Genomic prediction: some caveats

Introduction

In the early 2010s, Turchin and colleagues [6][1] studied the association between variation at SNP loci and height in humans. They showed that both individual alleles known to be associated with increased height and in genome-wide analysis are elevated in northern European populations compared to populations from southern Europe. They argued that these differences were consistent with weak selection at each of the loci ($s \approx [10^{-3}, 10^{-5}]$) rather than genetic drift alone.

Allele frequency comparisons

Turchin et al. used allele frequency estimates from the Myocardial Infarction Genetics consortium (MIGen) [2] and the Population Reference Sample (POPRES) [5]. For the MIGen analysis, they compared allele frequencies in 257 US individuals of northern European ancestry with those in 254 Spanish individusls at loci that are known to be associated with height based on GWAS analysis[2] and found differences greater than those expected based on 10,000 SNPs drawn at random and matched to allele frequencies at the target loci in each population. They performed a similar analysis with the POPRES sample and found similar results.

Turchin et al. were aware that the association could be spurious if ancestry was not fully accounted for in these analyses, so they also used data collected by the Genetic Investigation of ANthropometric Traits consortium (GIANT) [4].[3] They noted that "control" SNPs used in the preceding analysis, i.e., the 10,000 SNPs drawn at random from the genome, with a tendency to increase height in the GIANT analysis also tended to be more frequent in the northern European sample.

They compared the magnitude of the observed differences at the most strongly associated 1400 SNPs with what would be expected if they were due entirely to drift and what would

[1] *Michael* Turchin, not *Peter* Turchin of UConn's EEB department.

[2] See Turchin et al. for details.

[3] This includes the GWAS on height that I mentioned in the last lecture.

be expected if they were due to a combination of drift and selection. A likelihood-ratio test of the drift alone model *versus* the drift-selection model provided strong support for the drift-model.

Second thoughts

Within sample stratification

This all seems very promising, but a word of caution is in order. Berg et al. [1] re-examined these claims using new data available from the UK Biobank (`https://www.bdi.ox.ac.uk/research/uk-biobank`), which includes a host of information on individual phenotypes as well as genome-wide genotypes for the 500,000 individuals included in the sample.[4] They failed to detect evidence of a cline in polygenic scores in their analysis (Figure 2).

In thinking about this result, it's important to understand that Berg et al. [1] did something a bit different from what we did, but it's exactly what you'd want to do if polygenic scores worked. They estimated polygenic scores from each of the data sets identified in the figure. Then they used those scores to estimate polygenic scores for a new set of samples derived from the 1000 Genomes and Human Origins projects.[5] Since they did the same thing with all of the data sets, this difference from what we did doesn't account for the differences among data sets. As Berg et al. dug more deeply into the data, they concluded that all of the data sets "primarily capture real signals of association with height" but that the GIANT and R-15 sibs data sets, the ones that show the latitudinal (and longitudinal in the case of GIANT) associations do so because the estimated allelic effects in those data sets failed to fully remove confounding variation along the major geographic axes in Europe.

The Berg et al. analysis illustrates how difficult it is to remove confounding factors from GWAS and genomic prediction analyses. Turchin et al. are highly skilled population geneticists. If they weren't able to recognize the problem with stratification in the GIANT consortium data set, all of us should be concerned about recognizing it in our own. Indeed, I wonder whether the stratification within GIANT would ever have come to light had Berg et al. not had additional large data sets at there disposal in which they could try to replicate the results.

[4] Although all of the samples are from the UK, one of the data sets Berg et al. [1] studied included individuals of European, but non-UK, ancestry.

[5] See Berg et al. [1] for details.

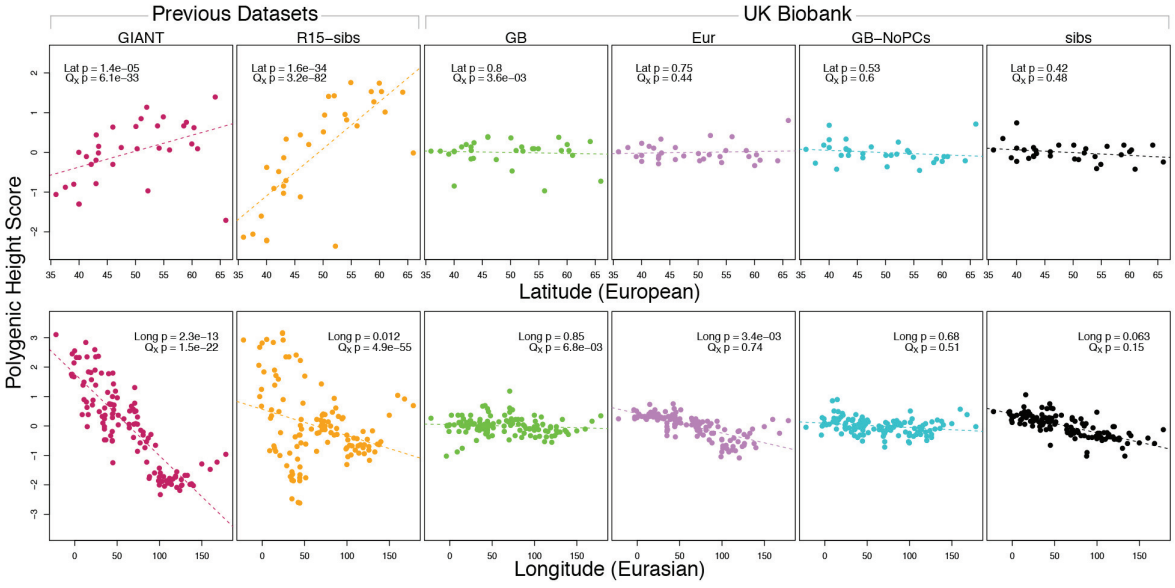

Figure 1: Polygenic score as a function of latitude and longitude for several different GWAS data sets. Each vertical column corresponds to a different data source. Notice that all of the UK Biobank samples fail to show either a latitudinal or a longitudinal cline in polygenic height score (from [1]).

Difficulties extrapolating polygenic scores

In one way the Berg et al. results are actually encouraging. They estimated effects in one set of data and used the genomic regressions estimated from those data to predict polygenic scores in a new data pretty successfully. Maybe it's difficult to be sure that the polygenic scores we estimate are useful for inferring anything about natural selection on the traits they predict, but if we could be sure that they allow us to predict phenotypes in populations we haven't studied yet, they could still be very useful. Can we trust them that far?

Unfortunately, the answer appears to be "No." Yair and Coop [7] recently studied the relationship between phenotypic stabilizing selection and genetic differentiation in isolated populations. They showed that even in a very simple model in which allelic effects at each locus are the same in both populations, polygenic scores estimated from one population may not perform very well in the other. Interestingly, as you can see in Figure **??**, the stronger the selection and the more stronly allelic differences influence the phenotype, the less well genomic predictions in one population work in the other.

That seems paradoxical, but interestingly it's not too difficult to understand if we think about what happens when we combine stabilizing selection with geographical isolation.[6] First, let's remind ourselves of a fundamental property of polygenic variation: Different genotypes can produce the same phenotype. Figure 3, which you've seen before, illustrates this when three loci influence the trait. While there is only one genotype that produces the dark red phenotype and only one that produces the white phenotype, there are four genotypes that produce the light red phenotype, four that produce the medium dark red phenotype, and six that produce the medium red phenotype. Goldstein and Holsinger [3] called this phenomenon *genetic redundancy*. As you can imagine, the number of redundant genotypes increases dramatically as the number of loci involved increases.[7]

Why does this redundancy matter? Let's consider what happens when we impose stabilizing selection on a polygenic trait, where

$$w(z) = \exp\left(\frac{-(z - z_0)^2}{2V_s}\right) \quad ,$$

where z_0 is the intermediate phenotype favored by selection, z is the phenotype of a particular individual, and V_s is the variance of the fitness function. If selection is weak ($V_s = 115.2$), then the relative fitness of a genotype 1 unit away from the optimum is 0.9957 while that of

[6] And the fun thing for me about this is that we get to finish out the course by returning to a paper I wrote with my first master's student more than 30 years ago.

[7] If the allelic effects are strictly additive, the number of genotypes corresponding to the intermediate phenotype is $\binom{2N}{N}$ where N is the number of loci. For $N = 10$, $\binom{2N}{N} = 184,756$. For $N = 100$, $\binom{2N}{N} = 9.05 \times 10^{58}$.

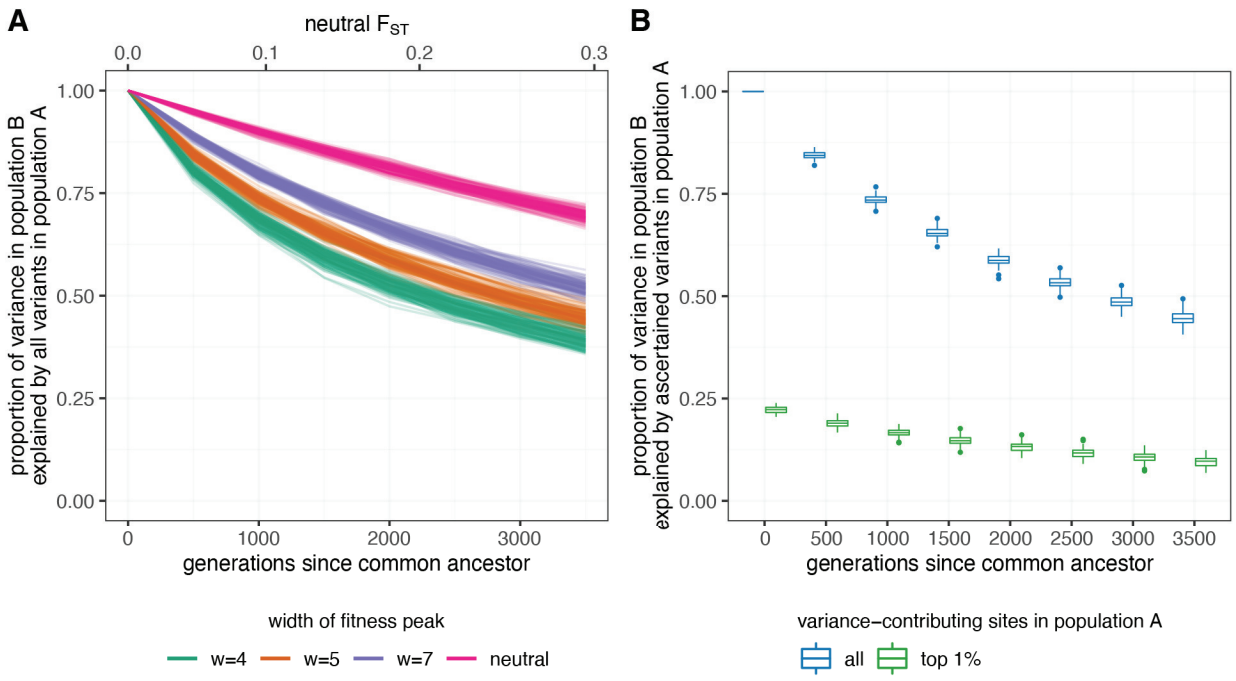

Figure 2: Polygenic score as a function of latitude and longitude for several different GWAS data sets. Each vertical column corresponds to a different data source. Notice that all of the UK Biobank samples fail to show either a latitudinal or a longitudinal cline in polygenic height score (from [7]).

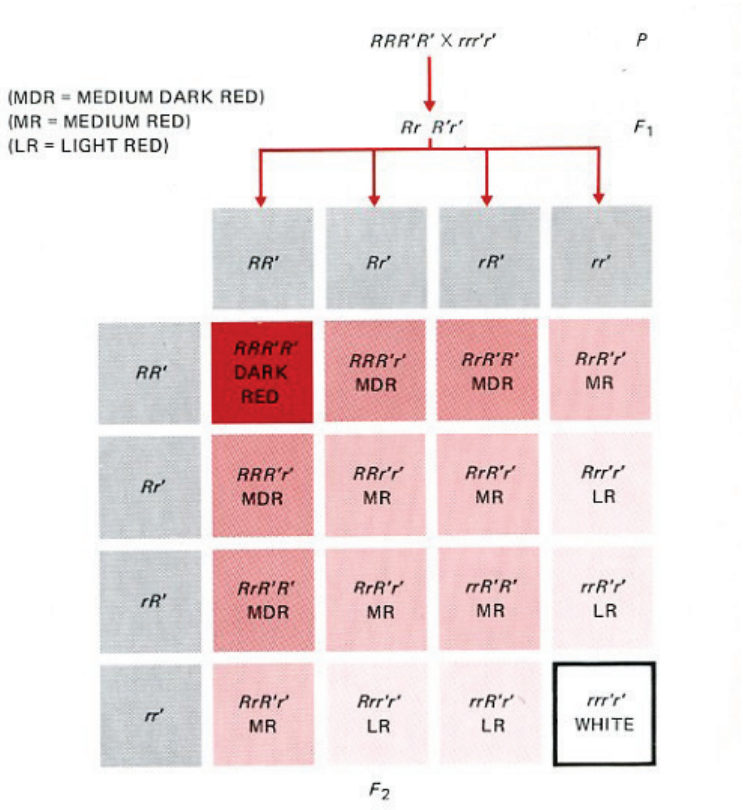

Figure 3: Results from one of Nilsson-Ehle's crosses illustrating polygenic inhertance of kernel color in wheat (from `http://www.biology-pages.info/Q/QTL.html`, accessed 9 April 2017).

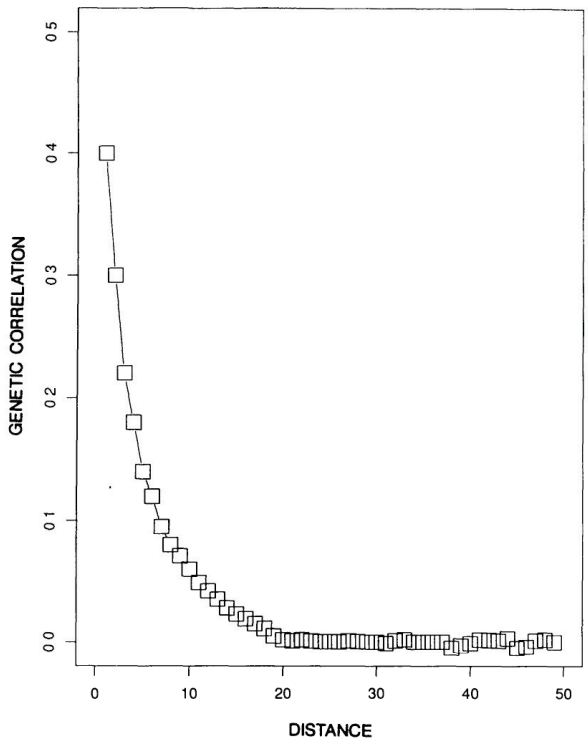

Figure 4: Isolation by distance with weak selection (from [3]).

a genotype 8 units away is only 0.7575. If 16 loci influence the trait, there are 601,080,390 genotypes that produce the optimum phenotype and have the same fitness. There are another 1,131,445,440 genotypes whose fitness within one percent of the optimum. Not only are there a lot of different genotypes with roughly the same fitness, the selection at any one locus is very weak.

Now suppose these genotypes are distributed in a large, continuous population. Because selection is pretty weak and because mating is primarily with close neighbors, allele frequency changes at each locus will be close to what they would be if the loci were neutral. The result is that the genetic correlation between individuals drops off rapidly as a function of the distance between them (Figure 4). Notice that in the simulation illustrated individuals separated by more than about 20 distance units are effectively uncorrelated. That means that their genotypes are essentially random with respect to one another, even though their phenotypes are similar because of the stabilizing selection.

Now think about what that means for polygenic scores. Imagine that we sampled two

ends of a large, continuously distributed population. To make things concrete, let's imagine that the population is distributed primarily North-South so that our samples come from a northern population and a southern one. Now imagine that we've done a GWAS in the northern population and we want to use the genomic predictions from that population to predict phenotypes in the southern population. What's going to happen?

The genotypes in the southern population will be a random sample from all of the possible genotypes that could produce the same optimal phenotype (or something close to the optimum) and that sample will be independent of the sample of genotypes represented in our northern population. As a result, there are sure to be loci that are useful for predicting phenotype in the northern population that aren't variable in the southern population, which will reduce the accuracy of our genomic prediction. That's precisely what Yair and Coop show.[8]

In short, it's to be expected that genomic predictions will be useful only within the population for which they are constructed. They can be very useful in plant and animal breeding, for example, but any attempt to use them in other contexts must be alert to the ways in which extrapolation from one population to another will be problematic.

References

[1] Jeremy J Berg, Arbel Harpak, Nicholas Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan A Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K Pritchard, and Graham Coop. Reduced signal for polygenic adaptation of height in UK Biobank. *bioRxiv*, pages 1–54, December 2018.

[2] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics*, 41:334–341, 2009.

[3] David B. Goldstein and Kent E. Holsinger. Maintenance of polygenic variation in spatially structured populations: roles for local mating and genetic redundancy. *Evolution*, 46(2):412–429, 1992.

[4] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R. Wood, Robert J. Weyant, Ayellet V. Segrè, Elizabeth K. Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian

[8] Although their results go much farther than Goldstein and Holsinger who did their simulations long before anyone was thinking about GWAS, much less genomic prediction and polygenic scores.

Yang, Daniel Gudbjartsson, Nancy L. Heard-Costa, Joshua C. Randall, Lu Qi, Albert Vernon Smith, Reedik Mägi, Tomi Pastinen, Liming Liang, Iris M. Heid, Jian'an Luan, Gudmar Thorleifsson, Thomas W. Winkler, Michael E. Goddard, Ken Sin Lo, Cameron Palmer, Tsegaselassie Workalemahu, Yurii S. Aulchenko, Åsa Johansson, M. Carola Zillikens, Mary F. Feitosa, Tõnu Esko, Toby Johnson, Shamika Ketkar, Peter Kraft, Massimo Mangino, Inga Prokopenko, Devin Absher, Eva Albrecht, Florian Ernst, Nicole L. Glazer, Caroline Hayward, Jouke-Jan Hottenga, Kevin B. Jacobs, Joshua W. Knowles, Zoltán Kutalik, Keri L. Monda, Ozren Polasek, Michael Preuss, Nigel W. Rayner, Neil R. Robertson, Valgerdur Steinthorsdottir, Jonathan P. Tyrer, Benjamin F. Voight, Fredrik Wiklund, Jianfeng Xu, Jing Hua Zhao, Dale R. Nyholt, Niina Pellikka, Markus Perola, John R. B. Perry, Ida Surakka, Mari-Liis Tammesoo, Elizabeth L. Altmaier, Najaf Amin, Thor Aspelund, Tushar Bhangale, Gabrielle Boucher, Daniel I. Chasman, Constance Chen, Lachlan Coin, Matthew N. Cooper, Anna L. Dixon, Quince Gibson, Elin Grundberg, Ke Hao, M. Juhani Junttila, Lee M. Kaplan, Johannes Kettunen, Inke R. König, Tony Kwan, Robert W. Lawrence, Douglas F. Levinson, Mattias Lorentzon, Barbara McKnight, Andrew P. Morris, Martina Müller, Julius Suh Ngwa, Shaun Purcell, Suzanne Rafelt, Rany M. Salem, Erika Salvi, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467:832, 2010.

[5] Matthew R. Nelson, Katarzyna Bryc, Karen S. King, Amit Indap, Adam R. Boyko, John Novembre, Linda P. Briley, Yuka Maruyama, Dawn M. Waterworth, Gérard Waeber, Peter Vollenweider, Jorge R. Oksenberg, Stephen L. Hauser, Heide A. Stirnadel, Jaspal S. Kooner, John C. Chambers, Brendan Jones, Vincent Mooser, Carlos D. Bustamante, Allen D. Roses, Daniel K. Burns, Margaret G. Ehm, and Eric H. Lai. The population reference sample, popres: A resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347–358, 2008.

[6] Michael C. Turchin, Charleston W. K. Chiang, Cameron D. Palmer, Sriram Sankararaman, David Reich, Joel N. Hirschhorn, and ANthropometric Traits Consortium Genetic Investigation of. Evidence of widespread selection on standing variation in europe at height-associated snps. *Nature Genetics*, 44(9):1015–1019, 2012. (GIANT).

[7] Sivan Yair and Graham Coop. Population differentiation of polygenic score predictions under stabilizing selection. *bioRxiv*, page 2021.09.10.459833, 2021.

Creative Commons License