individuals from the UK Biobank data before p-value cuts were applied to SNPs. The remaining individuals were used for LASSO training and the resulting model was applied to the individuals initially held back to check in-sample validity.

Out-of-sample validation is similar, except that we used a set of common SNPs for which state values can be imputed on the UKBB individuals and are also known for ARIC individuals. Initial

[1] We use a custom implementation in Julia [25] using safe screening ideas [26–29].

```
Data: $X_{jI}$ and $y_I$ with $j = 1, \ldots, p$ and $I = 1, \ldots, n$
Input: Penalty parameter $\lambda$, tolerance $\epsilon$, and (optionally) initial guess $\vec{\beta}_0$
Output: $\vec{\beta}$ solving LASSO optimization problem within convergence tolerance $\epsilon$
$\vec{\beta} \leftarrow \vec{\beta}_0$
repeat
    $\vec{\beta}_0 \leftarrow \vec{\beta}$
    for $j$ in $\{1, \ldots, p\}$ do
        $\beta_j \leftarrow \frac{1}{\langle x_j^2 \rangle} S\left(\langle x_j y \rangle - \sum_{k \neq j} \langle x_j x_k \rangle \beta_k,\, \lambda\right)$
    end
until $(\vec{\beta} - \vec{\beta}_0)^2 < \epsilon^2$
return $\vec{\beta}$
```

Algorithm 1: Basic coordinate descent algorithm for LASSO.

training of the model was performed using UKBB individuals, but its validity was then tested on the ARIC data. Results using the un-imputed dataset reached correlation of ~0.61 when testing within the UKBB. After selecting SNPs in common with ARIC, the correlation fell to ~0.58 while testing within the UKBB and achieved a correlation of ~0.54 on ARIC participants. The ARIC results are shown in Figure (6). Actual heights of most individuals in the ARIC validation set are within 4 cm or less of the predicted height.

The ARIC dataset [18] was composed of 12,772 caucasian and African-American individuals who were genotyped on the Affymetrix 6.0 chip with 841,820 SNPs. This was filtered to keep only caucasian individuals and SNPs with MAF larger than 1% and missing call rates below 5% with a final sample size of 9618 individuals with 705,956 SNPs. After filtering to only SNPs which were in common with the UKBB imputed data, the number of SNPs was reduced to 632,155.

We compare our activated predictor SNPs to known hits from GWAS collaborations studying the same phenotypes [3–6]. Specifically we compare our results for height with those of the GIANT collaboration, for educational attainment with SSGAC, and for Bone Density with GEFOS. We ordered activated SNPs (i.e., those assigned non-zero effect size β by the LASSO algorithm) by variance explained ($V_i = 2\nu(1-\nu)\beta_i^2$ where ν is the minor allele frequency), then scanned down this list and looked for a proxy match by distance in the corresponding dataset. For GIANT, we took the results published online and extracted the top 3000 hits ordered by p-value. For SSGAC, we used the published results and kept SNPs with $p < 10^{-6}$ - a total of 316. For GEFOS, we kept all SNPs with $p < 10^{-8}$ and then coarse grained SNPs within blocks of 10k base pairs, resulting in 3901 regions. These results are displayed in Figures (7), (8), (9). They show significant overlap between regions of the genome near previously known SNPs and regions identified by our algorithm. However, our activated SNPs are roughly uniformly distributed over the entire genome, and number in the many thousands for each trait. This means that many of our SNPs, including some of those that account for the most variance, are in regions not previously identified by earlier GWAS.

Figure (10) shows number of activated SNPs by *sign of effect of the minor allele* and minor allele frequency (MAF). The height of each bar represents the number of + or − SNPs in a MAF bin of width 0.005. The specific height predictor from which these SNPs are taken was built from 50k candidate SNPs and achieves a correlation between actual and predicted height of ~0.61. The curves, which are meant to aid visualization, are constructed by fitting a power law $n(\nu) = a\nu^{-b}$ to the range $\nu \in (0.025, 0.3)$ where ν is MAF and $n(\nu)$ is the number of nonzero effects. We exclude the smallest values of MAF because of incomplete discovery of SNPs in that region. The ± distributions are nearly symmetrical ($a_+ = 31.07, b_+ = 0.6553$; $a_- = 31.96; b_- = 0.6404$), even at very small MAF. There

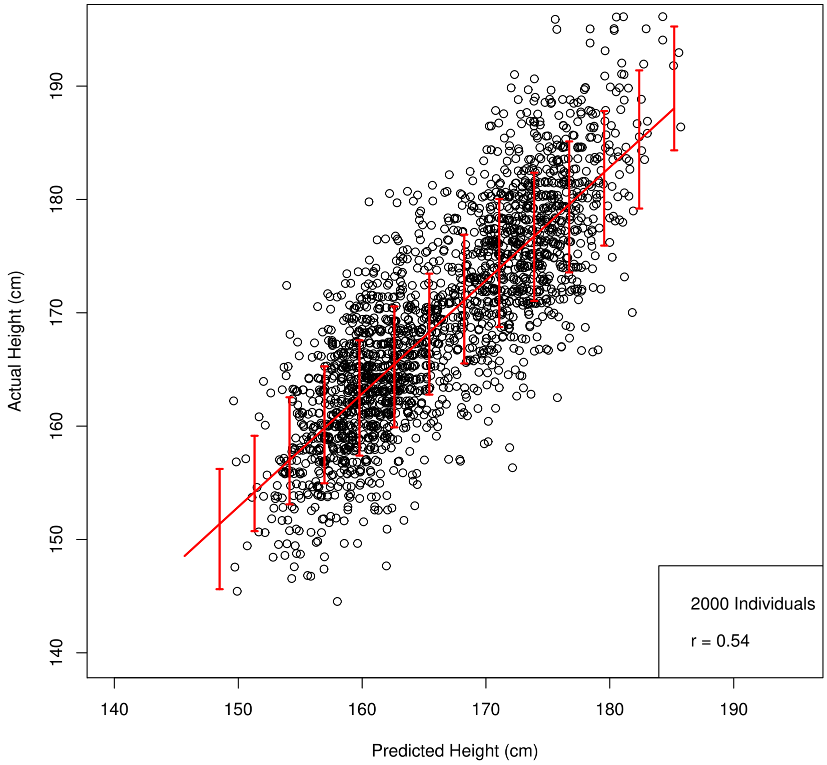

Figure 6: Actual height (cm) versus predicted height (cm) using 2000 randomly selected individuals (roughly equal numbers of M and F; no corrections for age or gender) from the ARIC dataset. Error bars indicate ±1 SD range computed using larger validation set.

does not appear to be a statistically significant deviation from random assignment of signs – the minor allele of an activated SNP is just equally likely to increase or decrease height.

References

[1] Jian Yang et al. “GCTA: A Tool for Genome-wide Complex Trait Analysis”. In: *The American Journal of Human Genetics* 88.1 (Jan. 2011), pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011. URL: https://doi.org/10.1016/j.ajhg.2010.11.011 (cit. on p. 2).

[2] Peter M. Visscher et al. “10 Years of GWAS Discovery: Biology, Function, and Translation”. In: *The American Journal of Human Genetics* 101.1 (July 2017), pp. 5–22. DOI: 10.1016/j.ajhg.2017.06.005. URL: https://doi.org/10.1016/j.ajhg.2017.06.005 (cit. on p. 2).

[3] Eirini Marouli et al. “Rare and low-frequency coding variants alter human adult height”. In: *Nature* 542.7640 (Feb. 2017), pp. 186–190. DOI: 10.1038/nature21039. URL: https://doi.org/10.1038/nature21039 (cit. on pp. 2, 12).

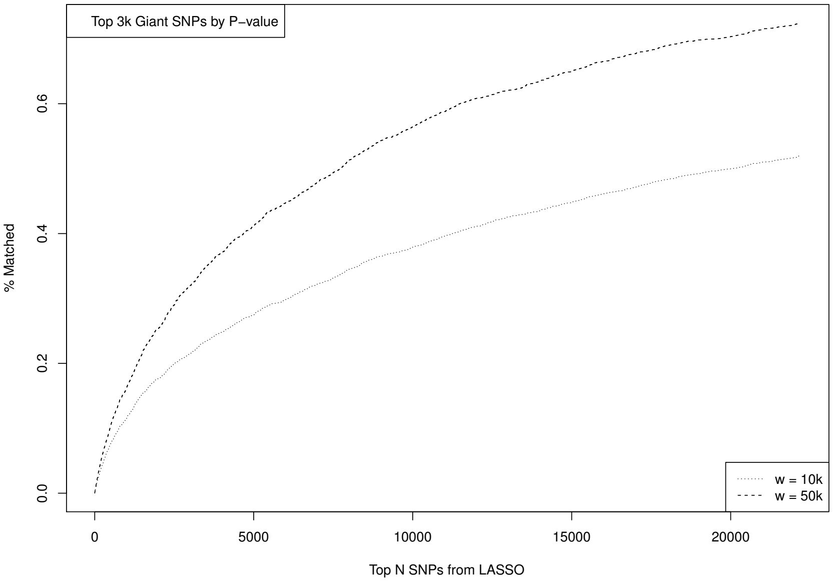

Figure 7: Matching between top SNPs activated in predictor model ordered by variance accounted for (x axis) and SNPs identified previously by GIANT GWAS (height). Percent of previously known SNPs matched is shown on y axis. Matching window size *w* given in bp.

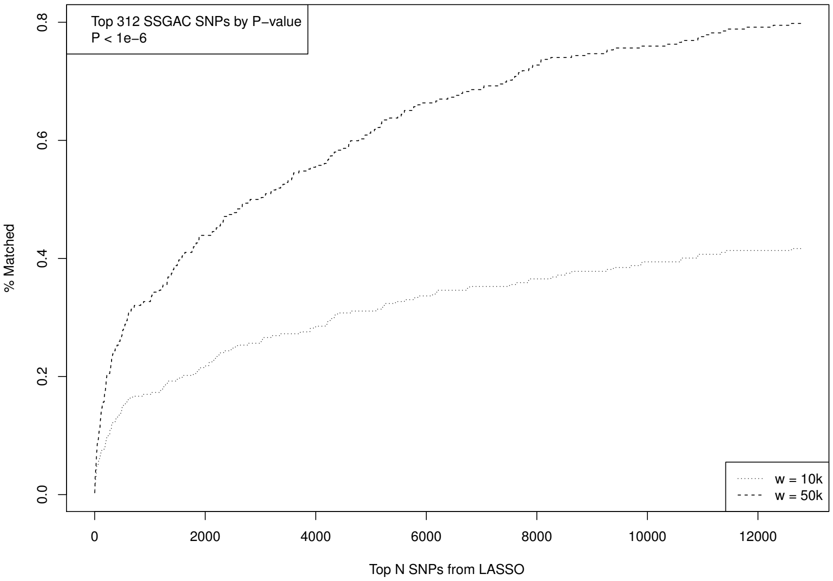

Figure 8: Matching between top SNPs activated in predictor model ordered by variance accounted for (x axis) and SNPs identified previously by SSGAC GWAS (Educational Attainment). Percent of previously known SNPs matched is shown on y axis. Matching window size *w* given in bp.

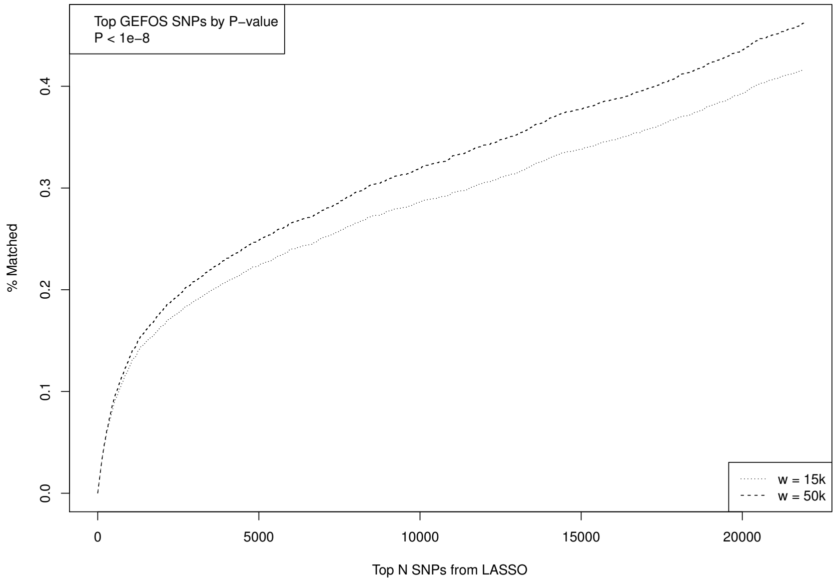

Figure 9: Matching between top SNPs activated in predictor model ordered by variance accounted for (x axis) and SNPs identified previously by GEFOS GWAS (Heel Bone Density). Percent of previously known SNPs matched is shown on y axis. Matching window size *w* given in bp.

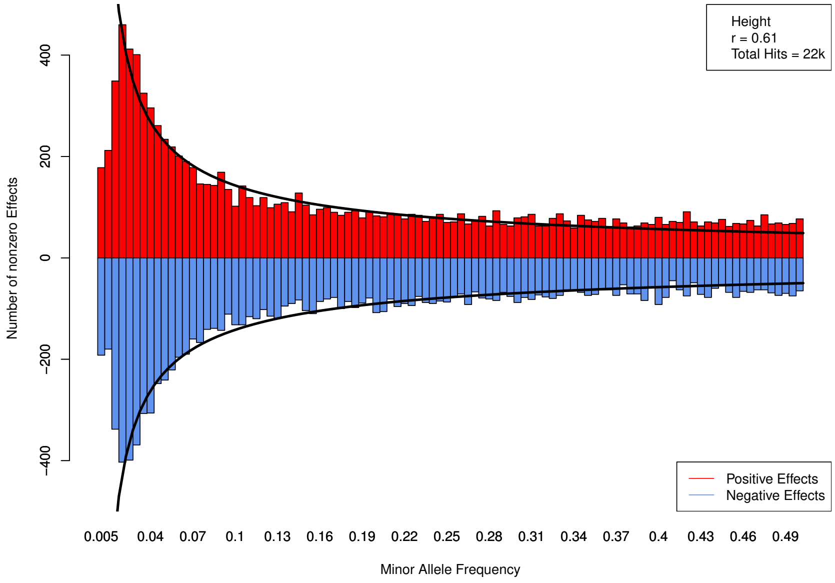

Figure 10: Number of SNPs with positive (red) and negative (blue) minor allele effect sizes. Curves are constructed by fitting a power law in MAF.

[4] Unnur Styrkarsdottir et al. “Multiple Genetic Loci for Bone Mineral Density and Fractures”. In: *New England Journal of Medicine* 358.22 (May 2008), pp. 2355–2365. DOI: 10.1056/nejmoa0801197. URL: https://doi.org/10.1056/nejmoa0801197 (cit. on pp. 2, 12).

[5] John A Morris et al. “Genome-Wide Association Study of Heel Bone Mineral Density Identifies 153 Novel Loci and Implicates Functional Involvement of GPC6 in Osteoporosis”. In: (). To appear in Nature Genetics (cit. on pp. 2, 12).

[6] Aysu Okbay et al. “Genome-wide association study identifies 74 loci associated with educational attainment”. In: *Nature* 533.7604 (May 2016), pp. 539–542. DOI: 10.1038/nature17671. URL: https://doi.org/10.1038/nature17671 (cit. on pp. 2, 12).

[7] Gustavo de los Campos et al. “Genomic Heritability: What Is It?” In: *PLOS Genetics* 11.5 (May 2015). Ed. by Gregory S. Barsh, e1005048. DOI: 10.1371/journal.pgen.1005048. URL: https://doi.org/10.1371/journal.pgen.1005048 (cit. on pp. 2, 4).

[8] Gustavo de los Campos et al. “Predicting genetic predisposition in humans: the promise of whole-genome markers”. In: *Nature Reviews Genetics* 11.12 (Nov. 2010), pp. 880–886. DOI: 10.1038/nrg2898. URL: https://doi.org/10.1038/nrg2898 (cit. on p. 2).

[9] Shashaank Vattikuti et al. “Applying compressed sensing to genome-wide association studies”. In: *GigaScience* 3.1 (2014), p. 10. ISSN: 2047-217X. DOI: 10.1186/2047-217X-3-10. URL: http://dx.doi.org/10.1186/2047-217X-3-10 (cit. on pp. 2, 7, 10).

[10] *UK Biobank*. Accessed: 2017-07-21. URL: http://www.ukbiobank.ac.uk/ (cit. on pp. 2, 9).

[11] Clare Bycroft et al. “Genome-wide genetic data on ~500, 000 UK Biobank participants”. In: (July 2017). DOI: 10.1101/166298. URL: https://doi.org/10.1101/166298 (cit. on pp. 2, 9).

[12] Robert Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. In: *Journal of the Royal Statistical Society, Series B* 58 (1994), pp. 267–288 (cit. on pp. 2, 10).

[13] Siddharth Krishna Kumar et al. “Limitations of GCTA as a solution to the missing heritability problem”. In: *Proceedings of the National Academy of Sciences* 113.1 (Dec. 2015), E61–E70. DOI: 10.1073/pnas.1520109113. URL: https://doi.org/10.1073/pnas.1520109113 (cit. on p. 4).

[14] Jian Yang et al. “GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs”. In: *Proceedings of the National Academy of Sciences* 113.32 (July 2016), E4579–E4580. DOI: 10.1073/pnas.1602743113. URL: https://doi.org/10.1073/pnas.1602743113 (cit. on p. 4).

[15] Siddharth Krishna Kumar et al. “Response to Commentary on "Limitations of GCTA as a solution to the missing heritability problem"”. In: (Feb. 2016). DOI: 10.1101/039594. URL: https://doi.org/10.1101/039594 (cit. on p. 4).

[16] Eric R Gamazon and Danny S Park. “SNP-based heritability estimation: measurement noise, population stratification, and stability”. In: (Feb. 2016). DOI: 10.1101/040055. URL: https://doi.org/10.1101/040055 (cit. on p. 4).

[17] Robert Makowsky et al. “Beyond Missing Heritability: Prediction of Complex Traits”. In: *PLoS Genetics* 7.4 (Apr. 2011). Ed. by Greg Gibson, e1002051. DOI: 10.1371/journal.pgen.1002051. URL: https://doi.org/10.1371/journal.pgen.1002051 (cit. on pp. 4, 7).

[18] “The decline of ischaemic heart disease mortality in the ARIC study communities. The ARIC Study Investigators”. In: *Int J Epidemiol* 18.3 Suppl 1 (1989), pp. 88–98 (cit. on pp. 4, 9, 12).

[19] Hwasoon Kim et al. "Will Big Data Close the Missing Heritability Gap?" In: (2017). To appear in Genetics. (cit. on p. 7).

[20] URL: https://www.snpedia.com/index.php/Heritability (cit. on p. 7).

[21] Shaun Purcell et al. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. DOI: 10.1086/519795. URL: https://doi.org/10.1086/519795 (cit. on p. 9).

[22] Chiu Man Ho and Stephen DH Hsu. "Determination of nonlinear genetic architecture using compressed sensing". In: *GigaScience* 4.1 (Sept. 2015). DOI: 10.1186/s13742-015-0081-6. URL: https://doi.org/10.1186/s13742-015-0081-6 (cit. on p. 10).

[23] Jerome Friedman et al. "Pathwise coordinate optimization". In: *The Annals of Applied Statistics* 1.2 (Dec. 2007), pp. 302–332. DOI: 10.1214/07-aoas131. URL: https://doi.org/10.1214/07-aoas131 (cit. on p. 11).

[24] Jerome Friedman et al. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010). DOI: 10.18637/jss.v033.i01. URL: https://doi.org/10.18637/jss.v033.i01 (cit. on p. 11).

[25] J. Bezanson et al. "Julia: A Fast Dynamic Language for Technical Computing". In: *ArXiv e-prints* (Sept. 2012). arXiv: 1209.5145 [cs.PL] (cit. on p. 11).

[26] L. El Ghaoui et al. "Safe Feature Elimination in Sparse Supervised Learning". In: *Pacific Journal of Optimization* 8 (4 Jan. 2012), pp. 667–698 (cit. on p. 11).

[27] Jun Liu et al. "Safe Screening with Variational Inequalities and Its Application to Lasso". In: *Proceedings of The 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. 1. JMLR Workshop and Conference Proceedings, Jan. 2014, pp. 289–297 (cit. on p. 11).

[28] O. Fercoq et al. "Mind the duality gap: safer rules for the Lasso". In: *ArXiv e-prints* (May 2015). arXiv: 1505.03410 [stat.ML] (cit. on p. 11).

[29] A. Malti and C. Herzet. "Safe screening tests for LASSO based on firmly non-expansiveness". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2016. DOI: 10.1109/icassp.2016.7472575. URL: https://doi.org/10.1109/icassp.2016.7472575 (cit. on p. 11).