Linear Mixed Models and Prediction

Hae Kyung Im, PhD

THE UNIVERSITY OF CHICAGO

February 15, 2021

LD Score Regression

$$E[\chi^2|l_j] = Nh^2 l_j/M + Na + 1$$

heritability

confounders (pop. structure)

$$E[z_{1j}z_{2}\ell_j] = \frac{\sqrt{N_1N_2}\,\varrho_g}{M}\ell_j + \frac{\varrho N_s}{\sqrt{N_1N_2}}$$

genetic correlation

2

Remember LD score regression allows us to decompose inflation into polygenic and confounder contributions. By running a regression of the chi2 statistics on ld scores, we can estimate heritability, effect of confounders (including population stratification and relatedness). By regressing the product of the zscores of two traits, one can also estimate the genetic correlation. Later, you will also see that by partitioning the ld scores into different functional categories, one can also partition the heritability of traits by functional category.

GWAS: Simple Linear Regression

In a GWAS we find one SNP at a time

$$\mathbf{Y} = \mu + a \cdot \mathbf{age} + \beta_1 \cdot \mathbf{X} + \epsilon$$

Find μ, a, β that minimizes squared error.
These are fixed parameters.

$$||\mathbf{Y} - \mu - a \cdot \mathbf{age} - \beta_1 \cdot \mathbf{X}||^2$$

3

Last class we saw that we can correct for population stratification by adding a random effects term to the linear model to capture the population structure and family structure. This was implemented in EMMAX. Today we will see how that random effect (u) is connected to the sum of the effects of all the SNPs in the genome. A typical GWAS may fit the model shown in this slide, with a mean μ, some covariates such as age, and the effect of a SNP X_1 with effect size β_1 and an error term that will "absorb" what the model is not able to capture. All parameters can be estimated by maximizing the likelihood, which is equivalent to minimizing the square difference between the phenotype and the regression function (if error term is

assumed to be normally distributed). We say that we minimize the L2 norm of the error term.

Minimizing the L2 norm

$$
\begin{aligned}
&||\mathbf{Y} - \mu - a \cdot \mathbf{age} - \beta_1 \cdot \mathbf{X}||^2 \\
&\quad = (y_1 - \mu - a \cdot \text{age}_1 - \beta_1 \cdot x_1)^2 + \\
&\quad\quad (y_2 - \mu - a \cdot \text{age}_2 - \beta_1 \cdot x_2)^2 + \\
&\quad\quad\quad \cdots \quad + \\
&\quad\quad (y_n - \mu - a \cdot \text{age}_n - \beta_1 \cdot x_n)^2
\end{aligned}
$$

Here we spell out what we mean by L2 norm.

Mixed Effects Modeling

Can we fit all SNPs at the same time?

$$Y = \mu + a \text{ age} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{1,000,000} X_{1,000,000}$$

Why can't we estimate betas by least squares?

Mixed Effects Modeling

Can we fit all SNPs at the same time?

$$Y = \mu + a \text{ age} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{1,000,000} X_{1,000,000}$$

Why can't we estimate betas by least squares?

Too many parameters and too few observations

Mixed Effects Modeling

Can we fit all SNPs at the same time?

$$Y = \mu + a \text{ age} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{1,000,000} X_{1,000,000}$$

Why can't we estimate betas by least squares?

Too many parameters and too few observations

Solution

Assume $\beta \sim N(0, \sigma_\beta^2)$ and estimate just the σ_β^2

It is reasonable to think that we could fit all the SNPs at the same time. The problem we encounter when we try to do that is that there are too many parameters and not enough data points. We typically have millions of SNPs and only thousands of individuals. Even with sample sizes growing the estimates would be overfitting the data and not work very well in new individuals. So instead of fitting millions of β's as fixed effects, we can consider them to be random and estimate their distribution, i.e. consider
β to be normally distributed with mean 0 and variance σ^2 and only estimate the variance parameter σ^2 .

Mixed Effects Modeling

$$\begin{aligned} Y &= \text{fixed effects} + \text{random effects} + \text{noise} \\ &= \text{fixed effects} + \sum \beta_k X_k + \epsilon \end{aligned}$$

$\beta_k' s$ are random

$$\beta_k \sim N(0, \sigma_\beta^2)$$

** this is one form of Regularization, more on this later

Connection to EMMAX Used To Account for Population Structure?

$$Y = \text{fixed effects} + \sum \beta_k X_k + \epsilon$$

Recall EMMAX

- $Y = x_{test} \cdot \beta_{test} + u + \varepsilon$
- $u \sim N(0, \sigma^2 \cdot \mathbf{K})$

$$Y = X_{\text{test}} \cdot \beta_{\text{test}} + \underbrace{\sum_k X_k \beta_k}_{u} + \epsilon$$

9

Here we can see the connection between the EMMAX' random effect u and the sum of the effects of all the snps. To demonstrate that Emmax random effect is the same as the sum of the effects of all the snps, all we need to do is to shown that they have the same covariance matrix, also equal to the genetic relatedness matrix.

Calculate K

Kernel
Similarity matrix
Genetic Relatedness Matrix

10

Leave One Chromosome Out

$$Y = X_{\text{test}} \cdot \beta_{\text{test}} + \underbrace{\sum_k X_k \beta_k}_{u} + \epsilon$$

Initially, EMMAX was calculating K using all SNPs

Issue: deflation due to proximal contamination
Solution: LOCO, leave one chromosome out

11

The equivalence between the random effect u and the sum of the effects of all the snps provide an explanation to the deflation seen in EMMAX results.

The effect of a snp is being explained by both the fixed effects and the random effect, so the power to detect the effect is diluted and absorbed by the random effect component. LOCO (leave one chromosome out) is an easy solution to this problem, for each test SNP, only use the variants outside of the chromosome where theist SNP is located, ensuring that there will be no LD between SNPs that make up u and the test SNP.

Biobank-Scale Ready LMM Methods

- EMMAX
  - original mixed effects modeling proposal to correct for population and family structure
  - $Y = x_{test} \cdot \beta_{test} + u + \varepsilon$
  - $u \sim N(0, \sigma^2 \cdot \boldsymbol{K})$

Example $\boldsymbol{K}$

- BOLT-LMM (Loh et al, 2015, 2018)
  - $Y = x_{test} \cdot \beta_{test} + u + \varepsilon$
  - $u \sim \pi \cdot N(0, \sigma^2_{small} \cdot \boldsymbol{K}) + (1 - \pi) \cdot N(0, \sigma^2_{large} \cdot \boldsymbol{K})$
- fastGWA (Jiang et al 2019)
  - $Y = x_{test} \cdot \beta_{test} + PC \cdot \beta_{pc} + u + \varepsilon$
  - $u \sim N(0, \sigma^2 \cdot \boldsymbol{K})$, $\boldsymbol{K}$ only kinship by rounding off elements < 0.05

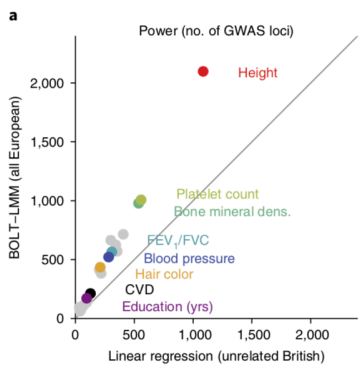
Kang et al (2010). Variance component model to account for sample structure in genome-wide association studies. Nature Genetics.

Loh et al (2018). Mixed-model association for biobank-scale datasets. Nature Genetics.

Jiang et al (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nature Genetics.

The main reason mixed effects models were not adopted more broadly is the computational cost. For example, most publications using the UK Biobank data use only unrelated individuals which means going from a sample size of 450K down to ~330k.
To address this problem, several biobank-scale ready methods that reduce the computational burden have been published. Two prominent ones are BOLT-LMM and fastGWA.

BOLT LMM Power Gain in UK Biobank



459,327 all British
vs.
337K unrelated

BOLT-LMM

$Y = x_{test} \cdot \beta_{test} + u + \varepsilon$

increased sample size + adjustment of polygenic component

Loh et al (2018). Mixed-model association for biobank-scale datasets. Nature Genetics.

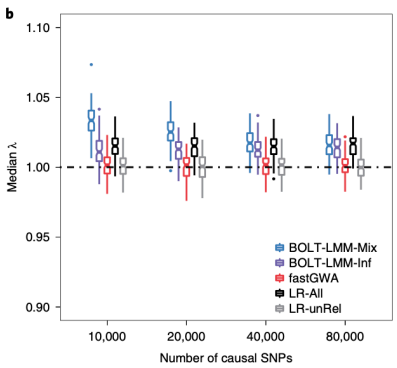
BOLT-LMM fits the SNP of interest, X_{test}, with the random effects term computed LOCO.

To gain computational speed, BOLT-LMM fits the model without the test SNP, i.e.the null model for the test SNP. This is done once per each chromosome.

Then the residual $Y - u$ to test whether X_{test} is associated with the Y.

The authors claim that by doing this they not only gain by sample size increase due to including the related individual, but also by adjusting for the polygenic component, which is captured by the u. In this figure boltlmm is shown to increase the number of discoveries by more than 80%.

fastGWA's Simulation Show Inflation of BOLT-LMM



Jiang et al (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nature Genetics.

fastGWA is another method capable of handling biobank scale GWAS with complex population and relatedness structure.

fastGWA adjusts for population structure using the more traditional approach of using genetic principal components as covariates. A random effect is used to adjust for relatedness by using a "rounded" version of the relatedness matrix, where all values below 0.05 are replaced by 0, yielding the kinship matrix.

They suggest that BOLT-LMM's increased discoveries may be due to inflation rather that its ability to leverage polygenicity.

fastGWA is FAST and Memory Efficient

Table 1 | Comparison of runtimes of fastGWA, BOLT-LMM, and PLINK2

Sample size	GCTA-fastGWA			BOLT-LMM			PLINK2	$\frac{Mem_{fastGWA}}{Mem_{BOLT-LMM}}$	$\frac{VMem_{fastGW}}{VMem_{BOLT-L}}$
	Para. est. (h)	Assoc. (h)	Total (h)	Para. est. (h)	Assoc. (h)	Total (h)	Total (h)		
50,000	0.00	0.03	0.03	0.88	1.05	1.93	0.07	15.9%	34.0%
100,000	0.00	0.04	0.04	2.09	2.07	4.16	0.15	10.5%	20.4%
200,000	0.01	0.07	0.08	5.34	4.16	9.50	0.37	6.3%	11.5%
300,000	0.01	0.14	0.15	9.51	6.24	15.75	0.81	5.2%	10.0%
400,000	0.02	0.23	0.25	13.85	8.44	22.29	1.15	4.9%	8.3%

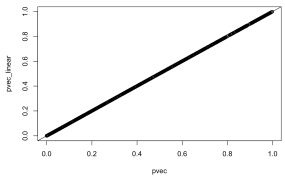
Jiang et al (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nature Genetics.

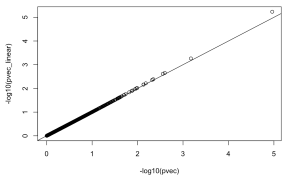
fastGWA is remarkably fast, even faster than plink2, which only corrects for population structure with genetic PCs but not for relatedness, so no random effect.

Linear Approximation to Logistic Regression

works well with balanced case-control designs

recall the homework problem when we simulated case control and compared linear vs. logistic regression





boltLMM and fastGWA assume linearity of the trait, even for diseases. This is justified for balanced studies, where there are similar numbers of cases and controls. But in the UK Biobank, a cohort study, not selected by diseases status, can have highly unbalanced ratios of cases and controls. In extreme cases, this unbalance can go very wrong. SAIGE next, address this problem.

Generalized Mixed Models for Unbalanced Studies

- SAIGE
  - Scalable and Accurate Implementation of GEneralized mixed model
  - unbalanced case control studies
  - log(p / (1-p)) = xtest · βtest + u + ε

Zhou, W. et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics, 50(9), 1–12.

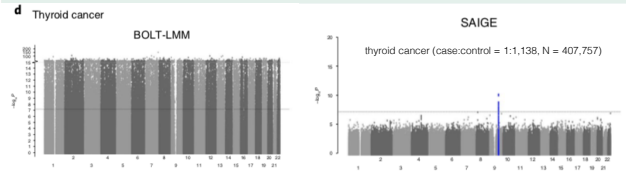
to speed up computation, BOLT-LMM and fastGWA use linear regression, which is a good approximation to logistic regression when the proportion of cases and controls are similar. With unbalanced studies with much smaller number of cases relative to controls or vice-versa, the approximation starts to fail and logistic regression must be used. Logistic mixed effects models can be difficult to deal with but Zhou et al developed a method that addresses the problems and yields a calibrated method.

For Unbalanced Case/Control Studies: SAIGE

Table 1 | Comparison of different methods for genome-wide association studies with mixed effect models

		Method features					Algorithm complexity				Benchmarks for UK Biobank data coronary artery disease (PheCode 411)	
		Does not require a precomputed genetic relationship matrix	Feasible for large sample sizes	Developed for binary traits	Accounts for unbalanced case-control ratio	Tests quantitative traits	Time complexity		Memory usage (GB)		Time CPU hours	Memory
							Step 1	Step 2	Step 1	Step 2		
Logistic mixed model	SAIGE	✓	✓	✓	✓	✓	$O(PM_rN^{1.5})^a$	$O(MN)$	$M_rN/4$	N	517	10.3 GB
	GMMAT			✓		✓	$O(PN^2)$	$O(MN^2)$	FN^2	FN^2	NA	NA
Linear mixed model	BOLT-LMM	✓	✓			✓	$O(PM_rN^{1.5})^a$	$O(MN)$	$M_rN/4$	N	360	10.9 GB
	GEMMA					✓	$O(N^3)$	$O(MN^2)$	FN^2	FN^2	NA	NA

N, number of samples; P, number of iterations required to reach convergence; M_r, number of markers used to construct the kinship matrix; M, total number of markers to be tested; F, byte for floating number. CPU, central processing unit; NA, not applicable. [a]Number of iterations in PCG is assumed as $O(N^{0.5})$[1].



In the UKB there are only 1138 cases of thyroid cancer vs 407K controls. BOLT-LMM would yield a badly inflated association as shown in the figure bottom left. SAIGE's results look much better calibrated.

Prediction of Complex Traits

Prediction of complex traits can help us better taylor treatment of patients.

Simple Polygenic Score

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\mathrm{GWAS}} X_k$$

Just use GWAS effect sizes

20

Polygenic risk scores are simple to calculate with unexpectedly good prediction performance.

Best Linear Unbiased Prediction (BLUP)/Ridge

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

AJHG 2011

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\mathrm{Ridge}} X_k$$

Penalized regression

Ridge

$$\|Y - \sum_k X_k \beta_k\|_2 + \lambda_2 \|\beta_2\|_2$$

21

More sophisticated methods will use betas estimated jointly. As we discussed earlier, to make this work we can use a random effects approach. This can be shown to be equivalent to using a penalized likelihood, also known as regularization. Ridge regression approach minimizes the likelihood with a penalty on the L2 norm of the effect size vector, i.e. it tries to minimize the mean square error while still keeping the length of the effect size vector small.

LASSO/Elastic Net Prediction

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie

Stanford University, USA

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\mathrm{E\text{-}N}} X_k$$

Penalized regression

LASSO

Elastic Net

$$\|Y - \sum_k X_k \beta_k\|_2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta_2\|_2$$

22

LASSO penalizes sum of the absolute values of the effect sizes, i,e,the L1 norm of the effect size vector. These tend to yield sparse models, a few SNPs rather than polygenic models.

Elastic net mixes both L1 and L2 norms yielding less sparse models, although not quite polygenic ones.

Whole Genome Prediction Approaches

OPEN ACCESS Freely available online PLOS | GENETICS

Polygenic Modeling with Bayesian Sparse Linear Mixed Models

Xiang Zhou[1]*, Peter Carbonetto[1], Matthew Stephens[1,2]*

$$Y = \sum_{k=1}^{M} \beta_k^L X_k + \sum_{k=1}^{M} \beta_k^S X_k + \epsilon$$

$$\beta_k^L \sim N(0, \sigma_L^2)$$

$$\beta_k^S \sim N(0, \sigma_S^2)$$

MultiBLUP: improved SNP-based prediction for complex traits

Doug Speed and David J Balding

Genome Res. published online June 24, 2014
Access the most recent version at doi:10.1101/gr.169375.113

Other approches for prediction include BSLMM, multiBLUP, OmicKriging.
BSLMM models the genetic effects as coming from a mixture of normals instead of just one normal distribution. One with small variance captures the polygenic component whereas the large variance component captures the sparse effects (a few SNPs with large effects). By selecting the right can sparsity can be enforced.

Advantages of Polygenic Scores

Main advantage easy to get or calculate, scalable

GWAS results publicly available

vs. multivariate estimates need individual data

although some fine-mapping methods allow inferring multivariate regression results

Current Methods for Improving Polygenic Scores

- Pruning and thresholding (PRSice)
- Lasso-sum (Mak et al)
- LD-Pred (Vilhjálmsson)
- RSS (Zhu)
- S-BayesR (Lloyd-Jones)
- PRS-CS

RSS and S-BayesR are likelihood-based methods, different priors on $\beta's$

Zhu, X., & Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. AOAS
Vilhjálmsson et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. AJHG
Mak,et al (2017). Polygenic scores via penalized regression on summary statistics. Genetic Epidemiology. 41(6), 469–480.
Luke R. Lloyd-Jones (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. BioRxiv.
Ge, T., Chen, CY., Ni, Y. et al. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 10, 1776 (2019). https://doi.org/10.1038/s41467-019-09718-5

Importance of Having Good LD Reference Data

All the methods listed in the previous page rely on having good LD reference data.

With increasing sample sizes, methods that use summary statistics and infer results similar to having individual level data are critical.

- Summary statistics from GWAS are being widely shared.
- LD reference from the same study is not, this is something that needs to change

Clinical Utility of Genetic Predictions

Genomic Prediction of Height in UK Biobank

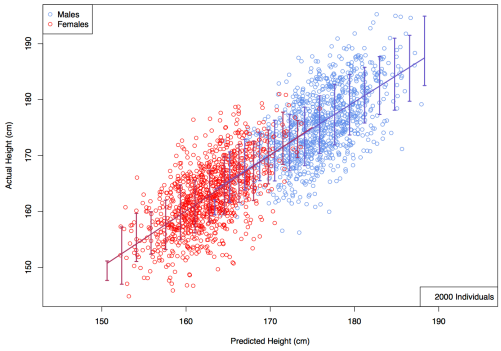
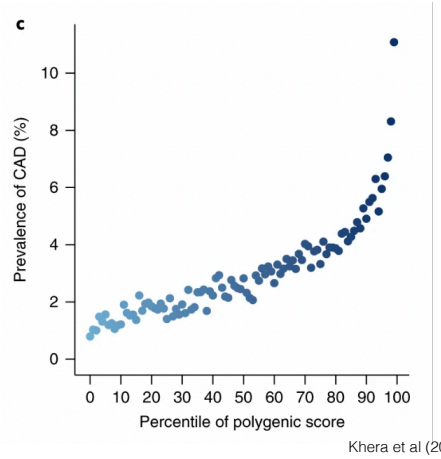


Figure 4: Actual height (cm) versus predicted height (cm) using 2000 randomly selected individuals held back from predictor optimization. Error bars indicate ±1 SD range computed using larger validation set. (Roughly equal numbers of males and females; no corrections of actual height for age or gender.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

Lello et al (2018). Accurate Genomic Prediction of Human Height. Genetics

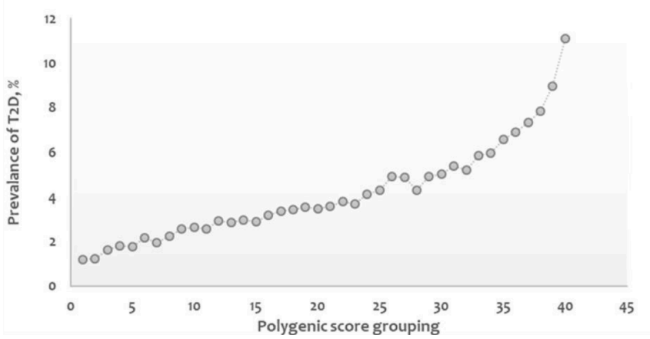
with biobank scale data, we are able to predict height quite well using common variants

Prevalence of Coronary Artery Disease Increases with PRS



Khera et al (2018) Nature Genetics

Prevalence of Type 2 Diabetes Increases with PRS



Mahajan et al (2019) Nature Genetics

Do PRS work for everyone?

Portability of Prediction Across Ancestries

PERSPECTIVE
https://doi.org/10.1038/s41588-019-0379-x
nature genetics

Clinical use of current polygenic risk scores may exacerbate health disparities

Alicia R. Martin[1,2,3]*, Masahiro Kanai[1,2,3,4,5], Yoichiro Kamatani[5,6], Yukinori Okada[5,7,8], Benjamin M. Neale[1,2,3] and Mark J. Daly[1,2,3,9]

Ancestry Composition of Current GWAS

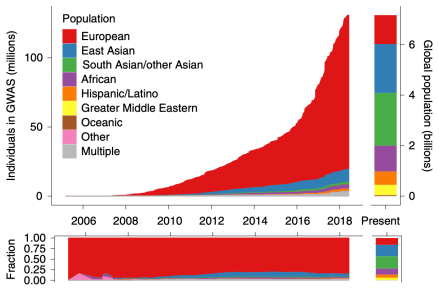


Fig. 1 | Ancestry of GWAS participants over time, as compared with the global population. Cumulative data, as reported by the GWAS catalog[76]. Individuals whose ancestry is 'not reported' are not shown.

The majority of the GWAS have been performed in individuals of European descent.

Allele Frequency and LD Differ Across Ancestries

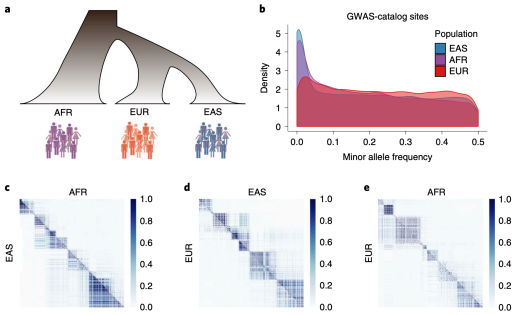


Fig. 2 | Demographic relationships, allele frequency differences and local LD patterns between population pairs. Data analyzed from 1000 Genomes. Population labels: AFR, continental African; EUR, European; EAS, East Asian. **a**, Cartoon relationships among AFR, EUR and EAS populations. **b**, Allele frequency distributions in AFR, EUR and EAS populations of variants from the GWAS catalog. **c**–**e**, Color axis shows LD scale (r^2) for the indicated LD comparisons between pairs of populations; the same region of the genome for each comparison (representative region is chromosome 1, 51572–52857 kilobases) among pairs of SNPs polymorphic in both populations is shown, illustrating that different SNPs are polymorphic across some population pairs and that these SNPs have variable LD patterns across populations.

Difference in LD are likely to make the transfer of PRS difficult.
Many of the significant variants are likely to be proxy to the causal ones. With different LD proxies will vary across population contributing the wrong value to the PRS.

PRS Does Not Transfer Well Across Populations

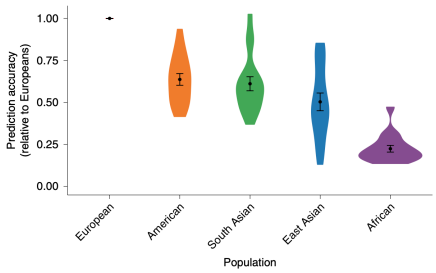


Fig. 3 | Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB. All phenotypes shown here are quantitative anthropometric and blood-panel traits, as described in Supplementary Table 6, which includes discovery-cohort sample sizes. Prediction target individuals do not overlap with the discovery cohort and are unrelated; sample sizes are shown in Supplementary Table 7. Violin plots show distributions of relative prediction accuracies, points show mean values, and error bars show s.e.m. values. Prediction R^2 for each trait and population are shown in Supplementary Fig. 12.

This loss of prediction performance is what was reported by Martin et al.

Ongoing Efforts to Diversify GWAS Studies

Genomic Privacy

Surge of Genomic Data Since First Draft of Human Genome

- New era of biomedical research massive amounts data
- Huge potential for new discoveries
- “Few blockbuster new cures” (NY Times)
- For full advantage, broad sharing of data and results is needed
- However, privacy of study participants has to be protected

Challenges in Sharing Genomic Results

- Summary statistics in large studies considered safe to publish
  - proportion of females vs. males,
  - average LDL cholesterol levels, etc.
- Genome wide association studies GWAS
  - for millions of SNPs
  - differential mutations frequencies in cases vs. controls are generated
- Frequency of mutations in cases and controls used to be publicly available

Forensic Study Revealed Vulnerability

- Forensic application - Homer et al (2008) Plos Genetics
- Effciency of new genotyping chips in forensic application
  - DNA sample from crime scene
  - DNA from suspect
  - Determine whether suspect’s DNA is part of the sample

	Id 1	Id 2	Id 3	Id 4	Sample	Popul	Suspect
SNP 1	1	2	0	0	0.75	1.10	0
SNP 2	1	0	0	1	0.50	1.25	1
	...	...	...	...	...	...	...
SNP M	1	0	1	2	1.00	1.50	2

Quantitative Trait GWAS - What Are the Risks of Sharing?

$$Y_i = \alpha_j + \beta_j X_{i,j} + e_i$$

$$\hat{\beta}_j = \left(\tilde{\mathbf{X}}_j' \tilde{\mathbf{X}}_j\right)^{-1} \tilde{\mathbf{X}}_j' \tilde{\mathbf{Y}}$$

We wanted to share the summary results but wanted mathematical proof that it would not allow re-identification of subjects.

Betas and Genotypes Are Known

$\hat{\beta}_1$	$X_{I,1}$
$\hat{\beta}_2$	$X_{I,2}$
$\vdots$	$\vdots$
$\hat{\beta}_M$	$X_{I,M}$

Average the product

$$\frac{1}{M}\sum_{j=1}^{M} \hat{\beta}_j X_{I,j}$$

Testing the Yhat Statistic in GoKinD Data

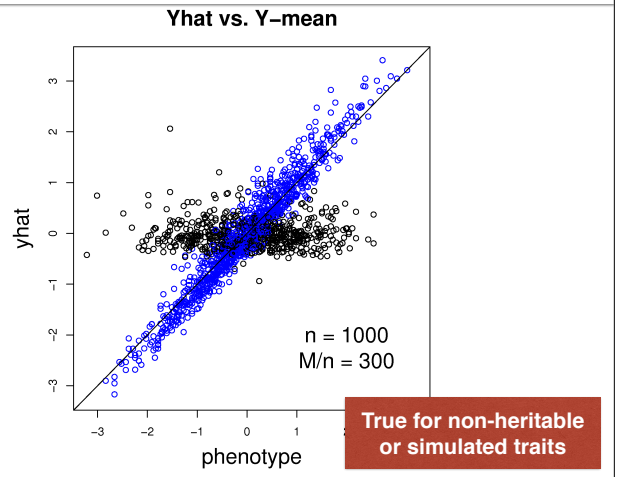
- Dataset from The Genetics of Kidneys in Diabetes
  - Study long-term Type 1 diabetes adults (GoKinD)
- Phenotype: rank normalized cholesterol level
- n = 1600
- Random sample of 1000 individuals
- 600 used as reference
- Using only the 1000 sample ran GWAS

$$\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta_M}$$

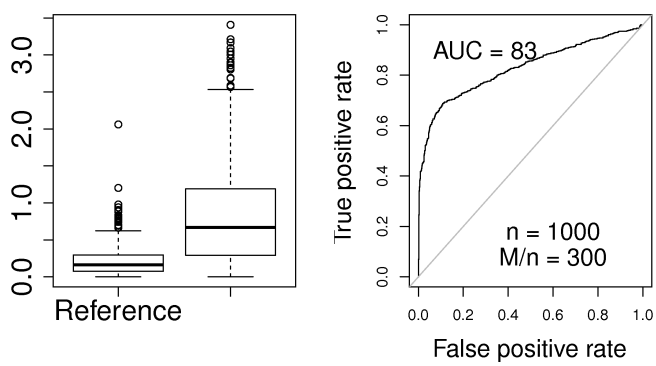
- Computed the statistic for all 1600

$$\text{Yhat}_I = \frac{1}{M}\sum_{j=1}^{M} \hat{\beta}_j X_{I,j}$$

Yhat as Predictor of Y - GoKinD data



Distribution of Yhat and Performance - GoKind data



if we use a threshold to separate the "predicted" in the study and not in the study (horizontal line on the left box plot figure), a number of individuals will be true positives and a number will be false positives, these can be used use construct the ROC curve on the right.

Yhat Statistics

$$\hat{Y}_I = \frac{n}{M}\sum_{j=1}^{M}\hat{\beta}_j(X_{I,j} - \hat{X}_j)$$

M	# of SNPs
n	# of individuals in the test sample
$X_{I,j}$	allelic dosage of individual I at SNP j
$\hat{X}_j$	estimated mean using the reference group
$\hat{\beta}_j$	estimated β for $Y_i = \alpha_j + \beta_j X_{i,j} + e_i$

Conditional Distribution of Yhat

$$\mathbb{E}\ \hat{Y} \mid X_I, Y_I, \text{in} \quad \approx \quad (Y_I - \mu)$$

$$\mathbb{E}\ \hat{Y} \mid X_I, Y_I, \text{out} \quad \approx \quad O_p\left(\frac{n}{M}\right)$$

$$\text{Var}(\hat{Y}) \mid X_I, Y_I, \text{in} \quad \approx \quad \sigma^2 \frac{n}{M}$$

$$\text{Var}(\hat{Y}) \mid X_I, Y_I, \text{out} \quad \approx \quad \sigma^2 \frac{n}{M}$$

Power of the Method

$$\text{power} \approx \Phi\left(\frac{|Y_I - \mu|}{\sigma}\sqrt{\frac{M}{n}} - z_{\alpha/2}\right)$$

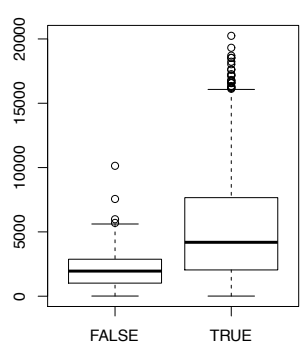
To be compared to power for binary traits

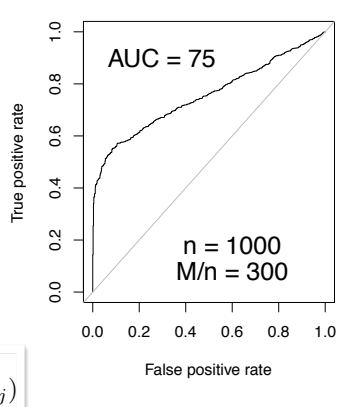
$$\text{power} \approx \Phi\left(\sqrt{\frac{M}{n}} - z_{\alpha}\right)$$

For 5% alpha, 90% power, and $Y_I = \mu + \sigma$

$$13 = \frac{M}{n}$$

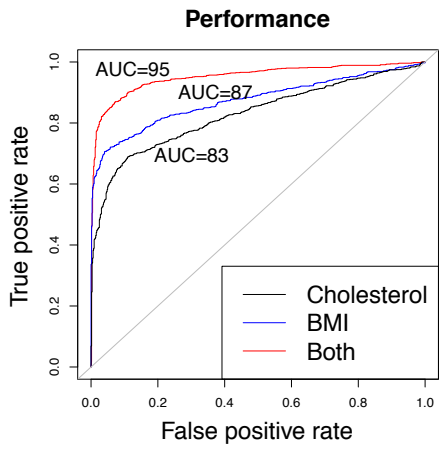
What if Only Direction of Effects is Known





$$\hat{S} = \sum_{j=1}^{M} \text{sign}(\hat{\beta})\text{sign}(X_{ij} - \hat{X}_j)$$

Performance Improves with Multiple Phenotypes



Summary of Genomic Privacy

- Showed that aggregate results from quantitative GWAS can reveal individual's participation and phenotype
- Computed power of the identification method
- Determined that the direction of effects contains most of the individual's information
- Established that identification becomes more accurate when results from multiple phenotypes are combined
- Thus, there is need to develop data sharing strategies that protect participant's privacy but also facilitate access to data
- Growing consensus now that re-identification risk is minor compared to benefit of sharing summary results