Multi-ethnic polygenic risk scores improve risk prediction in diverse populations

Carla Márquez-Luna[1], The SIGMA Type 2 Diabetes Consortium, Alkes L. Price[1,2,3]

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[2] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[3] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.

Correspondence should be addressed to C.M.L. (cmarquezluna@fas.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu).

Abstract

Methods for genetic risk prediction have been widely investigated in recent years. However, most available training data involves European samples, and it is currently unclear how to accurately predict disease risk in other populations. Previous studies have used either training data from European samples in large sample size or training data from the target population in small sample size, but not both. Here, we introduce a multi-ethnic polygenic risk score approach, MultiPRS, that combines training data from European samples and training data from the target population. We applied MultiPRS to predict type 2 diabetes in a Latino cohort using both publicly available European summary statistics in large sample size and Latino training data in small sample size, and observed a >70% relative improvement in prediction accuracy compared to methods that use only one source of training data, consistent with large relative improvements observed in simulations. Notably, this improvement is contingent on the use of ancestry-adjusted coefficients in MultiPRS. Our work reduces the gap in risk prediction accuracy between European and non-European target populations.

Introduction

Genetic risk prediction is an important and widely investigated topic because of its potential clinical application as well as its application to better understand the underlying genetic architecture of complex traits. Many polygenic risk prediction methods have been developed and applied to complex traits. These include polygenic risk scores (PRS)[1–7], which use summary association statistics as training data, and Best Linear Unbiased Predictor (BLUP) methods and their extensions[8–14], which require raw genotype/phenotype data·

However, all of these methods are inadequate for polygenic risk prediction in non-European populations, because they consider training data from only a single population. Indeed, ref. 7 reported a relative decrease of 53-89% in schizophrenia risk prediction accuracy in Japanese and African American populations compared to Europeans when applying PRS methods using European training data; this decrease is an expected consequence of different patterns of linkage disequilibrium (LD) in different populations[15]. An alternative is to use training data from the same population as the target population, but this would generally imply a lower sample size, reducing prediction accuracy.

To tackle this problem, we developed a method, MultiPRS, that combines PRS based on European training data with PRS based on training data from the target population. The method takes advantage of both the accuracy that can be achieved with large training samples[3,4] and the accuracy that can be achieved with training data containing the same LD patterns as the target population.

In simulations and application to predict type 2 diabetes (T2D) in Latino target samples in the SIGMA T2D data set[16], MultiPRS attains a >70% relative improvement in prediction accuracy compared to methods that use only one source of training data. This improvement is contingent on the use of ancestry-adjusted coefficients, as methods that use coefficients that are not adjusted for ancestry perform much worse than MultiPRS.

Methods

Overview of Methods

We explored 5 different prediction methods in our simulations and application to T2D: European training data only (EUR); Latino training data only (LAT); Latino training data and an ancestry predictor based on the proportion of European ancestry (LAT+ANC); optimal linear combination of predictions from European training data and Latino training data (EUR+LAT); and optimal linear combination of predictions from European training data, Latino training data and an ancestry predictor (MultiPRS). For each of the methods using Latino training data we used 10-fold cross-validation within the SIGMA T2D Latino data set. Latino effect sizes were adjusted for genome-wide ancestry in all primary analyses[17], but we also considered methods with unadjusted Latino effect sizes as a secondary analysis. For each method, we built predictions by LD-pruning the SNPs and then thresholding on SNPs with P-value below a threshold, a widely used approach[1–4]. We used squared correlation (R^2 on the observed scale) between predicted phenotype and phenotype as our primary measure of prediction accuracy.

Genetic Model

Let Y be the 1 x N phenotype vector with elements Y_j, and let X be the M x N genotype matrix with elements g_{ij}, where j = 1,...,N is the number of samples and i = 1,...,M is the number of genetic markers. We assume that the phenotype Y has E(Y)=0 and var(Y)=1, and that genotypes g_{ij} are mean-centered with missing data to 0. We assume that the phenotype is a linear combination of genetic and environmental effects: $Y_j = \sum_{i=1}^{M} b_i g_{ij} + \varepsilon$.

Polygenic risk score and LD-pruning + P-value thresholding.

We define the polygenic risk score[1] as $\hat{Y}_j = \sum_{i=1}^{M} \hat{b}_i g_{ij}$, where the marginal least square estimate is $\hat{b}_i = \dfrac{\sum_{j=1}^{N} g_{ij} Y_j}{\sum_{j=1}^{N} g_{ij}^2}$.

We use LD-pruning and P-value thresholding as follows. We first LD-prune the SNPs based on a pairwise threshold $R_{LD}{}^2$, and then restrict to SNPs with an association P-value below a fixed threshold P_T. We investigated different values of the pruning threshold $R_{LD}{}^2$, and ultimately fixed $R_{LD}{}^2$ at 0.8 since this value consistently performed best. We perform a grid search over different values of the P-value threshold P_T: 1.0, 0.8, 0.5, 0.4, 0.3, 0.2, 0.1, 0.08, 0.05, 0.02, 0.01, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}. We optimize P_T via an in-sample fit on validation samples, consistent with previous work[1–4]. We performed LD-pruning and P-value thresholding using PLINK2 (see Web Resources).

Prediction methods

We considered 5 different methods for prediction in Latino target samples: EUR, LAT, LAT+ANC, EUR+LAT and MultiPRS.

EUR builds predictions using estimated effect sizes from European training data, using LD-pruning and P-value thresholding as described above: $PRS_{EUR_j} = \sum_{i=1}^{M} \hat{b}_{EUR,i} g_{ij}$.

LAT builds predictions using estimated effect sizes from Latino training data, using LD-pruning and P-value thresholding: $PRS_{LAT_j} = \sum_{i=1}^{M} \hat{b}_{LAT,i} g_{ij}$. We employ 10-fold cross-validation, using 90% of the Latino samples to estimate effect sizes and the remaining 10% of the samples for validation. Latino effect sizes are estimated with adjustment for 2 PCs (except in the unadjusted case).

LAT+ANC builds predictions using the best linear combination of LAT and ancestry along the top 2 PCs (ANC): $PRS_{LAT+ANC,j} = \hat{\alpha}_1 PRS_j + \hat{\alpha}_2 PC_j$, where PC_j is a 2 x 1 vector containing the top 2 PCs. We optimize this linear combination via an in-sample fit on validation samples.

EUR+LAT builds predictions using the best linear combination of EUR and LAT: $PRS_{EUR+LAT,j} = \hat{\alpha}_1 PRS_{EUR,j} + \hat{\alpha}_2 PRS_{LAT,j}$. We optimize this linear combination via an in-sample fit on validation samples. We optimize different P-value thresholds for EUR and LAT, performing a 2-dimensional grid search.

MultiPRS builds predictions using the best linear combination of EUR, LAT, and ancestry along the top 2 PCs (ANC): $PRS_{MultiPRS,j} = \hat{\alpha}_1 PRS_{EUR,j} + \hat{\alpha}_2 PRS_{LAT,j} + \hat{\alpha}_3 PC_j$. We optimize this linear combination via an in-sample fit on validation samples. We optimize different P-value thresholds for EUR and LAT, performing a 2-dimensional grid search.

Simulations

We simulated quantitative phenotypes using real genotypes from European (WTCCC2) and Latino (SIGMA) data sets (see below). We fixed the proportion of causal markers at 1% and fixed SNP-heritability h_g^2 at 0.5, and sampled normalized effect sizes β_i from a normal distribution with variance equal to h_g^2 divided by the number of causal markers. We calculated per-allele effect sizes b_i as $b_i = \frac{\beta_i}{\sqrt{2p_i(1-p_i)}}$, where p_i is the minor allele frequency of SNP i in the European data set. We simulated phenotypes as $Y_j = \sum_{i=1}^{M} b_i g_{ij} + \varepsilon_j$, where $\varepsilon_j \sim N(0, 1-h_g^2)$.

In our primary simulations, we discarded the causal SNPs and used only the non-causal SNPs as input to the prediction methods (i.e. we simulated untyped causal SNPs, which we believe to be most realistic). We also considered simulations in which we included the causal SNPs as input to the prediction methods (typed causal SNPs) as a secondary analysis.

We also performed simulations in which Latino phenotypes were explicitly correlated to ancestry. In these simulations, we added a constant multiple of PC1 (representing European vs. Native American ancestry, with positive values representing higher European ancestry) to the Latino phenotypes such that the correlation between phenotype and PC1 was equal to -0.11, which is the correlation between T2D phenotype and PC1 in the SIGMA data set.

We performed simulations under 4 different scenarios: (i) using all chromosomes, (ii) using chromosomes 1-4, (iii) using chromosomes 1-2, and (iv) using chromosome 1 only. The motivation for performing simulations with a subset of chromosomes was to increase N/M, extrapolating to performance at larger sample sizes, as in previous work[7].

To assess the gap in risk prediction accuracy between European and Latino target populations, we also computed polygenic risk scores in Europeans via 10-fold cross-validation in the European data set, using LD-pruning and P-value thresholding.

WTCCC2, SIGMA and DIAGRAM data sets.

Our simulations used real genotypes from the WTCCC2 and SIGMA data sets. The WTCCC2 data set consists of 15,622 unrelated European samples genotyped at 360,557 SNPs after QC[18,19]. The SIGMA data set consists of 8,214 unrelated Latino samples genotyped at 2,440,134 SNPs after QC[16]. We restricted our simulations to 232,629 SNPs present in both data sets (with matched reference and variant alleles) after removing A/T and C/G SNPs due to strand ambiguity.

Our analyses of type 2 diabetes used summary association statistics from DIAGRAM data set and genotypes and phenotypes from the SIGMA data set. The DIAGRAM data set consists of 12,171 cases and 56,862 controls of European ancestry for which summary association statistics at 2,473,441 imputed SNPs are publicly available (see Web Resources)[20]. As noted above, the SIGMA data set consists of 8,214 unrelated Latino samples (3,848 type 2 diabetes cases and 4,366 controls) genotyped at 2,440,134 SNPs after QC. We restricted our analyses of type 2 diabetes to 776,374 SNPs present in both data sets (with matched reference and variant alleles) after removing A/T and C/G SNPs due to strand ambiguity. For the SIGMA data set, we used the top 2 PCs as computed in ref. [16]. We also performed an analysis of type 2 diabetes using imputed genotypes from the SIGMA T2D data set[16], restricting to 2,062,617 present in both data sets (with matched reference and variant alleles) after removing A/T and C/G SNPs due to strand ambiguity.

Results

Simulations.

We performed simulations using real genotypes and simulated phenotypes. We simulated continuous phenotypes under a non-infinitesimal model with 1% of markers chosen to be causal and SNP-heritability $h_g^2 = 0.5$ (see Methods); we report the average R^2 and standard deviation over 100 simulations. We used WTCCC2[18,19] data (15,622 samples after QC, see Methods) as the European training data, and the SIGMA data[16] (8,214 samples) as the Latino training and validation data (with 10-fold cross-validation). We simulated phenotypes using the 232,629 SNPs present in both data sets and built predictions from these SNPs excluding the causal SNPs, modeling the causal SNPs as untyped (see Methods).

Prediction accuracies (R^2) and optimal weights for the 5 main methods (EUR, LAT, LAT+ANC, EUR+LAT, MultiPRS) are reported in Table 1 (optimal P-value thresholds for each method are reported in Supplementary Table 1). The EUR and LAT methods, which each use only one source of training data, performed similarly. This reflects a tradeoff between the larger training sample size for EUR and the target-matched LD patterns for LAT (confirmed by higher accuracy when using EUR training data to predict in Europeans; Table 1 caption). EUR+LAT attained a >78% relative improvement over either EUR or LAT (and chose similar weights for EUR and LAT), highlighting the advantages of incorporating multiple sources of training data. When including an ancestry predictor, we observed a relative improvement of 20% for MultiPRS compared to EUR+LAT, reflecting small genetic effects of ancestry on phenotype that can arise from random genetic drift between populations at causal markers.

Predictions using Latino effect sizes that were not adjusted for genetic ancestry ($\mathrm{LAT_{unadj}}$, $\mathrm{EUR+LAT_{unadj}}$, $\mathrm{EUR+LAT_{unadj}}$+ANC, as compared to LAT, EUR+LAT, MultiPRS) were much less accurate (Supplementary Table 1), consistent with previous work[17]; this is a consequence of the fact that $\mathrm{LAT_{unadj}}$ predictions were dominated by genetic ancestry (R^2 = 0.86; Supplementary Table 2). We also observed a modest correlation (R^2 = 0.15) between the EUR prediction and genetic ancestry (Supplementary Table 2), again reflecting small genetic effects of ancestry on phenotype that can arise from random genetic drift between populations at causal markers. The relative performance of the different prediction methods was similar in simulations in which phenotypes explicitly contained an ancestry term, representing environmentally driven stratification (Supplementary Table 3).

We extrapolated the results in Table 1 to larger sample sizes by limiting the simulations to subsets of chromosomes, as in previous work[7] (Figure 1 and Supplementary Table 4). MultiPRS was the best performing method in each of these experiments. We also performed simulations using predictions from all SNPs including the causal SNPs (Supplementary Figure 1 and Supplementary Table 5). In these experiments, MultiPRS was once again the best performing method, but now EUR performed much better than LAT, consistent with the larger training sample size for EUR and the fact that differential tagging of causal SNPs is of reduced importance when causal SNPs are typed.

Analyses of type 2 diabetes.

We applied the same methods to predict T2D in Latino target samples from the SIGMA T2D data set. We used publicly available European summary statistics from DIAGRAM[20] as European training data (12,171 cases and 56,862 controls; effective sample size = 40,100) and SIGMA T2D genotypes and phenotypes[16] (3,848 cases and 4,366 controls; effective sample size = 8,181) as Latino training and validation data, employing 10-fold cross-validation.

Prediction accuracies (R^2) and optimal weights for the 5 main methods (EUR, LAT, LAT+ANC, EUR+LAT, MultiPRS) are reported in Table 2 (other prediction metrics are reported in Supplementary Table 6). EUR performed only 44% better than LAT despite the much larger training sample size, again reflecting a tradeoff between sample size and target-matched LD patterns. EUR+LAT attained improvements ranging from 72%-148% over EUR and LAT respectively (and chose slightly higher weights for EUR than for LAT), again highlighting the advantages of incorporating multiple sources of training data. Although adding an ancestry predictor to LAT produced a substantial improvement (LAT+ANC vs. LAT), adding an ancestry predictor to EUR+LAT produced only a very small improvement for MultiPRS compared to EUR+LAT; this can be explained by the large negative correlation between the European PRS (EUR) and European ancestry (R = -0.70; Supplementary Table 7), such that any predictor that includes EUR already includes effects of genetic ancestry. This correlation is far larger than analogous correlations due to random genetic drift in our simulations (Supplementary Table 2), suggesting that this systematically lower load of T2D risk alleles in Latino individuals with more European ancestry could be due to polygenic selection[21,22] in ancestral European and/or Native American populations; previous studies using top GWAS-associated SNPs have also reported continental differences in genetic risk for T2D[23,24]. As in our simulations, predictions using Latino effect sizes that were not adjusted for genetic ancestry ($\text{LAT}_{\text{unadj}}$, $\text{EUR+LAT}_{\text{unadj}}$, $\text{EUR+LAT}_{\text{unadj}}$+ANC, as compared to LAT, EUR+LAT, MultiPRS) were much less accurate (Supplementary Table 8), a consequence of the fact that these predictions were dominated by genetic ancestry (Supplementary Table 9). We also computed predictions for each method using imputed SNPs from the SIGMA T2D data set; this did not improve prediction accuracy, but MultiPRS was still the best performing method (Supplementary Table 10).

We investigated how the prediction accuracy of each method varied as a function of P-value thresholds, by varying either the EUR P-value threshold (Figure 2a and Supplementary Table 11) or the LAT P-value threshold (Figure 2b and Supplementary Table 12). In both cases, permissive P-value thresholds performed best, reflecting the relatively small sample sizes analyzed. However, the prediction accuracy of MultiPRS was relatively stable, with prediction $R^2 > 3.7\%$ across all EUR P-value thresholds (Figure 2a) and $R^2 > 3.4\%$ across all LAT p-value thresholds (Figure 2b). In Figure 2a, we observe that as the EUR P-value threshold becomes more stringent, the difference in prediction accuracy between MultiPRS and EUR+LAT increases, because EUR is less able to capture polygenic ancestry effects (see above).

Discussion

We have shown that MultiPRS attains a >70% improvement in prediction accuracy for type 2 diabetes in a Latino cohort compared to prediction methods that use training data from a single population. This improvement is consistent with simulations, and reduces the well-documented gap in risk prediction accuracy between European and non-European target populations[7]. Intuitively, MultiPRS leverages both large training sample sizes and training data with target-matched LD patterns. We note that the effects of differential tagging (or different causal effect sizes) in different populations can potentially be quantified using cross-population genetic correlation[25], and that leveraging data from a different population to improve predictions is a natural analogue to leveraging data from a correlated trait[12].

Despite the advantages of MultiPRS, our work is subject to several limitations. First, although we have demonstrated large relative improvements in prediction accuracy, absolute prediction accuracies are currently not large enough to be clinically useful, which will require larger sample sizes[3]. Second, it may be possible to attain higher prediction accuracy using methods that fit all markers simultaneously, such as Best Linear Unbiased Predictor (BLUP) methods and their extensions[8–14]. However, these methods require raw genotypes/phenotype data, which is not available for the European type 2 diabetes summary statistics that we analyzed here. Third, our LDpred risk prediction method[7], which analyzes summary statistics in conjunction with LD information from a reference panel, is more accurate in European populations than the LD-pruning + p-value thresholding approach employed by MultiPRS. However, we elected not to employ LDpred in the current setting due to the complexities of admixture-LD when using LD information from a reference panel in admixed populations[26]. Fourth, our work has highlighted the importance of validating predictions using a separate cohort (instead of cross-validation within the same cohort) in order to avoid inflation in prediction accuracies due to population stratification effect[7]. However, we believe this issue is not substantially affecting our conclusions, both because our empirical results are consistent with simulations and because the type 2 diabetes prediction accuracy we obtained in a Latino-only (LAT) prediction scheme with no P-value thresholding ($R^2 = 0.017$; see Supplementary Table 13) imply values of SNP-heritability ($h_{g,obs}^2 = 0.39$; $h_{g,liab}^2 = 0.38$ assuming 8% prevalence[2]) that are consistent with previous estimates[2]. Fifth, MultiPRS includes an ancestry predictor, but in some cases it may be preferable to construct and evaluate a predictor that does not benefit from ancestry information; we note that very similar results for type 2 diabetes were obtained using a prediction method (EUR+LAT) that does not include the ancestry predictor. Sixth, we optimize P-value thresholds and weights for each predictor via an in-sample fit on validation samples, consistent with widely used LD-pruning and P-value thresholding methods[1–4]; an additional layer of validation using samples not used to fit those parameters would formally be most statistically appropriate, but is unlikely to impact our results given the small number of parameters (up to two P-value thresholds and weights for EUR, LAT, and two ANC predictors) and fairly large validation sample (N=8,214), as we have demonstrated in previous work[17]. Seventh, we have not considered here how to improve prediction accuracy in data sets with related individuals[13]. Eighth, we focused our analyses on common variants, but future work may wish to consider rare variants as well. Finally, we have only considered a Latino target population, but it is also of interest to apply MultiPRS to other non-European target populations.

Consortia

Members of the SIGMA Type 2 Diabetes Consortium are Amy L. Williams[1,2], Suzanne B. R. Jacobs[1], Hortensia Moreno-Macías[3], Alicia Huerta-Chagoya[4,5], Claire Churchouse[1], Carla Márquez-Luna[6], Humberto García-Ortíz[6], María José Gómez-Vázquez[4,7], Stephan

Ripke[1,15], Alisa K. Manning[1], Benjamin Neale[1,15], David Reich[1,2], Daniel O. Stram[11], Juan Carlos Fernández-López[6], Sandra Romero-Hidalgo[6], Nick Patterson[1], Suzanne B. R. Jacobs[1], Claire Churchhouse[1], Shuba Gopal[22], James A. Grammatikos[22], Ian C. Smith[23], Kevin H. Bullock[22], Amy A. Deik[22], Amanda L. Souza[22], Kerry A. Pierce[22], Clary B. Clish[22], Irma Aguilar-Delfín[6], Angélica Martínez-Hernández[6], Federico Centeno- Cruz[6], Elvia Mendoza-Caamal[6], Cristina Revilla-Monsalve[16], Sergio Islas-Andrade[16], Emilio Córdova[6], Eunice Rodríguez-Arellano[17], Xavier Soberón[6], María Elena González-Villalpando[8] , Brian E. Henderson[11], Kristine Monroe[11], Lynne Wilkens[18], Laurence N. Kolonel[18], and Loic Le Marchand[18], Laura Riba[5], María Luisa Ordóñez-Sánchez[4], Rosario Rodríguez-Guillén[4], Ivette Cruz-Bautista[4], Maribel Rodríguez-Torres[4], Linda Liliana Muñoz-Hernández[4], Donají Gómez[4], Ulises Alvirde[4], Olimpia Arellano[4], Robert C. Onofrio[19], Wendy M. Brodeur[19], Diane Gage[19], Jacquelyn Murphy[1], Jennifer Franklin[19], Scott Mahan[19], Kristin Ardlie[19], Andrew T. Crenshaw[19], and Wendy Winckler[19] , Maria L. Cortes[53], Noël P. Burtt[1], Carlos A. Aguilar-Salinas[4], Clicerio González-Villalpando[8], Jose C. Florez[1,9,10], Lorena Orozco[6], Christopher A. Haiman[11], Teresa Tusié-Luna[4,5], David Altshuler[1,2,9,10,12,13,14]

1 Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA
2 Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA
3 Universidad Autonoma Metropolitana, Mexico City, Mexico
4 Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico
5 Instituto de Investigaciones Biomédicas, UNAM. Unidad de Biología Molecular y Medicina Genómica, UNAM/INCMNSZ, Mexico City, Mexico
6 Instituto Nacional de Medicina Genómica, Mexico City, Mexico
7 Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León 66451, México
8 Centro de Estudios en Diabetes, Unidad de Investigacion en Diabetes y Riesgo Cardiovascular, Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Mexico City, Mexico
9 Center for Human Genetic Research and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston, Massachusetts, USA
10 Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA
11 Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California,, USA
12 Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA
13 Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts, USA
14 Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
15 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA 16 Instituto Mexicano del Seguro Social SXXI, Mexico City, Mexico
17 Instituto de Seguridad y Servicios Sociales para los Trabajadores del Estado, Mexico City, Mexico

18 Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA
19 The Genomics Platform, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA
20 Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany
21 Palaeolithic Department, Institute of Archaeology and Ethnography, Russian Academy of Sciences, Siberian Branch, 630090 Novosibirsk, Russia
22 The Metabolite Profiling Platform, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA
23 Cancer Biology Program, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA 24 University of Minnesota, Minneapolis, Minnesota, USA 25 University of California San Francisco, San Francisco, California, USA 26 Duke-National University of Singapore Graduate Medical School, Singapore, Singapore
27 Saw Swee Hock School of Public Health, National University of Singapore, Singapore
28 Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore,
Singapore
29 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK 30 Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor,
Michigan, USA
30 Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor,
Michigan, USA
31 Department of Medicine, Department of Genetics, Albert Einstein College of Medicine, Bronx, New York, USA
32 Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas, USA
33 Center for Genomics and Personalized Medicine Research, Center for Diabetes Research, Department of Biochemistry, Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA
34 Department of Epidemiology and Biostatistics, Imperial College London, London, UK
35 Imperial College Healthcare NHS Trust, London, UK.
36 Ealing Hospital National Health Service (NHS) Trust, Middlesex, UK
37 Department of Biomedical Science, Hallym University, Chuncheon, Gangwon-do, Korea
38 Endocrinology and Metabolism Service, Hadassah-Hebrew University Medical School, Jerusalem, Israel
39 Israel Diabetes Research Group (IDRG), Israel 40 Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA
40 Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA
41 National Heart and Lung Institute (NHLI), Imperial College London, Hammersmith Hospital, London, UK

42 Department of Medicine, University of Eastern Finland, Kuopio Campus and Kuopio University Hospital, Kuopio, Finland
43 Center for Genome Science, National Institute of Health, Osong Health Technology Administration Complex, Chungcheongbuk-do, Cheongwon-gun, Gangoe-myeon, Yeonje-ri, Korea
44 Department of Epidemiology and Public Health, National University of Singapore, Singapore, Singapore
45 Centre for Molecular Epidemiology, National University of Singapore, Singapore, Singapore 46 Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore
46 Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore
47 Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore, Singapore
48 Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore.
49 Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, USA
50 Division of Nephrology, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas
51 Division of Diabetes, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas
52 Division of Clinical Epidemiology, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas
53 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

Acknowledgements

We are grateful to B. Vilhjalmsson for helpful discussions. This research was funded by NIH grant R01 GM105857 (A.L.P.) and a Harvard/CONACYT/FMH scholarship (C.M.L.).

Web Resources

PLINK2: https://www.cog-genomics.org/plink2.
DIAGRAM summary association statistics: http://www.diagram-consortium/org/.

References

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

2. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44,** 483–489 (2012).
3. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45,** 400–405, 405e1–3 (2013).
4. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9,** e1003348 (2013).
5. Shah, S. *et al.* Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am. J. Hum. Genet.* **97,** 75–85 (2015).
6. Palla, L. & Dudbridge, F. A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. *Am. J. Hum. Genet.* **97,** 250–259 (2015).
7. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).
8. de los Campos, G., Gianola, D. & Allison, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* **11,** 880–886 (2010).
9. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9,** e1003264 (2013).
10. Golan, D. & Rosset, S. Effective genetic-risk prediction using mixed models. *Am. J. Hum. Genet.* **95,** 383–393 (2014).
11. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24,** 1550–1557 (2014).

12. Maier, R. *et al.* Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* **96,** 283–294 (2015).
13. Tucker, G. *et al.* Two-Variance-Component Model Improves Genetic Prediction in Family Datasets. *Am. J. Hum. Genet.* **97,** 677–690 (2015).
14. Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* **11,** e1004969 (2015).
15. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11,** 356–366 (2010).
16. SIGMA Type 2 Diabetes Consortium *et al.* Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506,** 97–101 (2014).
17. Chen, C.-Y., Han, J., Hunter, D. J., Kraft, P. & Price, A. L. Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. *Genet. Epidemiol.* **39,** 427–438 (2015).
18. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).
19. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46,** 100–106 (2014).
20. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44,** 981–990 (2012).
21. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44,** 1015–1019 (2012).

22. Robinson, M. R. *et al.* Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* **47,** 1357–1362 (2015).
23. Chen, R. *et al.* Type 2 Diabetes Risk Alleles Demonstrate Extreme Directional Differentiation among Human Populations, Compared to Other Diseases. *PLOS Genet* **8,** e1002621 (2012).
24. Corona, E. *et al.* Analysis of the Genetic Basis of Disease in the Context of Worldwide Human Relationships and Migration. *PLOS Genet* **9,** e1003447 (2013).
25. Brown, B. C., Asian Genetic Epidemiology Network-Type 2 Diabetes (AGEN-T2D) Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic correlation estimates from summary statistics. *Am. J. Hum. Genet.* (in press).
26. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

Tables

Model	**Average weight (s.e.) associated to each predictor.**		**Average R^2 (s.e.)**
	European PRS	**Latino PRS**	
EUR	0.141 (0.004)		0.0218 (0.001)
LAT		0.155 (0.002)	0.0248 (0.001)
LAT+ANC		0.154 (0.002)	0.0345 (0.002)
EUR+LAT	0.133 (0.004)	0.147 (0.002)	0.0437 (0.001)
MultiPRS	0.148 (0.002)	0.145 (0.002)	0.0521 (0.002)

Table 1. Accuracy of 5 prediction methods in simulations. We report average R^2 on the observed scale over 100 simulations for each of the 5 main prediction methods. We assessed prediction accuracy in European target sample using 10-fold cross-validation and obtained R^2 =0.031 (s.e.=0.0006).

Model	**Weights associated to each predictor**		***R^2***
	European PRS	**Latino PRS**	
EUR	0.161		0.0259
LAT		0.134	0.0179
LAT+ANC		0.143	0.0339
EUR+LAT	0.171	0.137	0.0444
MultiPRS	0.148	0.138	0.0448

Table 2. Accuracy of 5 prediction methods in analyses of type 2 diabetes. We report R^2 on the observed scale for each of the 5 main prediction methods. We obtained similar relative results using Nalgerkerke R^2, R^2 on the liability scale and AUC (Supplementary Table 6).

Figures

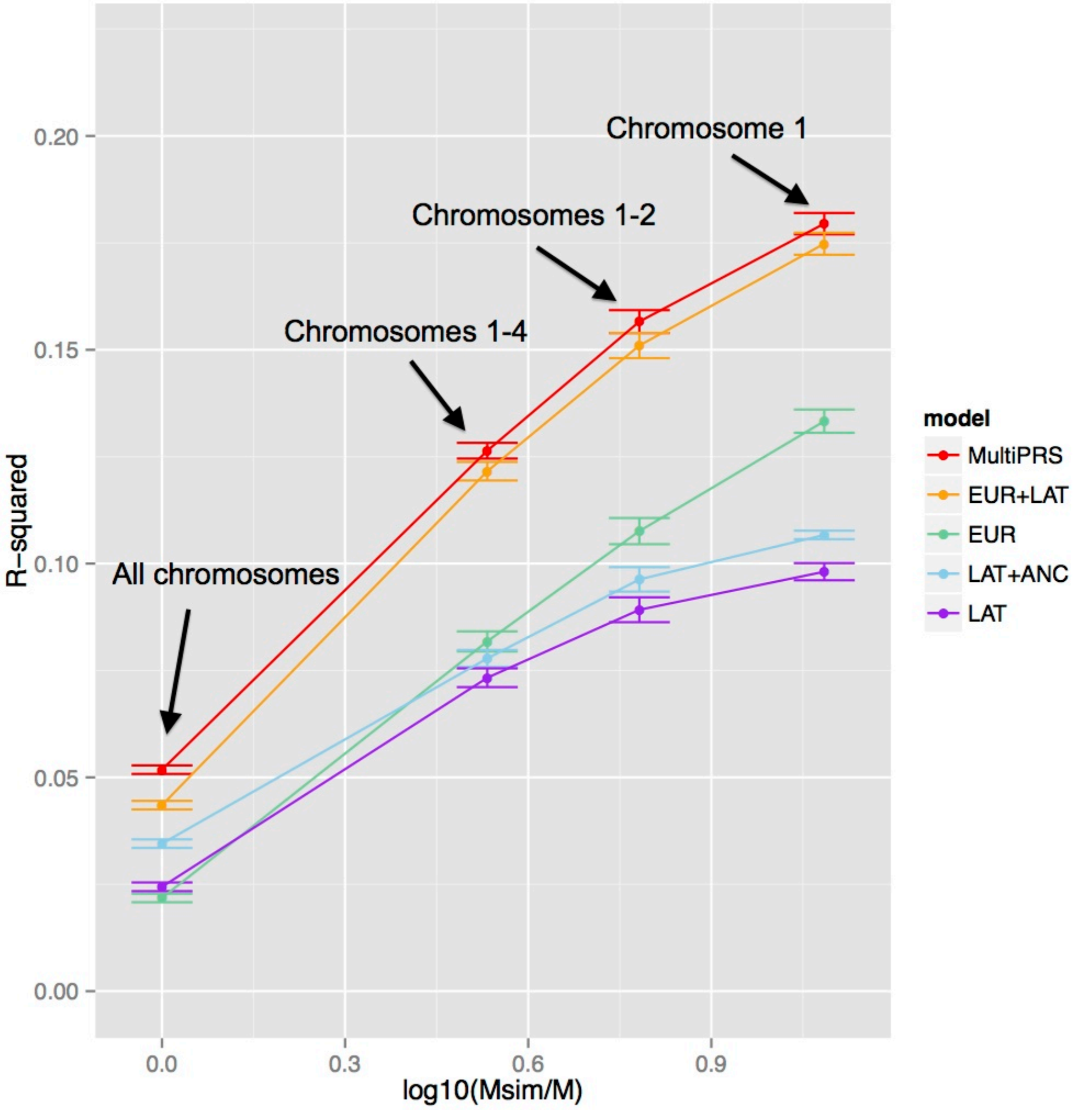


Figure 1. Accuracy of 5 prediction methods in simulations using subsets of chromosomes. We report prediction accuracies for each of the 5 main prediction methods as a function of Msim/M, where M=232,629 is the total number of SNPs and Msim is the actual number of SNPS used in each simulation: 232,629 (all chromosomes), 68,188 (chromosomes 1-4), 38412 (chromosomes 1-2), and 19087 (chromosome 1). Numerical results are provided in Supplementary Table 4.

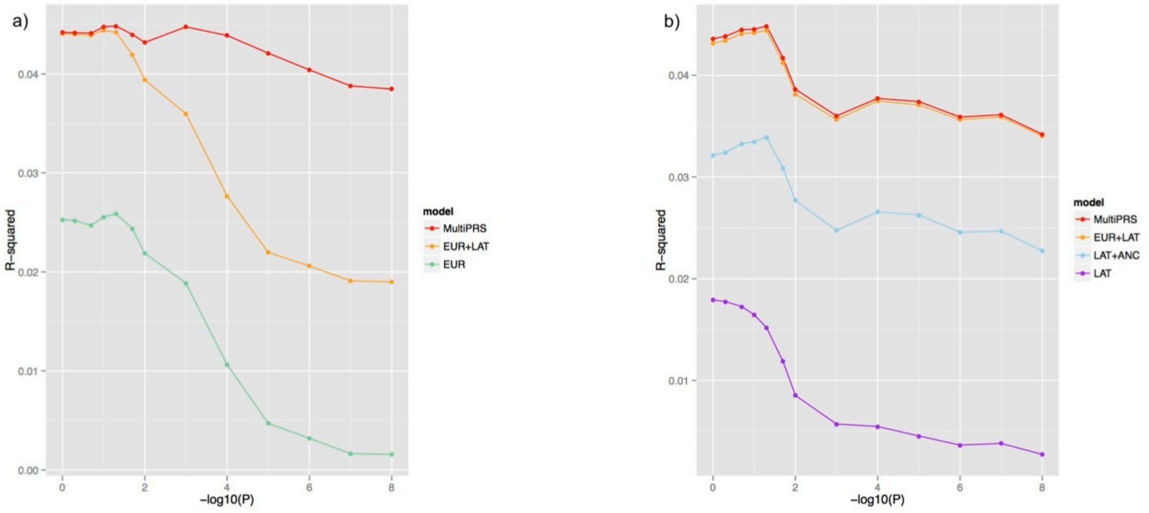


Figure 2. Accuracy of 5 prediction methods in analyses of type 2 diabetes as a function of P-value thresholds. We report prediction accuracies for each of the 5 main prediction methods as a function of (a) EUR P-value threshold, where applicable (with optimized LAT P-value threshold, where applicable) and (b) LAT P-value threshold, where applicable (with optimized EUR P-value threshold, where applicable). Numerical results are provided in Supplementary Table 11 and Supplementary Table 12.

Supplementary Tables

Model	Average weight (s.e.) associated to each predictor.		Average R^2 (s.e.)	European training	Latino training
	European PRS	Latino PRS		P value threshold	P value threshold
EUR	0.141 (0.004)		0.0218 (0.001)	0.01	
LAT$_{\text{unadj}}$		0.104 (0.005)	0.0133 (0.002)		0.01
LAT$_{\text{unadj}}$+ANC		0.385 (0.022)	0.0168 (0.001)		0.01
LAT		0.155 (0.002)	0.0248 (0.001)		0.1
LAT+ANC		0.155 (0.002)	0.0349 (0.002)		0.08
EUR+LAT$_{\text{unadj}}$	0.151 (0.002)	0.089 (0.007)	0.0325 (0.002)	0.01	0.01
EUR+LAT$_{\text{unadj}}$+ANC	0.155 (0.002)	0.359 (0.021)	0.0361 (0.001)	0.01	0.01
EUR+LAT	0.133 (0.004)	0.147 (0.002)	0.0437 (0.001)	0.01	0.3
MultiPRS	0.148 (0.002)	0.146 (0.002)	0.0521 (0.002)	0.01	0.2

Supplementary Table 1. Accuracy of 9 prediction methods in simulations. We report prediction accuracies for methods using both ancestry-adjusted Latino effect sizes (LAT) and ancestry-unadjusted Latino effect sizes (LAT$_{\text{unadj}}$). Reported values are mean R^2 on the observed scale over 100 simulations.

Model	**Average *R^2* (s.e.)**
EUR	0.152 (0.017)
$\text{LAT}_{\text{unadj}}$	0.861 (0.024)
$\text{LAT}_{\text{unadj}}$+ANC	0.479 (0.038)
LAT	0.030 (0.004)
LAT+ANC	0.232 (0.026)
EUR+$\text{LAT}_{\text{unadj}}$	0.310 (0.021)
EUR+$\text{LAT}_{\text{unadj}}$+ANC	0.235 (0.023)
EUR+LAT	0.078 (0.011)
MultiPRS	0.166 (0.017)

Supplementary Table 2. *R^2* with European ancestry for 9 prediction methods in simulations. European ancestry is represented by PC1 in the SIGMA data set. Reported values are mean *R^2*over 100 simulations. The average R^2 between ancestry and phenotype was 0.011.

Model	**Average weight (s.e.) associated to each predictor.**		**Average R^2 (s.e.)**	**European training**	**Latino training**
	European PRS	**Latino PRS**		**P-value threshold**	**P-value threshold**
EUR	0.135 (0.004)		0.0201 (0.001)	0.01	
$\text{LAT}_{\text{unadj}}$		0.115 (0.001)	0.0133 (0.0001)	0.01	
$\text{LAT}_{\text{unadj}}$+ANC		0.51 (0.022)	0.0159 (0.0002)	0.01	
LAT		0.151 (0.002)	0.0237 (0.001)		0.08
LAT+ANC		0.154 (0.002)	0.0359 (0.001)		0.05
EUR+$\text{LAT}_{\text{unadj}}$	0.154 (0.002)	0.103 (0.006)	0.0329 (0.0004)	0.01	0.01
EUR+$\text{LAT}_{\text{unadj}}$+ANC	0.156 (0.002)	0.471 (0.02)	0.0355 (0.0005)	0.01	0.01
EUR+LAT	0.127 (0.004)	0.145 (0.003)	0.0414 (0.001)	0.01	0.08
MultiPRS	0.147 (0.002)	0.146 (0.001)	0.0533 (0.001)	0.01	0.08

Supplementary Table 3. Accuracy of 9 prediction methods in simulations with ancestry-correlated phenotypes. We report prediction accuracies for methods using both ancestry-adjusted Latino effect sizes (LAT) and ancestry-unadjusted Latino effect sizes ($\text{LAT}_{\text{unadj}}$). Reported values are mean R^2 on the observed scale over 100 simulations .

Model	Chr 1	Chr 1-2	Chr 1-4	Chr 1-22
EUR	0.133 (0.002)	0.108 (0.003)	0.082 (0.002)	0.022 (0.001)
LAT	0.098 (0.002)	0.089 (0.003)	0.073 (0.002)	0.024 (0.001)
LAT+ANC	0.107 (0.001)	0.096 (0.003)	0.078 (0.002)	0.035 (0.002)
EUR+LAT	0.175 (0.002)	0.151 (0.003)	0.122 (0.002)	0.044 (0.001)
MultiPRS	0.180 (0.002)	0.157 (0.003)	0.126 (0.002)	0.052 (0.002)

Supplementary Table 4. Numerical values of results displayed in Figure 1. We report prediction accuracies for each of the 5 main prediction methods, for each subset of chromosomes.

Model	**Chr 1**	**Chr 1-2**	**Chr 1-4**	**Chr 1-22**
EUR	0.277 (0.003)	0.247 (0.003)	0.207 (0.002)	0.079 (0.003)
LAT	0.143 (0.003)	0.130 (0.003)	0.113 (0.002)	0.042 (0.001)
LAT+ANC	0.158 (0.003)	0.141 (0.003)	0.120 (0.002)	0.052 (0.002)
EUR+LAT	0.295 (0.003)	0.267 (0.003)	0.232 (0.002)	0.106 (0.002)
MultiPRS	0.301 (0.002)	0.275 (0.003)	0.243 (0.002)	0.122 (0.002)

Supplementary Table 5. Numerical values of results displayed in Supplementary Figure 1. We report prediction accuracies for each of the 5 main prediction methods, for each subset of chromosomes, in simulations including the causal SNPs.

Model	**Observed-scale R^2**	**Nagelkerke R^2**	**Liability-scale R^2**	**AUC**
EUR	0.0259	0.0346	0.0258	0.58841
LAT	0.0179	0.0239	0.0178	0.5764
LAT+ANC	0.0309	0.0412	0.0309	0.5971
EUR+LAT	0.0444	0.0593	0.0447	0.6192
MultiPRS	0.0445	0.0595	0.0448	0.6191

Supplementary Table 6. Accuracy of 5 prediction methods in analyses of type 2 diabetes, using alternate prediction metrics. Liability-scale R^2 was computed assuming a disease prevalence of K=0.08.

R	**EUR**	**LAT**	**European ancestry**	**T2D**
EUR	1	0.005	-0.699	0.161
LAT	0.005	1	0.031	0.133
European ancestry	-0.699	0.031	1	-0.112
T2D	0.161	0.133	-0.112	1

Supplementary Table 7. Pairwise correlations (*R*) between EUR and LAT polygenic risk scores, European ancestry and T2D phenotype in analyses of T2D. European ancestry is represented by PC1 in the SIGMA data set.

Model	Weight associated to each predictor		R^2	European training	Latino training
	European PRS	Latino PRS		P-value threshold	P-value threshold
EUR	0.161		0.0259	0.05	
$\mathrm{LAT}_{\mathrm{unadj}}$		0.117	0.0125		0.02, 0.2
$\mathrm{LAT}_{\mathrm{unadj}}$+ANC		0.674	0.0155		1
LAT		0.134	0.0179		1
LAT+ANC		0.143	0.0339		1
EUR+$\mathrm{LAT}_{\mathrm{unadj}}$	0.159	0.180	0.0259	0.05	0.02
EUR+$\mathrm{LAT}_{\mathrm{unadj}}$ +ANC	0.159	0.702	0.0279	0.05	1
EUR+LAT	0.171	0.137	0.0444	0.1	0.05
MultiPRS	0.148	0.138	0.0448	0.05	0.05

Supplementary Table 8. Accuracy of 9 prediction methods in type 2 diabetes analyses. We report prediction accuracies for methods using both ancestry-adjusted Latino effect sizes (LAT) and ancestry-unadjusted Latino effect sizes ($\mathrm{LAT}_{\mathrm{unadj}}$).

Model	***R***	***R*²**
EUR	-0.699	0.488
$\text{LAT}_{\text{unadj}}$	-0.996	0.992
$\text{LAT}_{\text{unadj}}$+ANC	-0.999	0.997
LAT	0.016	0.0003
LAT+ANC	-0.634	0.402
EUR+$\text{LAT}_{\text{unadj}}$	-0.703	0.494
EUR+$\text{LAT}_{\text{unadj}}$ +ANC	-0.690	0.475
EUR+LAT	-0.493	0.243
MultiPRS	-0.529	0.280

Supplementary Table 9. *R* and *R*² with European ancestry for 9 prediction methods in analyses of type 2 diabetes. European ancestry is represented by PC1 in the SIGMA data set.

Model	**Weights associated to each predictor**		***R*^{2}**
	EUR	**LAT**	
EUR	0.159		0.0255
LAT		0.149	0.0222
LAT+ANC		0.131	0.0303
EUR+LAT	0.148	0.129	0.0422
MultiPRS	0.149	0.129	0.0426

Supplementary Table 10. Accuracy of 5 prediction methods in analyses of type 2 diabetes, using imputed genotypes. We report R^2 on the observed scale for each of the 5 main prediction methods.

	P-value Threshold												
Model	**10^{-8}**	**10^{-7}**	**10^{-6}**	**10^{-5}**	**10^{-4}**	**10^{-3}**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.5**	**1**
EUR	0.0016	0.0016	0.0032	0.0047	0.0106	0.0189	0.0219	0.0244	0.0259	0.0255	0.0247	0.0252	0.0253
EUR+LAT	0.0190	0.0191	0.0206	0.0220	0.0277	0.0360	0.0394	0.0419	0.0442	0.0444	0.0439	0.0440	0.0441
MultiPRS	0.0374	0.0377	0.0393	0.0411	0.0428	0.0437	0.0426	0.0434	0.0445	0.0445	0.0440	0.0441	0.0441

Supplementary Table 11. Numerical values for results displayed in Figure 2a. We report R^2 on the observed scale for each of the 3 prediction methods that include the EUR predictor.

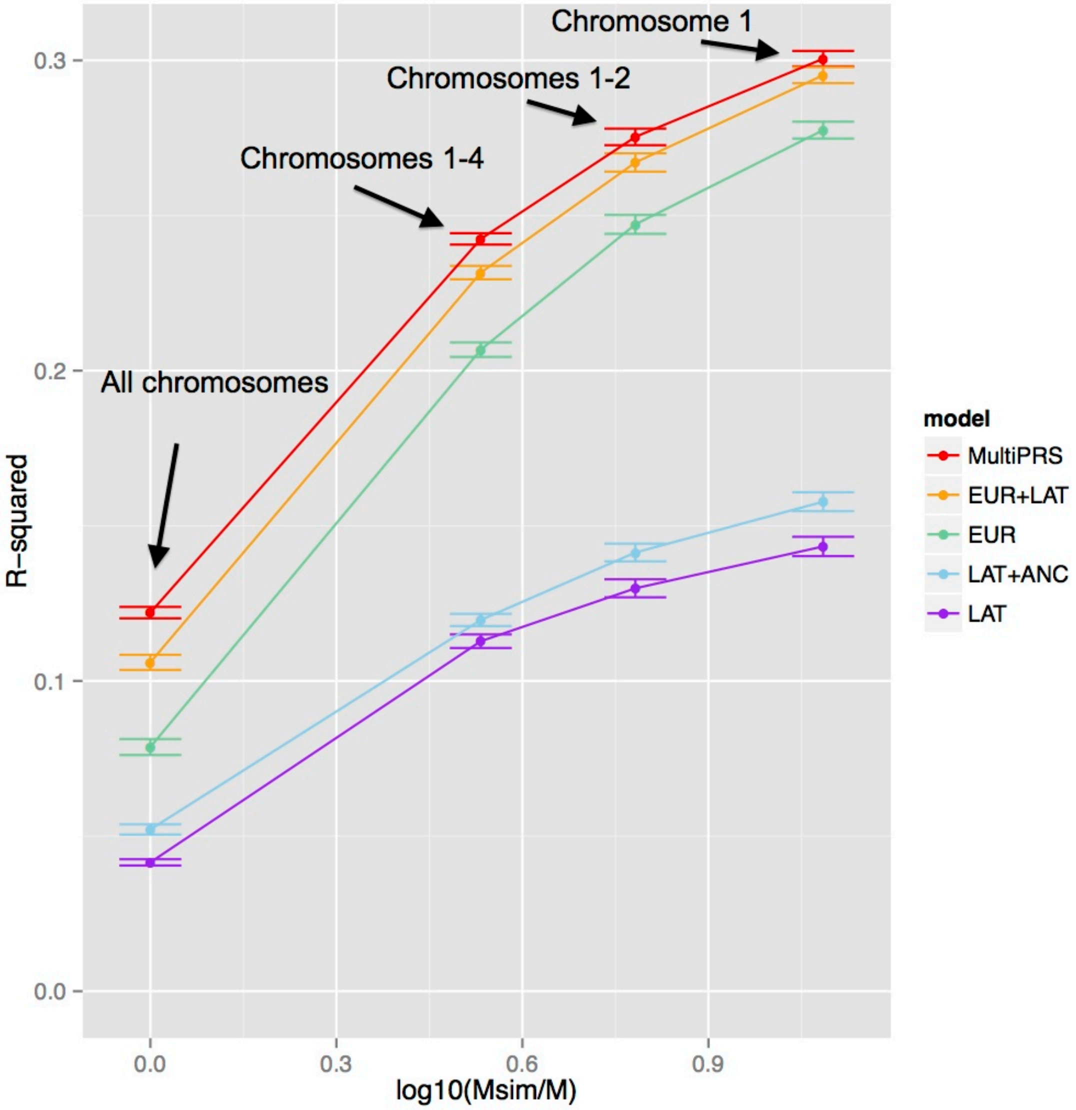
	P-value Threshold												
Model	**10^{-8}**	**10^{-7}**	**10^{-6}**	**10^{-5}**	**10^{-4}**	**10^{-3}**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.5**	**1**
LAT	0.0027	0.0038	0.0037	0.0045	0.0055	0.0057	0.0085	0.0119	0.0152	0.0165	0.0173	0.0178	0.0179
LAT+ANC	0.0215	0.0235	0.0233	0.0249	0.0252	0.0234	0.0264	0.0296	0.0326	0.0322	0.0320	0.0312	0.0309
EUR+LAT	0.0340	0.0359	0.0357	0.0371	0.0375	0.0357	0.0382	0.0412	0.0444	0.0442	0.0441	0.0434	0.0431
MultiPRS	0.0342	0.0361	0.0359	0.0374	0.0377	0.0357	0.0383	0.0414	0.0445	0.0442	0.0441	0.0434	0.0432

Supplementary Table 12. Numerical values for results displayed in Figure 2b. We report R^2 on the observed scale for each of the 4 prediction methods that include the LAT predictor.

Model	**Weights associated to each predictor**		R^2
	European PRS	**Latino PRS**	
EUR	0.159		0.0253
LAT		0.134	0.0179
LAT+ANC		0.135	0.0321
EUR+LAT	0.157	0.132	0.0426
MultiPRS	0.16	0.132	0.0427

Supplementary Table 13. Accuracy of 5 prediction methods in analyses of type 2 diabetes with no P-value thresholding. We report R^2 on the observed scale for each of the 5 main prediction methods with no P-value thresholding, i.e. $P_T = 1$.

Supplementary Figures



Supplementary Figure 1. Accuracy of 5 prediction methods in simulations using subsets of chromosomes, including the causal SNPs. We report prediction accuracies for each of the 5 main prediction methods as a function of Msim/M, where M=232,629 is the total number of SNPs and Msim is the actual number of SNPS used in each simulation: 232,629 (all chromosomes), 68,188 (chromosomes 1-4), 38412 (chromosomes 1-2), and 19087 (chromosome 1). Numerical results are provided in Supplementary Table 5.