Estimating Prevalence of Human Traits Among Populations From Polygenic Risk Scores

Britney Graham
CWRU: Case Western Reserve University
Brian Plotkin
CWRU: Case Western Reserve University
Louis Muglia
Burroughs Wellcome Fund
Jason Moore
University of Pennsylvania
Scott Williams (smw154@case.edu)
Case Western Reserve University

Research Article

Keywords: Disease prevalence, PRS transferability, Universal risk variants, Genetic architecture

Posted Date: March 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-341766/v1

Version of Record: A version of this preprint was published at Human Genomics on December 1st, 2021.
See the published version at https://doi.org/10.1186/s40246-021-00370-z.

Estimating prevalence of human traits among populations from polygenic risk scores

Britney E. Graham[1,2], Brian Plotkin[1], Louis Muglia[3], Jason H. Moore[4], Scott M. Williams[1]

[1] Departments of Population and Quantitative Health Sciences and Genetics and Genome Scenes, Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio 44106, USA

[2]Systems Biology and Bioinformatics, Case Western Reserve University, Cleveland, Ohio 44106, USA

[3] Burroughs Wellcome Fund, Research Triangle Park, NC 27614 and University of Cincinnati College of Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229

[4]Department of Biostatistics, Epidemiology, & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Correspondence:

Scott M. Williams, PhD

Department of Population and Quantitative Health Sciences

Case Western Reserve University, Cleveland, Ohio 44106, USA

ORCID: 0000-0002-4835-9544

Email: smw154@case.edu

Keywords: Disease prevalence; PRS transferability; Universal risk variants; Genetic architecture

Abstract:

The genetic basis of phenotypic variation across populations has not been well explained for most traits. Several factors may cause disparities, from variation in environments to divergent population genetic structure. We hypothesized that a population level polygenic risk score (PRS) can explain phenotypic variation among geographic populations based solely on risk allele frequencies. We applied a population specific PRS (psPRS) to 26 populations from the 1000 Genomes to four phenotypes: lactase persistence (LP), melanoma, multiple sclerosis (MS) and height. Our models assumed additive genetic architecture among the polymorphisms in the psPRS, as is convention. Linear psPRSs explained a significant proportion of trait variance ranging from 0.32 for height in men to 0.88 for melanoma. The best models for LP and height were linear, while those for melanoma and MS were nonlinear. As not all variants in a PRS may confer similar, or even any, risk among diverse populations, we also filtered out SNPs to assess whether variance explained was improved using psPRSs with fewer SNPs. Variance explained usually improved with fewer SNPs in the psPRS and were as high as 0.99 for height in men using only 548 of the initial 4208 SNPs. That reducing SNPs improves psPRS performance may indicate that missing heritability is partially due to complex architecture that does not mandate additivity, undiscovered variants, or spurious associations in the databases. We demonstrated that PRS based-analyses can be used across diverse populations and phenotypes for population prediction and that these comparisons can identify the universal risk variants.

Introduction

The prevalence of many phenotypes differs across populations. The causes of population disparity, though not always well understood, can be partially due to different frequencies of common causative alleles that are shared among populations and/or variation in environmental exposure across these same populations. However, it is also possible that population specific alleles affect prevalence. One way to increase our understanding of a trait's genetic architecture and population differences in disease prevalences is to determine if variants associated with risk in one or a few populations can be extrapolated to the phenotypic burden for other populations across the world. For example, some variants that are extremely common in some populations are very rare in others despite having large phenotypic effects (Jorde and Wooding 2004). It has been suggested that most heritability can be explained by variants associated with a specific phenotype that are not in the group of "core" variants thought to affect trait characteristics (Boyle et al. 2017). These "core" variants may, however, not necessarily be the same as variants determined to be most statistically associated, although there may be overlap. Examining variants that do or do not transfer among populations may help elucidate the concept of "core" genes.

Models of genetic architecture often assume that the effects of a trait's genetic components are additive, without interaction, and highly similar across populations. The assumption of additivity disregards potential complexity but can be implicitly tested by assessing how well a genetic model explains the genotype to phenotype relationship (Greene et al. 2009). One additive model used to predict phenotypic status is the Polygenic risk score (PRS) that can be somewhat informative in elucidating an individual's risk of a specific phenotype (Khera et al. 2018; Torkamani et al. 2018) A PRS is the sum of the known risk associated loci of a phenotype

based on the presence of the risk alleles in an individual and the average effects of these alleles. Commonly, calculated PRSs also assume that a trait's genetic architecture is additive and that neither gene-gene nor gene-environment interactions are important factors. This method has seen some success, but often fails to predict an individual's disease status, especially at intermediate values of the PRS (Igo et al. 2019), possibly because translating population level data to individual status is problematic and risks falling into the ecological fallacy. PRSs may also not be comparable among populations (Martin et al. 2017). In addition, PRSs may not predict individual risk across populations well because many of the variants and their effect sizes are derived from a limited ancestral group, as most reaserch is done in European populations (Sirugo et al. 2019b). This may lead to a lack of PRS transferablity across populations (Martin et al. 2017). Nonetheless, when a trait is multigenic or polygenic, a Polygenic Risk Score is becoming an often used risk estimator. The role of the PRS to estimate prevelances among populations has not been explored as much as for individual risk but it may point to key factors that are common. One study on height in admixed European/African populations found that the prediction ability of a Polygenic Risk Score (PRS) for height was a function of the amount of European ancestry, supporting the idea that population specific effect sizes and allelle frequencies are important to its utility (Bitarello and Mathieson 2020). Another study by Evans et al. used individual PRSs to estimate population level disease prevalence(Evans et al. 2020), but the idea of using PRSs at a population level remains novel.

Diseases vary widely in complexity. Under simple architecture (i.e. additivity), a PRS weighted by allele frequencies at the population level should enable prediction of relative disease prevalence among populations, if the heritability is high and risk alleles are common among populations(Visscher et al. 2008). Therefore, in theory, we should be able to predict ranking of

disease prevalence for simple traits, for which our understanding is relatively comprehensive with respect to the number and effect of risk loci. This will, presumably, allow us to predict population prevalence based on the correlation between prevalence and risk allele frequency using a weighted risk score and assuming locus additivity. We hypothesize that risk allele frequencies as compiled into a PRS are proportional to population prevalence and the change in prevalence based on specific variants is proportional to their importance in disease presentation, i.e., effect size, among populations. In this paper, we assessed the ability of PRSs, in traits of varying presumed genetic complexity, to explain population differences in prevalence and to assess whether the components in a PRS act additively in their contribution to disease prevalence. We also examined whether SNPs have universal effects by adjusting the number in each PRS.

Materials and Methods

Phenotypes

As proof of principle, we explored four phenotypes of differing presumed genetic complexity: lactase persistence, melanoma, multiple sclerosis, and height. As lactase persistence is monogenic, albeit with allelic heterogeneity, it is a genetically simple trait. Melanoma is dependent on both environment and a small number of known loci and is therefore likely oligogenic. Multiple sclerosis is a presumably moderately complex polygenic trait, with hundreds of associated alleles and several environmental factors. Height is a highly complex and heritable phenotype with thousands of associating alleles, making it essentially omnigenic.

Lactase Persistence

Lactase persistence into adulthood is a monogenic autosomal dominant trait caused by one or more of several mutations affecting the expression of lactase (*LCT*), the gene responsible for the encoding of lactase. Lactase persistence is reasonably well understood genetically in some, but not all, populations. Lactase is the enzyme that our bodies produce to help breakdown lactose, the sugar found in milk. The production of lactase usually decreases after weaning, in some cases leading to an intolerance of lactose. Lactase persistence shows strong evidence of selection, although why and when is a matter of debate(Gerbault et al. 2011; Plantinga et al. 2012; Segurel and Bon 2017; Segurel et al. 2020). It is, however, believed to be associated with the advent of dairy farming. Individuals who are lactose intolerant can often consume a moderate amount of dairy, especially if processed into foods such as cheese and yogurt.

In Europe, two alleles upstream of the *LCT* gene, −13910*T (rs4988235) and −22018*G (rs182549), have been identified as conferring lactase persistence. In populations outside of Europe, other alleles have been associated with lactase persistence, where it exists(Jones et al. 2013; Liebert et al. 2017; Ranciaro et al. 2014; Tishkoff et al. 2007). A total of 11 SNPs has been associated with lactase persistence (Table S1). The prevalence of lactase persistence varies among populations around the world. For example, 92% of people in Great Britain are lactase persistent, whereas, in Vietnam the prevalence is only 2% (Table 1). We tested the expected relative frequency of lactase persistence based on a PRS, including all of the variants known, to date, to see if we could predict relative prevalences, especially in populations that appear to carry the less penetrant alleles.

Melanoma

A moderately complex oligogenic disease with 39 associated GWAS SNPs (Table S2), melanoma is a skin cancer that is both heritable and dependent, to an extent, on environmental factors, especially sun exposure. Although considered rare, melanoma is responsible for most skin cancer deaths and the incidence is increasing, due partially to improved diagnosis(Chang et al. 2014). Most cases of melanoma are caused by somatic mutations from exposure to ultraviolet light, although the above noted germline variants have been identified as conferring risk.

There is significant variation in melanoma prevalence globally, with the lowest rate in Vietnam and highest in Finland (Table 1). As melanin is protective, melanoma is higher in prevalence in populations of lighter skin color. However, non-European populations have a higher risk of mortality, possibly because melanoma is harder to detect in darker skin, and detection and treatment is late in the course of the disease(Dimitriou et al. 2018). There is some indication, also, that skin color modifies the genetic architecture of melanoma(Hulur et al. 2017).

The heritability of melanoma ranges from 19% to 58% (Lu et al. 2014; Mucci et al. 2016; Shekar et al. 2009). However, while known melanoma predisposing genes range in penetrance and frequency, the heritability in families explained by known genes is still only ~50%, indicating missing heritability and uncertain genetic architecture(Read et al. 2016).

Multiple Sclerosis

Multiple sclerosis (MS) is an autoimmune neurologic disorder affecting the central nervous system. It is a relatively complex phenotype, dependent on both environmental exposures and genetics. Environmental factors include past Epstein-Barr virus infection, vitamin D insufficiency(Pierrot-Deseilligny and Souberbielle 2013, 2017) and cigarette smoking. MS

also has a "latitude-gradient effect", i.e. the prevalence of MS is greater at higher latitudes, but there are some exceptions within Italy and Scandinavia(Simpson et al. 2011).

372 SNPs have been identified by GWAS as associating with MS (Table S3). Estimates of both prevalence and heritability vary among studies. MS is more common in women (70%-75% of cases)(Schwendimann and Alekseeva 2007) and people of European descent(Milo and Kahana 2010). Studies vary on the heritability of MS, one done in Australia, multiple European countries, and US states, showing moderate heritability (~20%)(International Multiple Sclerosis Genetics Consortium. Electronic address and International Multiple Sclerosis Genetics 2018), although a Swedish study showed a much higher heritability of 64% (36%–76%)(Westerlind et al. 2014).

Height

As a truly polygenetic trait, human height is both complex and highly heritable(Lango Allen et al. 2010; Lettre 2011). In addition to the 4,388 variants currently found to associate with this phenotype by GWAS, height is also dependent on environmental factors, including diet (Table S4)(Yeboah 2017). There are also differences in average height between men and women and between global populations. The average height for men ranges from 163.8 cm in Bangladesh to 179.6 cm in Finland. For women, average height ranges from 150.8 to 165.9, also in Bangladesh and Finland, respectively (Table 1). Height is less heritable in women than men (0.68 to 0.84 vs. 0.87 to 0.93, respectively)(Silventoinen et al. 2003). Male and female population average heights are highly, but not completely, correlated (r^2=0.84), potentially leading to some differences in the genetic models between sexes.

Allele and Prevalence Data Collection

Associated alleles for each phenotype were identified by a literature search and accessing the alleles that have been identified by GWAS from the GWAS catalog at $p < 1 \times 10^{-5}$. We chose to use this as the threshold for significance in our initial analyses, but report difference by p-value threshold as well. Prevalence data for each phenotype in each population came, similarly, from literature searches and from databases devoted to specific traits (cancer, height). For lactase persistence, it was necessary to subtract the proportion of lactose intolerance in a population from 1. An attempt was made to keep the sources as similar as possible for each population (Table 1).

1000 Genomes

To assess the role of PRS in predicting population phenotype distributions we chose to use only the populations included in The International Genome Sample Resource (IGSR) from the 1000 Genomes Project (Table 2) as our populations to study. Each ethnic population in the IGSR belongs to a larger super-population defined as: East Asian (EAS), South Asian (SAS), European (EUR), African (AFR) and Ad Mixed American (AMR). The allele frequencies of known risk alleles defined in the GWAS catalogue and literature were extracted from the 1000 genomes data using the Ensembl REST API.

Polygenic risk scores

Under the assumption that the genetic architecture of a phenotype is additive, we used a PRS to account for the genetic risk in each of our study populations, based on the frequency of the disease-causing alleles to estimate the relative presence of the phenotype in that population. As previously mentioned, in individuals this is done by simply summing the number of risk alleles that an individual possesses, usually GWAS hits, for the specific phenotype. Another approach is to weight each allele in the score by the effect size and/or the allele frequency.

However, for a population specific PRS (psPRS), effect sizes may not be transferable(Sirugo et al. 2019a) and as long as the direction of effect is the same, the role that any variant plays in prevalence should be proportional to the frequency of the risk allele in that population. We have structured psPRS without effect size weighting, as there is often little to no information on effect sizes/OR of the risk alleles in different populations. Therefore, we calculated our psPRS only by the population allele frequencies. In addition, many of the associating SNPs do not have reported effect sizes in the data sources available. Our expression for the psPRS is simply the sum of the frequencies for the risk alleles in each population. For a population in the 1000 Genomes database, psPRS is the PRS for that population and F_i is the allele frequency of SNP_i:

$$psPRS = \sum_{1}^{i} F_i$$

We then performed a linear regression for each phenotype to establish the relationship between the population specific psPRS and the population prevalence of that phenotype. (Table S5).

Maximization of the coefficient of determination sensitivity analysis

We performed a sensitivity analysis, filtering SNPs based on maximizing the coefficient of determination (r^2), or the square of the coefficient of correlation (r). This method prioritizes the SNPs that shift the populations closer to the regression line. To identify the alleles that make the relationship between the population psPRS and the phenotypic prevalence or trait mean the strongest, we maximized the coefficient of determination (r^2). This was done assessing the effect of removing SNPs from the psPRS and ordering each SNP by the r^2 value calculated for the linear regression between the population PRS without that SNP and the population prevalence.

We then permanently removed SNPs that resulted in the model with the remaining SNPs having the largest r^2. We then recalculated the r^2 values for the model with only the remaining SNPs (Tables S1-S4, S6 and Figures S6-S10). We repeated this process until the r^2 value reached a maximum. Under the assumption of additivity, the model with the largest r^2 was expected to include all truly associating SNPs with universal effects (Table 3). Our approach tested this implicitly.

Results

Lactase Persistence

We identified 11 SNPs associated with lactase persistence in the literature (Table S1). We used these SNPs to build our LP PRS for each population, using allele frequencies from the 1000 genomes project. We found a strong relationship between the PRS and the population prevalence of lactase persistence with a r^2 value of 0.65 (Figure 1A, p-value: 1.84×10^{-06}).

The relationship was especially strong amongst European populations, but less so for South Asian and Amerindian populations, However, in East Asian and African populations, the PRS failed to account for much, if any, of the relationship between the known lactase persistence alleles and the population prevalence (Figure S1A). Our sensitivity analysis (Figure S6, Table S1) based on r^2 maximization showed that keeping only 4 specific SNPs (Table S7) maximized the r^2 ($r^2 = 0.67$, p-value: 9.13×10^{-07}) and, although the r^2 did not increase by much (0.65 – 0.67), the slope of the linear regression changed from 0.45 to 0.92 (Figure 1B). The position of the populations that had high European allele content changed quite a bit, as one of the alleles that was filtered out of the PRS was one of the two alleles originally identified in Europeans. This is not a surprise, because the European alleles are in strong LD (in the European populations), and of the alleles tested these are the only ones in LD in Europe.

Within super populations, the relationships varied considerably. In the African subpopulations, the trend of the linear regression was slightly negative before maximization but slightly positive after (Figure S1B). The admixed African American population (ASW) has the highest PRS, but relatively low prevalence of LP (25%). This is due to the presence of the European alleles in the ASW population that are not present in West Africa. In the east Asian populations (EAS), the trend is also negative, but after the maximization of r^2, there were no SNPs retained that existed in the EAS populations. In the European populations (EUR), the trend was positive and stayed positive after maximization. In the south Asian populations (SAS), the trend was again positive and stayed so after maximization.

Melanoma

37 of the 39 GWAS SNPs were also in the 1000 Genomes Project (Table S2). The relationship of the melanoma PRS with these 37 associating SNPs to the population prevalence appears to be nonlinear (Figure 2A). We applied three different types of regression: linear, polynomial and exponential. The one that explained the relationship the best was the second order polynomial regression ($r^2 = 0.78$, p-value: 2.19×10^{-07}); the exponential model was next best ($r^2 = 0.66$, p-value 2.7×10^{-06}) and the linear the worst ($r^2 = 0.59$; p-value: 1.71×10^{-05}), although all were significant. The overall relationship of the psPRSs and the population prevalences reflects the fact that the highest prevalence and psPRSs are in European populations. East Asian populations had the lowest PRSs and prevalence. South Asian populations clustered with some Amerindian populations with low to medium PRSs. African populations had medium PRSs, but low melanoma prevalence.

With the 16 SNPs that remained after the maximization analysis (Table S8, Figure S7), the relationship between the melanoma population PRS and the population prevalence appeared

to remain nonlinear, similar to the original model, but with an improved explanation of variance and significance (linear regression: $r^2 = 0.88$, p-value: 2.81×10^{-11}) (Figure 2B). We also explored both polynomial ($r^2 = 0.94$, p-value: 7.36×10^{-13}) and exponential relationships ($r^2 = 0.77$, p-value: 3.39×10^{-08}). These models all performed better than the full PRS model.

When we separated populations according to their super populations, we observed that, apart from the Asian populations, the correlations weere positive, but of varying strength (Figure S2A). However, none of the relationships were significant, perhaps due to the relatively small sample size. These results indicate that the significant correlation is driven by relationship among the continental populations that are not identical to each other. After maximization, the positive and negative trends were as described above, with the Asian populations staying negative and the EUR, AFR, and AMR remaining positive (Figure S2B). The correlations did not improve substantially within super populations and remained non-significant using the reduced number of SNPs.

Multiple Sclerosis

For the full psPRS-prevalence multiple sclerosis model, we used 368 SNPs associated with MS that were in both the GWAS catalog and the 1000 Genomes project (Table S3). The resulting relationship appears to be nonlinear (Figure 3A). We explored three different models for the regression: linear, polynomial and exponential. As with melanoma, the model that explained the largest proportion of the variance was the second order polynomial ($r^2 = 0.80$, p-value: 3.94×10^{-08}). The worst was the linear model ($r^2 = 0.47$, p-value: 2.12×10^{-04}), while the exponential model was intermediate ($r^2 = 0.64$, p-value: 2.59×10^{-06}).

After the r^2 maximization sensitivity analysis (Table S9), the filtered PRS model included 131 SNPs and appears to be best modeled linearly (Figure 3B, $r^2 = 0.98$, p-value: 9.9×10^{-11}).

The linearity remains, even when the European populations are removed. Within the super populations, the prevalences and psPRSs become more highly correlated and the relationships, apart from the South Asian populations, are significant (Figure S3A).

The super populations clustered, with the European populations having the highest prevalence and psPRSs. The African populations had the lowest PRSs and prevalences, with the east and south Asian mixed with the admixed Amerindian with medium prevalences and PRSs. When the super populations were examined individually, the linear correlations were all positive, with strengths ranging from EAS ($r^2 = 0.0459$) to AFR ($r^2 = 0.4336$). However, again, none of these relationships were significant (Figure S3B).

Height

Because height has quite different ranges for men (~164 cm - ~180 cm) and women (~151 cm - ~166 cm) (Table 1), we examined the relationship between population average height and population PRS in each sex separately. The full PRS- population average height model included 4208 SNPs from the GWAS catalog and the 1000 genomes project for both men and women (Table S4). The relationships for both male and female between the population PRS and the population average height (cm) appear to be linear (Figure 4A and 4B). However, the regressions for men and women are different, with noticeable differences in the slopes of the regression lines, the correlations and the significance of the relationships (male: $r^2 = 0.32$, P-value: 2.55×10^{-03}; female: $r^2 = 0.11$, P-value: 0.0992).

The populations generally clustered by super populations, with European populations being both the tallest and having the largest psPRSs for both men and women. The south Asian and Amerindian were the shortest groups, but with medium PRSs. African and east Asian populations had medium to tall height, but the lowest PRSs.

Within the African super population, the relationship between average height and population PRS was positive in both males and females. Both south Asian and Amerindian populations had positive relationships as well. However, surprisingly the European and east Asian populations had negative relationships (Figures S4A and S5A).

The sensitivity analysis reduced the number of SNPs for the male model to 548 and for the female model to 188 (Figures 4C and 4D, S9-S10, Tables S6, S10-S11). The reduced male and female linear models changed substantially (male: slope from 0.06 to 3.92; female: slope from 0.03 to 3.86). The correlation strengthened for both male and female (male: $r^2 = 0.99$;

female: 0.98) and in males the relationship became more significant and became significant in females (male and female: P-value: $<2x10^{-16}$).

After the maximization filtering, the positions of the populations shifted significantly. The south Asians had lower PRSs to match their lower average height. Europeans still had the highest psPRSs and the African and east Asian populations were mixed. Within the super populations, the relationships all became positive for both men and women (Figures S4B and S5B).

Effect of P-value thresholds for SNP selections

As we used only a moderately stringent threshold for the SNPs from the GWAS catalog, we wished to know if the maximization analysis selected SNPs that were statistically significant, i.e., with p-values of genome wide significance. We found, using the Fisher's exact test, that there was no significant enrichment of GWAS SNPs with a p-value less than 5×10^{-8}, except for height in women (Table 4).

Discussion

Overall, our psPRS method estimated population prevalence quite well. This indicates that the population PRS is a reasonably good indicator of disease presence in a population. For lactase persistence, we found that the psPRSs and the prevalence were strongly correlated, even before SNP filtering. For melanoma and MS, we also found strong correlations, albeit non-linear ones. However, for height the correlations, while linear, were weaker. As expected, the complexity of the phenotype did affect the ability of the full PRS model to predict the population prevalence, sometimes being far from what would have been expected and being non-linear, i.e., melanoma and MS. Also, the SNP pruned models improved the explained variance over the complete psPRSs, sometimes substantially, and the relationships achieved or approached

linearity when the complete models were not. Although this can be viewed as “cherry picking”, it does reveal that not all detected SNPs have similar effects across populations and that some may reflect effects that are universal as opposed to population specific. Our results show that the European populations often skew the overall full model and that, with the exception of height, fit the PRS predictions best. This is not surprising as most of the SNPs were discovered in populations of European descent(Sirugo et al. 2019b). We also repeatedly observed that there were not as many significant correlations within the super populations, but there were between super populations, which may reflect the paucity of data within them.

Generally, the model of LP followed what was expected, as it is a monogenic disease. For LP in the African populations, the disparity between the observed prevalence in some populations and our psPRS model, shows that our ability to predict prevalence is likely impacted by unidentified associating alleles or other mechanisms by which lactose is digested, perhaps acquired gut microbiome activity(Goodrich et al. 2017). This is supported by the negative and weak relationship in the full data set, although likely impacted by the admixed ASW population, where the European alleles exist but do not seem to confer lactase persistence to the extent that the psPRS would predict; nonetheless, this African descent population still had the highest prevalence of lactase persistence. Another possible reason for the psPRS not predicting prevalence in Africa well is that there may be context dependent effects. For example, it has been found that the 13915*G DNA polymorphism that is associated with lactase persistence in Africa interacts with *Oct-1*(Olds et al. 2011).

Given the known impact of environmental exposure on the development of melanoma(Dimitriou et al. 2018), the observed nonlinearity of the relationship between the population PRSs and the prevalence of the disease was not unexpected. That the nonlinearity

continued after the filtering, implies that the actual relationship between the psPRS and the prevalence may be non-additive, and that we are missing key factors, either genetic interactions, environmental interactions, or both. Because we did not consider environmental factors in this study, we were not able to differentiate between the two. It has, however, been shown that at least one SNP pair at the *TERF1* and the *AFAP1L2* loci does interact to affect risk of melanoma(Brossard et al. 2015).

While the relationship observed with the full MS PRS model was nonlinear, the model after filtering of SNPs resulted in a strong linear model. This might indicate that there is some genetic interaction in MS, especially given that the r^2 improves as we drop SNPs from the psPRS model. Indeed, a *DDX39B* variant interacts with allelic variants in *IL7R* exon 6 to increase MS risk(Galarza-Munoz et al. 2017). Interaction with environmental factors has also been shown. Specifically, latitude, EBV infection, smoking and adolescent obesity interact with risk alleles at the *HLA* locus to increase risk of MS(Olsson et al. 2017).

While the relationship of the psPRS to population average height with the full model shows a relatively weak, though significant, positive correlation, the result of the maximization shows a very strongly correlated relationship. Although height is highly heritable, this was not expected given the foreknowledge of the impact of environment on height, especially in women(Silventoinen et al. 2003). It may be that some of the variants left in the final model are correlated with environmental parameters due to past selection. Also, there may be epistasis in the genetic architecture of height. For example, genetic interaction was found between loci 6p21 and 2q21 to account some of the variation in height(Liu et al. 2006).

We infer from our results that the maximizing r^2 sensitivity analysis is filtering out the SNPs that are not distributed as the population prevalence distribution of the phenotype in some

but maybe not all populations. This is, in effect, similar to a previous method Evolutionary Triangulation(Huang et al. 2016), where we filtered SNPs based specifically on their distribution relative to disease prevalence. Our results showing that pruning the SNPs in the model improves performance may be revealing heterogenous effect sizes that may present due to context dependent effects, such as epistasis or gene X environment interactions, spurious associations, or other population specific effects. This may explain why a filtered model is superior in some cases to a model with all associating SNPs included. This approach is, in essence, removing noisy data. psPRSs provide some explanation of population differences but are less effective when all SNPs are included. This indicates that PRSs have value but must be refined to improve prediction.

Our investigation as to whether GWAS significance was a useful threshold for inclusion indicated that it was not good at predicting which SNPs would end up in our pruned SNP set. As shown by our investigation of whether our model enriched for SNPs with a smaller p-value in the GWAS Catalog, we can conclude that GWAS p-value is not always the best indicator of the value of a SNP in the PRS model. This does justify, to some extent, our use of SNPs that were not genome wide significant at 5×10^{-08} and indicates that some care should be used in determining the importance of SNPs in models based solely on significance of p-values.

Understanding the relationship between allele frequency and disease prevalence will lead to further understanding of genetic influence, environmental pressure and gene-environment interactions. The effects of genetic variation on public health presents challenges for the exploration and management of these phenotypes worldwide, as most traits are primarily considered in the context of European descent. This blind spot, due partially to a lack of diversity in biomedical research, is not only detrimental to those populations that are understudied, but to

the understanding of the underlying genetic basis, or genetic architecture, of the trait itself, thereby, possibly affecting understanding in all populations. Nonetheless, some of our results indicate that even SNPs discovered primarily in Europeans are useful, when included in a psPRS, for predicting trait variation, e.g., height.

Our results help to identify the populations in which we are missing the most information regarding genetic foundations of trait variation. This is underlined by some of our results where the population PRSs do not match the population prevalences, i.e. where the prevalence is high or medium and the psPRS is low, as in the cases of height and LP in African populations. That using a reduced number of SNPs improves the psPRS likely indicates a certain portion of missing heritability is due to more complex architecture, i.e., genetic interaction, possibly differing by population and that there are still undiscovered variants. However, our method helps to define the areas of the genetic landscape where our knowledge of genetic architecture is relatively complete and where it is not.

Declarations

Competing Interests. There are no competing interests.

Funding. This works was supported by the March of Dimes Ohio Collaborative Prematurity Research Center and National Institutes of Health grant R01 LM010098.

Author Contributions. BG, LM, JHM and SMW conceived the idea; BEG, LM and SMW designed the research; BEG, BP and SMW performed research; BEG and SMW carried out statistical analysis; BEG and SMW wrote the manuscript.

Conflict of Interest. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data Availability. All data is publicly available at sites described in the manuscript.

Animal Research. Not applicable

Consent to Participate. Not applicable

Consent to Publish. Not applicable

Figure 1. The correlation between lactase persistence and psGRS. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model ($r^2 = 0.65$; p-value: 1.84×10^{-06}). B) After maximization ($r^2 = 0.67$, p-value: 9.13×10^{-07}).

Figure 2. The correlation between melanoma and psGRS with regression lines for linear (red), polynomial (blue) and exponential (green) relationships. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model with three regressions: polynomial ($r^2 = 0.78$, p-value: 2.19×10^{-07}), exponential (r2 = 0.66, p-value 2.7×10^{-06}) and linear ($r^2 = 0.59$; p-value: 1.71×10^{-05}). B) After maximization: linear regression ($r^2 = 0.88$, p-value: 2.81×10^{-11}), polynomial (r2= 0.94, p-value: 7.36×10^{-13}) and exponential (r2 = 0.77, p-value: 3.39×10^{-08}).

Figure 3. The correlation between multiple sclerosis and psGRS with linear (red), polynomial (blue) and exponential (green) regressions lines. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model with three regression types: polynomial ($r^2 = 0.80$, p-value: 3.94×10^{-08}), linear, ($r^2 = 0.47$, p-value: 2.12×10^{-04}) and exponential ($r^2 = 0.64$, p-value: 2.59×10^{-06}). B) Linear regression after maximization ($r^2 = 0.98$, p-value: 9.9×10^{-11}).

Figure 4. The correlation between height (cm) and psGRS. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model male linear regression ($r^2 = 0.32$, P-value: 2.55×10^{-03}). B) Full model female linear regression ($r^2 = 0.11$, p-value: 0.0992). C) Male linear regression after maximization ($r^2 = 0.99$, P-value: $<2 \times 10^{-16}$). D) Female linear regression after maximization (r^2.98, P-value: $<2 \times 10^{-16}$).

Appendix

Supplemental Data

The supplemental data includes ten figures and eleven tables.

Figure S1. Lactase persistence separated by super population. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model by super population: AFR ($r^2 = 0.0021$, p-value: 0.9314), AMR ($r^2 = 0.2608$, p-value: 0.659), EAS ($r^2 = 0.077$, p-value: 0.6514), EUR ($r^2 = 0.9734$, p-value: 0.00185) and SAS ($r^2 = 0.3847$, p-value: 0.2643). B) Super populations after maximization: AFR ($r^2 = 0.0177$, p-value: 0.8017), AMR ($r^2 = 0.2284$, p-value: 0.683), EAS (no data), EUR ($r^2 = 0.9747$, p-value: 0.00172) and SAS ($r^2 = 0.3914$, p-value: 0.0580).

Figure S2. Melanoma separated by super population. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model by super population: AFR ($r^2 = 0.1178$, p-value: 0.5718), AMR ($r^2 = 0.6664$, p-value: 0.1837), EAS ($r^2 = 0.4958$, p-value: 0.1844), EUR ($r^2 = 0.0421$, p-value: 0.7949) and SAS ($r^2 = 0.5914$, p-value: 0.1285). B) Super populations after maximization: AFR ($r^2 = 0.1767$, p-value: 0.481), AMR ($r^2 = 0.7766$, p-value: 0.1187), EAS ($r^2 = 0.2399$, p-value: 0.4022), EUR ($r^2 = 0.6268$, p-value: 0.2083) and SAS ($r^2 = 0.4324$, p-value: 0.2278).

Figure S3. Multiple sclerosis separated by super population. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model super populations: AFR ($r^2 = 0.4336$, p-value: 0.155), AMR ($r^2 = 0.1958$, p-value: 0.5575), EAS ($r^2 = 0.0459$, p-value: 0.7293), EUR ($r^2 = 0.3676$, p-value: 0.3937) and SAS ($r^2 = 0.3775$, p-value: 0.270). B) Super populations after maximization: AFR ($r^2 = 0.7781$, p-value: 0.02003), AMR ($r^2 = 0.9821$, p-value: 0.008995), EAS ($r^2 = 0.8775$, p-value: 0.0189), EUR ($r^2 = 0.9988$, p-value: 0.000617) and SAS ($r^2 = 0.2356$, p-value: 0.407) .

Figure S4. Male height separated by super population. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Super populations with full model: AFR ($r^2 = 0.7835$, p-value: 0.00806), AMR ($r^2 = 0.8628$, p-value: 0.0522), EAS ($r^2 = 0.1003$, p-value: 0.9254), EUR ($r^2 = 0.578$, p-value: 0.551) and SAS: $r^2 = 0.1162$, p-value: 0.2812. B) super populations after performing maximization: AFR ($r^2 = 0.5534$, p-value: 6.549×10^{-7}), AMR ($r^2 = 0.8556$, p-value: 0.001876), EAS ($r^2 = 0.0163$, p-value: 2.052×10^{-5}), EUR ($r^2 = 0.9158$, p-value: 0.01064) and SAS ($r^2 = 0.0475$, p-value: 0.002888).

Figure S5. Female height separated by super population. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Super populations with full model: AFR ($r^2 = 0.5534$, p-value: 0.05523), AMR ($r^2 = 0.8556$, p-value: 0.07501), EAS ($r^2 = 0.0163$, p-value: 0.8379), EUR ($r^2 = 0.2222$, p-value: 0.4229) and SAS ($r^2 = 0.0475$, p-value: 0.7246). B) Super populations after maximization: AFR ($r^2 = 0.9533$, p-value: 0.0001627), AMR ($r^2 = 0.9917$, p-value: 0.004174), EAS ($r^2 = 0.963$, p-value: 0.003058), EUR ($r^2 = 0.754$, p-value: 0.05619) and SAS ($r^2 = 0.9761$, p-value: 0.001584).

Figure S6. Lactase persistence maximization analysis r^2 values.

Figure S7. Melanoma maximization analysis r^2 values.

Figure S8. Multiple sclerosis maximization analysis r^2 values.

Figure S9. Male height maximization analysis r^2 values.

Figure S10. Female height maximization analysis r^2 values.

Table S1. Lactase persistence full data set. SNP rs number and minor allele are included, as well as the r^2 values from the sensitivity analysis. The columns headed with the 1000 Genomes population codes are the allele frequencies for each SNP in those populations. The SNPs are listed in order of removal in the sensitivity analysis.

Table S2. Melanoma full data set. SNP rs number and minor allele are included, as well as the r^2 values from the sensitivity analysis. The columns headed with the 1000 Genomes population codes are the allele frequencies for each SNP in those populations. The SNPs are listed in order of removal in the sensitivity analysis.

Table S3. Multiple sclerosis full data set. SNP rs number and minor allele are included, as well as the r^2 values from the sensitivity analysis. The columns headed with the 1000 Genomes population codes are the allele frequencies for each SNP in those populations. The SNPs are listed in order of removal in the sensitivity analysis.

Table S4. Height full data set. SNP rs number and minor allele are included, as well as the r^2 values from the sensitivity analysis. The columns headed with the 1000 Genomes population codes are the allele frequencies for each SNP in those populations. The SNPs are listed in order of removal in the sensitivity analysis.

Table S5. PRS values for each population, before and after maximization.

Table S6. Female height r^2 values from the maximization analysis.

Table S7. Lactase persistence filtered data set.

Table S8. Melanoma filtered data set.

Table S9. Multiple sclerosis filtered data set.

Table S10. Male Height filtered data set.

Table S11. Female Height filtered data set.

Declaration of Interests

None

Table 1. Relative trait distribution among populations.						
Super population	Population	Lactase Persistence	Melanoma	Multiple Sclerosis	Male Height (cm)	Female Height (cm)
AFR	ACB	-	$4.2x10^{-5}$ (Ferlay J 2020)	$1.36x10^{-4}$ (Collaborators 2019; United Nations 2019)	175.92(Collaboration 2016)	165.28(Collaboration 2016)
	ASW	0.25 (Bayless et al. 2017)	$2.9x10^{-5}$ (Group 2020)	-	175.5(Fryar et al. 2018)	162.6(Fryar et al. 2018)
	ESN	0.13 (Storhaug et al. 2017)	$5.7x10^{-6}$ (Ferlay J 2020)	$3.71x10^{-5}$ (Collaborators 2019; United Nations 2019)	165.91(Collaboration 2016)	156.32(Collaboration 2016)
	GWD	0.43	0 (Ferlay J 2020)	$3.35x10^{-5}$ (Collaborators 2019; United Nations 2019)	165.40(Collaboration 2016)	160.94(Collaboration 2016)
	LWK	0.61 (Storhaug et al. 2017)	$1.4x10^{-5}$ (Ferlay J 2020)	$3.30x10^{-5}$ (Collaborators 2019; United Nations 2019)	169.64(Collaboration 2016)	158.16(Collaboration 2016)
	MSL	0.52	$5.6x10^{-6}$ (Ferlay J 2020)	$2.89x10^{-5}$ (Collaborators 2019; United Nations 2019)	164.41(Collaboration 2016)	156.60(Collaboration 2016)
	YRI	0.13 (Storhaug et al. 2017))	$5.7x10^{-6}$ (Ferlay J 2020)	$3.71x10^{-5}$ (Collaborators 2019; United Nations 2019)	165.91(Collaboration 2016)	156.32(Collaboration 2016)

	CLM	0.2[(Storhaug et al. 2017)]	$1.08x10^{-4}$ (Ferlay J 2020)	$5.53x10^{-5}$ (Collaborators 2019; United Nations 2019)	169.50(Collaboration 2016)	156.85(Collaboration 2016)
	MXL	0.52[(Storhaug et al. 2017)]	$6.9x10^{-5}$ (Ferlay J 2020)	$1.08x10^{-4}$ (Collaborators 2019; United Nations 2019)	169.00(Collaboration 2016)	156.85(Collaboration 2016)
	Px10L	0.06[(Bayless et al. 2017)]	$8.3x10^{-5}$ (Ferlay J 2020)	$6.98x10^{-5}$ (Collaborators 2019; United Nations 2019)	165.23(Collaboration 2016)	152.93(Collaboration 2016)
AMR	PUR	-	$1.11x10^{-4}$ (Ferlay J 2020)	$1.9x10^{-4}$ (Collaborators 2019; United Nations 2019)	172.08(Collaboration 2016)	159.20(Collaboration 2016)
	CDX	0.15[(Storhaug et al. 2017)]	$1.5x10^{-5}$ (Ferlay J 2020)	$7.30x10^{-5}$[(Collaborators 2019; United Nations 2019)]	171.83(Collaboration 2016)	159.71(Collaboration 2016)
	CHB	0.15[(Storhaug et al. 2017)]	$1.5x10^{-5}$ (Ferlay J 2020)	$7.30x10^{-5}$ (Collaborators 2019; United Nations 2019)	171.83(Collaboration 2016)	159.71(Collaboration 2016)
	CHS	0.15[(Storhaug et al. 2017)]	$1.5x10^{-5}$ (Ferlay J 2020)	$7.30x10^{-5}$ (Collaborators 2019; United Nations 2019)	171.83(Collaboration 2016)	159.71(Collaboration 2016)
EAS	JPT	0.27[(Storhaug et al. 2017)]	$5.0x10^{-5}$ (Ferlay J 2020)	$3.62x10^{-4}$ (Collaborators 2019; United Nations 2019)	170.82(Collaboration 2016)	158.31(Collaboration 2016)

	KHV	0.02[(Storhaug et al. 2017)]	$4.3x10^{-6}$ (Ferlay J 2020)	$4.41x10^{-5}$ (Collaborators 2019; United Nations 2019)	164.45(Collaboration 2016)	153.59(Collaboration 2016)
EUR	CEU	0.87	$1.3x10^{-3}$ (Group 2020)	-	177.4(Fryar et al. 2018)	163.3(Fryar et al. 2018)
	FIN	0.81(Storhaug et al. 2017)	$1.04x10^{-3}$ (Ferlay J 2020)	$1.49x10^{-3}$ (Collaborators 2019; United Nations 2019)	179.59(Collaboration 2016)	165.90(Collaboration 2016)
	GBR	0.92(Storhaug et al. 2017)	$9.39x10^{-4}$ (Ferlay J 2020)	$1.61x10^{-3}$ (Collaborators 2019; United Nations 2019)	177.49(Collaboration 2016)	164.40(Collaboration 2016)
	IBS	0.71(Storhaug et al. 2017)	$3.92x10^{-4}$ (Ferlay J 2020)	$9.41x10^{-4}$ (Collaborators 2019; United Nations 2019)	176.59(Collaboration 2016)	163.40(Collaboration 2016)
	TSI	0.28(Storhaug et al. 2017)	$7.12x10^{-4}$ (Ferlay J 2020)	$1.19x10^{-3}$ (Collaborators 2019; United Nations 2019)	177.77(Collaboration 2016)	164.61(Collaboration 2016)
SAS	BEB	0.175	$6.5x10^{-6}$ (Ferlay J 2020)	$1.42x10^{-4}$ (Collaborators 2019; United Nations 2019)	163.81(Collaboration 2016)	150.79(Collaboration 2016)
	GIH	0.39(Storhaug et al. 2017)	$5.4x10^{-6}$ (Ferlay J 2020)	$1.54x10^{-4}$ (Collaborators 2019; United Nations 2019)	164.95(Collaboration 2016)	152.59(Collaboration 2016)

	ITU	0.39(Storhaug et al. 2017)	$5.4x10^{-6}$ (Ferlay J 2020)	$1.54x10^{-4}$ (Collaborators 2019; United Nations 2019)	164.95(Collaboration 2016)	152.59(Collaboration 2016)
	PJL	0.42(Storhaug et al. 2017)	$4.6x10^{-6}$ (Ferlay J 2020)	$1.46x10^{-04}$ (Collaborators 2019; United Nations 2019)	166.95(Collaboration 2016)	153.84(Collaboration 2016)
	STU	0.27(Storhaug et al. 2017)	$1.4x10^{-5}$ (Ferlay J 2020)	$3.35x10^{-05}$ (Collaborators 2019; United Nations 2019)	165.69(Collaboration 2016)	154.56(Collaboration 2016)

Table 2. 1000 Genomes populations.		
Super population	Population Code	Population Description
East Asian	CHB	Han Chinese in Beijing, China
	JPT	Japanese in Tokyo, Japan
	CHS	Southern Han Chinese
	CDX	Chinese Dai in Xishuangbanna, China
	KHV	Kinh in Ho Chi Minh City, Vietnam
European	CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
	TSI	Toscani in Italia
	FIN	Finnish in Finland
	GBR	British in England and Scotland
	IBS	Iberian Population in Spain
African	YRI	Yoruba in Ibadan, Nigeria
	LWK	Luhya in Webuye, Kenya
	GWD	Gambian in Western Divisions in the Gambia
	MSL	Mende in Sierra Leone
	ESN	Esan in Nigeria
	ASW	Americans of African Ancestry in SW USA
	ACB	African Caribbean in Barbados
Ad Mixed American	MXL	Mexican Ancestry from Los Angeles USA
	PUR	Puerto Ricans from Puerto Rico
	CLM	Colombians from Medellin, Colombia
	PEL	Peruvians from Lima, Peru
South Asian	GIH	Gujarati Indian from Houston, Texas
	PJL	Punjabi from Lahore, Pakistan
	BEB	Bengali from Bangladesh
	STU	Sri Lankan Tamil from the UK
	ITU	Indian Telugu from the UK

Table 3. Sensitivity analysis.			
Phenotype	GWAS SNPs	1000 Genomes SNPs	Reduced SNPs
Lactase Persistence	11	NA	4
Multiple Sclerosis	372	368	131
Melanoma	39	37	16
Height male	4388	4209	547
Height female	4388	4209	188

Table 4. P-value enrichment analysis.								
Phenotype	Melanoma		Multiple Sclerosis		Male Height		Female Height	
Threshold	unfiltered	filtered	unfiltered	filtered	unfiltered	filtered	unfiltered	filtered
GWAS significant[1]	27	12	199	71	3552	478	3552	188
GWAS not significant[2]	10	4	169	60	656	69	656	0
p-value	1		1		0.07644		6.73E-14	

[1] $P \leq 5 \times 10^{-8}$

[2] $1 \times 10^{-5} > P > 5 \times 10^{-8}$

References

Bayless TM, Brown E, Paige DM (2017) Lactase Non-persistence and Lactose Intolerance. Curr Gastroenterol Rep 19: 23. doi: 10.1007/s11894-017-0558-9

Bitarello BD, Mathieson I (2020) Polygenic Scores for Height in Admixed Populations. G3 (Bethesda) 10: 4027-4036. doi: 10.1534/g3.120.401658

Boyle EA, Li YI, Pritchard JK (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169: 1177-1186. doi: 10.1016/j.cell.2017.05.038

Brossard M, Fang S, Vaysse A, Wei Q, Chen WV, Mohamdi H, Maubec E, Lavielle N, Galan P, Lathrop M, Avril MF, Lee JE, Amos CI, Demenais F (2015) Integrated pathway and epistasis analysis reveals interactive effect of genetic variants at TERF1 and AFAP1L2 loci on melanoma risk. Int J Cancer 137: 1901-1909. doi: 10.1002/ijc.29570

Chang C, Murzaku EC, Penn L, Abbasi NR, Davis PD, Berwick M, Polsky D (2014) More skin, more sun, more tan, more melanoma. Am J Public Health 104: e92-9. doi: 10.2105/AJPH.2014.302185

Collaboration NCDRF (2016) A century of trends in adult human height. Elife 5. doi: 10.7554/eLife.13410

Collaborators GBDMS (2019) Global, regional, and national burden of multiple sclerosis 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 18: 269-285. doi: 10.1016/S1474-4422(18)30443-5

Dimitriou F, Krattinger R, Ramelyte E, Barysch MJ, Micaletto S, Dummer R, Goldinger SM (2018) The World of Melanoma: Epidemiologic, Genetic, and Anatomic Differences of Melanoma Across the Globe. Curr Oncol Rep 20: 87. doi: 10.1007/s11912-018-0732-8

Evans BD, Słowiński P, Hattersley AT, Jones SE, Sharp S, Kimmitt RA, Weedon MN, Oram RA, Tsaneva-Atanasova K, Thomas NJ (2020) Estimating population level disease prevalence using genetic risk scores. medRxiv: 2020.02.20.20025528. doi: 10.1101/2020.02.20.20025528

Ferlay J EM, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2020) Global Cancer Observatory: Cancer Today. International Agency for Research on Cancer. https://gco.iarc.fr/today. Accessed 22 November 2020

Fryar CD, Kruszon-Moran D, Gu Q, Ogden CL (2018) Mean Body Weight, Height, Waist Circumference, and Body Mass Index Among Adults: United States, 1999-2000 Through 2015-2016. Natl Health Stat Report: 1-16.

Galarza-Munoz G, Briggs FBS, Evsyukova I, Schott-Lerner G, Kennedy EM, Nyanhete T, Wang L, Bergamaschi L, Widen SG, Tomaras GD, Ko DC, Bradrick SS, Barcellos LF, Gregory SG, Garcia-Blanco MA (2017) Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. Cell 169: 72-84 e13. doi: 10.1016/j.cell.2017.03.007

Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG (2011) Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond B Biol Sci 366: 863-77. doi: 10.1098/rstb.2010.0268

Goodrich JK, Davenport ER, Clark AG, Ley RE (2017) The Relationship Between the Human Genome and Microbiome Comes into View. Annu Rev Genet 51: 413-433. doi: 10.1146/annurev-genet-110711-155532

Greene CS, Penrod NM, Williams SM, Moore JH (2009) Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS One 4: e5639. doi: 10.1371/journal.pone.0005639

Group USCSW (2020) U.S. Cancer Statistics Data Visualizations Tool, based on 2019 submission data (1999–2017): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. www.cdc.gov/cancer/dataviz. Accessed February 15, 2021

Huang M, Graham BE, Zhang G, Harder R, Kodaman N, Moore JH, Muglia L, Williams SM (2016) Evolutionary triangulation: informing genetic association studies with evolutionary evidence. BioData Min 9: 12. doi: 10.1186/s13040-016-0091-7

Hulur I, Skol AD, Gamazon ER, Cox NJ, Onel K (2017) Integrative genetic analysis suggests that skin color modifies the genetic architecture of melanoma. PLoS One 12: e0185730. doi: 10.1371/journal.pone.0185730

Igo RP, Jr., Kinzy TG, Cooke Bailey JN (2019) Genetic Risk Scores. Curr Protoc Hum Genet 104: e95. doi: 10.1002/cphg.95

International Multiple Sclerosis Genetics Consortium. Electronic address ccye, International Multiple Sclerosis Genetics C (2018) Low-Frequency and Rare-Coding Variation Contributes to Multiple Sclerosis Risk. Cell 175: 1679-1687 e7. doi: 10.1016/j.cell.2018.09.049

Jones BL, Raga TO, Liebert A, Zmarz P, Bekele E, Danielsen ET, Olsen AK, Bradman N, Troelsen JT, Swallow DM (2013) Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. Am J Hum Genet 93: 538-44. doi: 10.1016/j.ajhg.2013.07.008

Jorde LB, Wooding SP (2004) Genetic variation, classification and 'race'. Nat Genet 36: S28-33. doi: 10.1038/ng1435

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50: 1219-1224. doi: 10.1038/s41588-018-0183-z

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-8. doi: 10.1038/nature09410

Lettre G (2011) Recent progress in the study of the genetics of height. Hum Genet 129: 465-72. doi: 10.1007/s00439-011-0969-x

Liebert A, Lopez S, Jones BL, Montalva N, Gerbault P, Lau W, Thomas MG, Bradman N, Maniatis N, Swallow DM (2017) World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. Hum Genet 136: 1445-1453. doi: 10.1007/s00439-017-1847-y

Liu YZ, Guo YF, Xiao P, Xiong DH, Zhao LJ, Shen H, Liu YJ, Dvornyk V, Long JR, Deng HY, Li JL, Deng HW (2006) Epistasis between loci on chromosomes 2 and 6 influences human height. J Clin Endocrinol Metab 91: 3821-5. doi: 10.1210/jc.2006-0348

Lu Y, Ek WE, Whiteman D, Vaughan TL, Spurdle AB, Easton DF, Pharoah PD, Thompson DJ, Dunning AM, Hayward NK, Chenevix-Trench G, Q M, Investigators A, Anecs S, Ukops S, Consortium B, Macgregor S (2014) Most common 'sporadic' cancers have a significant germline genetic component. Hum Mol Genet 23: 6112-8. doi: 10.1093/hmg/ddu312

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE (2017) Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 100: 635-649. doi: 10.1016/j.ajhg.2017.03.004

Milo R, Kahana E (2010) Multiple sclerosis: geoepidemiology, genetics and the environment. Autoimmun Rev 9: A387-94. doi: 10.1016/j.autrev.2009.11.010

Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, Graff RE, Holst K, Moller S, Unger RH, McIntosh C, Nuttall E, Brandt I, Penney KL, Hartman M, Kraft P, Parmigiani G, Christensen K, Koskenvuo M, Holm NV, Heikkila K, Pukkala E, Skytthe A, Adami HO, Kaprio J, Nordic Twin Study of Cancer C (2016) Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA 315: 68-76. doi: 10.1001/jama.2015.17703

Olds LC, Ahn JK, Sibley E (2011) 13915*G DNA polymorphism associated with lactase persistence in Africa interacts with Oct-1. Hum Genet 129: 111-3. doi: 10.1007/s00439-010-0898-0

Olsson T, Barcellos LF, Alfredsson L (2017) Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nature Reviews Neurology 13: 25-36. doi: 10.1038/nrneurol.2016.187

Pierrot-Deseilligny C, Souberbielle JC (2013) Contribution of vitamin D insufficiency to the pathogenesis of multiple sclerosis. Ther Adv Neurol Disord 6: 81-116. doi: 10.1177/1756285612473513

Pierrot-Deseilligny C, Souberbielle JC (2017) Vitamin D and multiple sclerosis: An update. Mult Scler Relat Disord 14: 35-45. doi: 10.1016/j.msard.2017.03.014

Plantinga TS, Alonso S, Izagirre N, Hervella M, Fregel R, van der Meer JW, Netea MG, de la Rua C (2012) Low prevalence of lactase persistence in Neolithic South-West Europe. Eur J Hum Genet 20: 778-82. doi: 10.1038/ejhg.2011.254

Ranciaro A, Campbell MC, Hirbo JB, Ko WY, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, Tishkoff SA (2014) Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am J Hum Genet 94: 496-510. doi: 10.1016/j.ajhg.2014.02.009

Read J, Wadt KA, Hayward NK (2016) Melanoma genetics. J Med Genet 53: 1-14. doi: 10.1136/jmedgenet-2015-103150

Schwendimann RN, Alekseeva N (2007) Gender issues in multiple sclerosis. Int Rev Neurobiol 79: 377-92. doi: 10.1016/S0074-7742(07)79017-7

Segurel L, Bon C (2017) On the Evolution of Lactase Persistence in Humans. Annu Rev Genomics Hum Genet 18: 297-319. doi: 10.1146/annurev-genom-091416-035340

Segurel L, Guarino-Vignon P, Marchi N, Lafosse S, Laurent R, Bon C, Fabre A, Hegay T, Heyer E (2020) Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. PLoS Biol 18: e3000742. doi: 10.1371/journal.pbio.3000742

Shekar SN, Duffy DL, Youl P, Baxter AJ, Kvaskoff M, Whiteman DC, Green AC, Hughes MC, Hayward NK, Coates M, Martin NG (2009) A population-based study of Australian twins with melanoma suggests a strong genetic contribution to liability. J Invest Dermatol 129: 2211-9. doi: 10.1038/jid.2009.48

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, De Lange M, Harris JR, Hjelmborg JV, Luciano M, Martin NG, Mortensen J, Nistico L, Pedersen NL, Skytthe A, Spector TD, Stazi MA, Willemsen G, Kaprio J (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6: 399-408. doi: 10.1375/136905203770326402

Simpson S, Jr., Blizzard L, Otahal P, Van der Mei I, Taylor B (2011) Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis. J Neurol Neurosurg Psychiatry 82: 1132-41. doi: 10.1136/jnnp.2011.240432

Sirugo G, Williams SM, Tishkoff SA (2019a) The Missing Diversity in Human Genetic Studies. Cell 177: 26-31. doi: 10.1016/j.cell.2019.02.048

Sirugo G, Williams SM, Tishkoff SA (2019b) The Missing Diversity in Human Genetic Studies. Cell 177: 1080. doi: 10.1016/j.cell.2019.04.032

Storhaug CL, Fosse SK, Fadnes LT (2017) Country, regional, and global estimates for lactose malabsorption in adults: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2: 738-746. doi: 10.1016/S2468-1253(17)30154-1

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40. doi: 10.1038/ng1946

Torkamani A, Wineinger NE, Topol EJ (2018) The personal and clinical utility of polygenic risk scores. Nat Rev Genet 19: 581-590. doi: 10.1038/s41576-018-0018-x

United Nations DoEaSA, Population Division (2019) (2019) World Population Prospects 2019, Online Edition. Rev. 1.

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9: 255-66. doi: 10.1038/nrg2322

Westerlind H, Ramanujam R, Uvehag D, Kuja-Halkola R, Boman M, Bottai M, Lichtenstein P, Hillert J (2014) Modest familial risks for multiple sclerosis: a registry-based study of the population of Sweden. Brain 137: 770-8. doi: 10.1093/brain/awt356

Yeboah J (2017) Diet, height, and health. Am J Clin Nutr 106: 443-444. doi: 10.3945/ajcn.117.161562

Figures

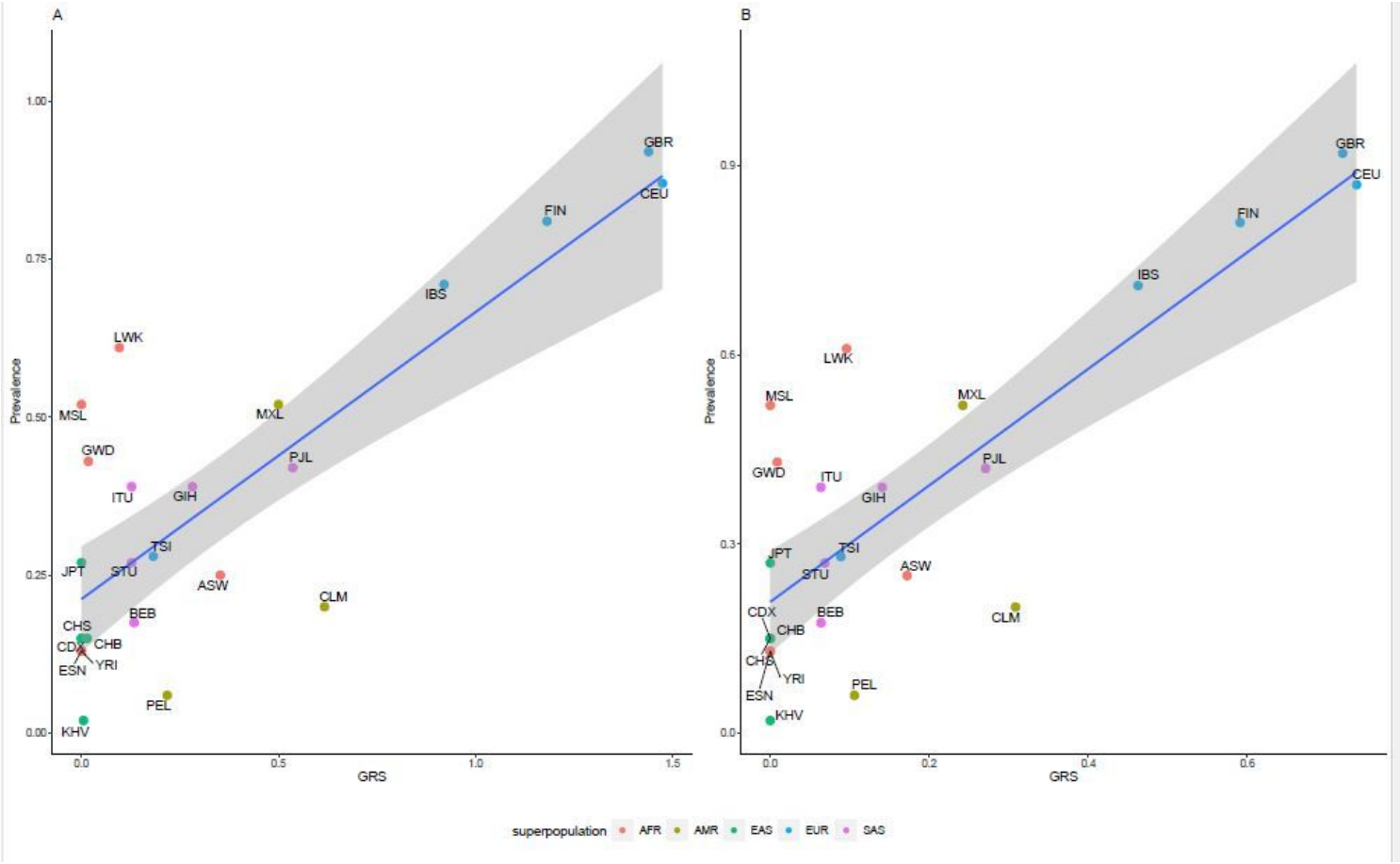

Figure 1

The correlation between lactase persistence and psGRS. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model (r2 = 0.65; p-value: 1.84 x 10-06). B) After maximization (r2 = 0.67, p-value: 9.13 x 10-07).

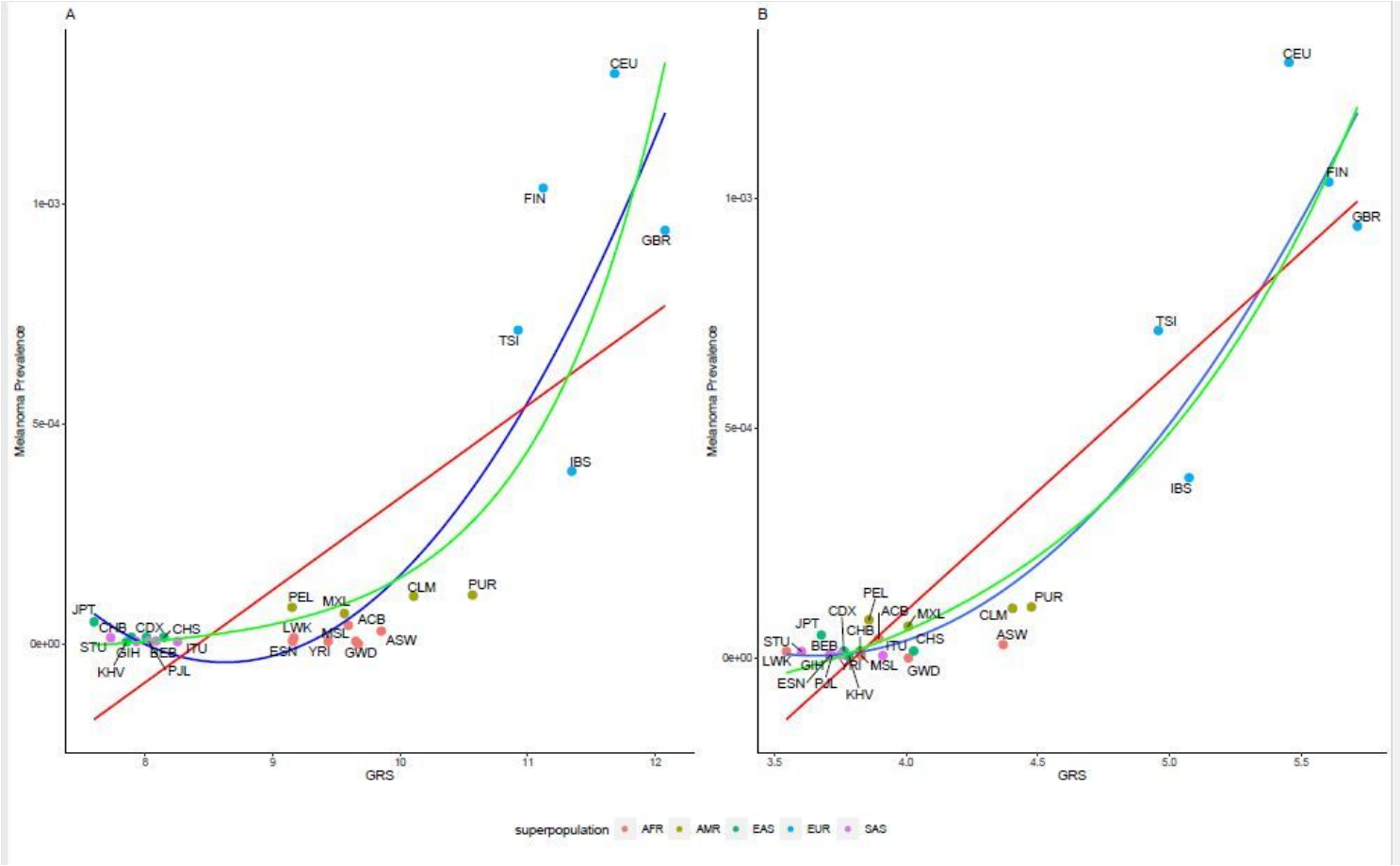

Figure 2

The correlation between melanoma and psGRS with regression lines for linear (red), polynomial (blue) and exponential (green) relationships. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model with three regressions: polynomial (r^2 = 0.78, p-value: 2.19x10-07), exponential (r2 = 0.66, p-value 2.7x10-06) and linear (r^2 = 0.59; p-value: 1.71 x 10-05). B) After maximization: linear regression (r^2 = 0.88, p-value: 2.81x10-11), polynomial (r2= 0.94, p-value: 7.36x10-13) and exponential (r2 = 0.77, p-value: 3.39x10-08).

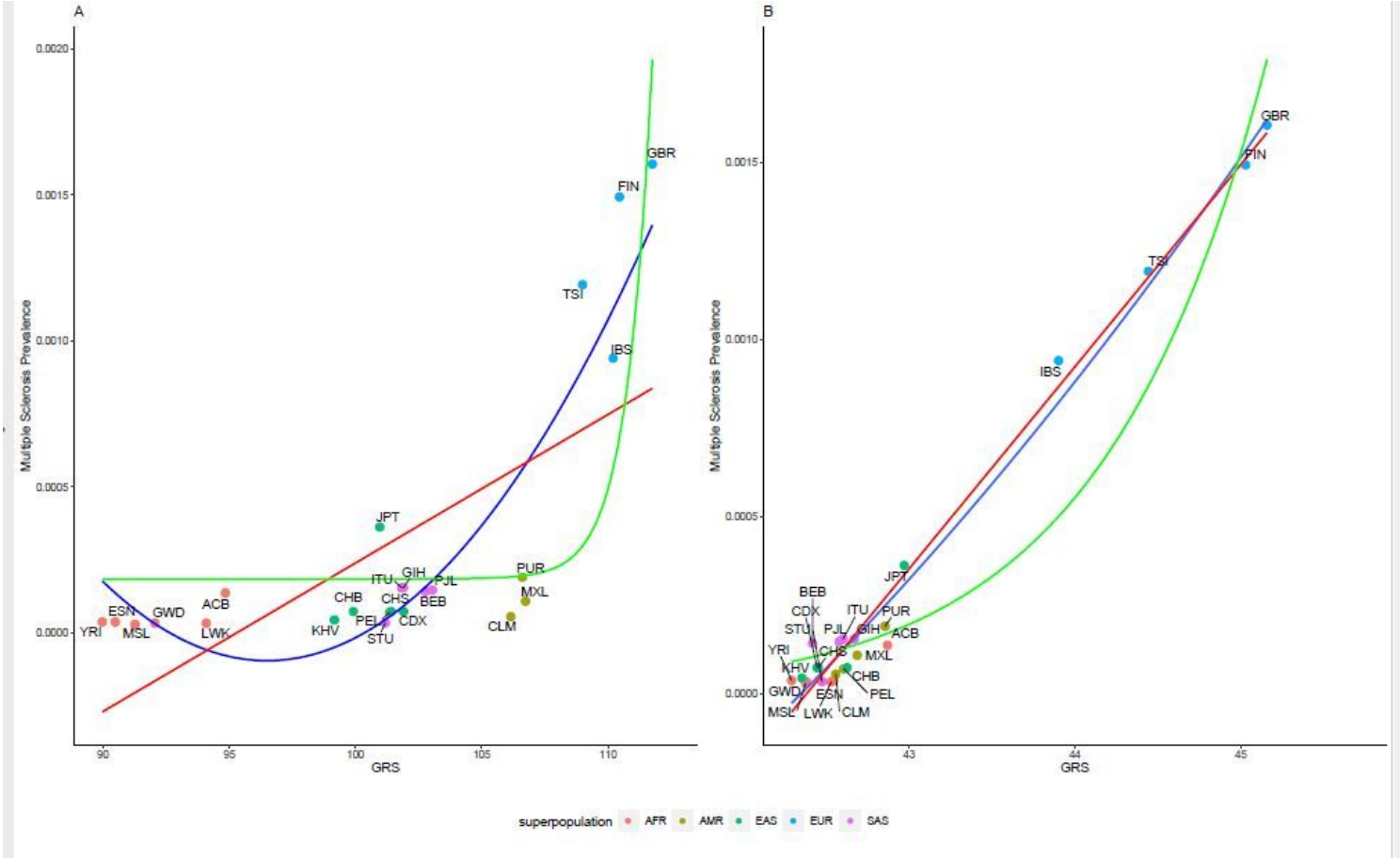

Figure 3

The correlation between multiple sclerosis and psGRS with linear (red), polynomial (blue) and exponential (green) regressions lines. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model with three regression types: polynomial (r^2 = 0.80, p-value: 3.94 x 10-08), linear, (r^2 = 0.47, p-value: 2.12 x 10-04) and exponential (r^2 = 0.64, p-value: 2.59 x 10-06). B) Linear regression after maximization (r^2 = 0.98, p-value: 9.9 x 10-11).

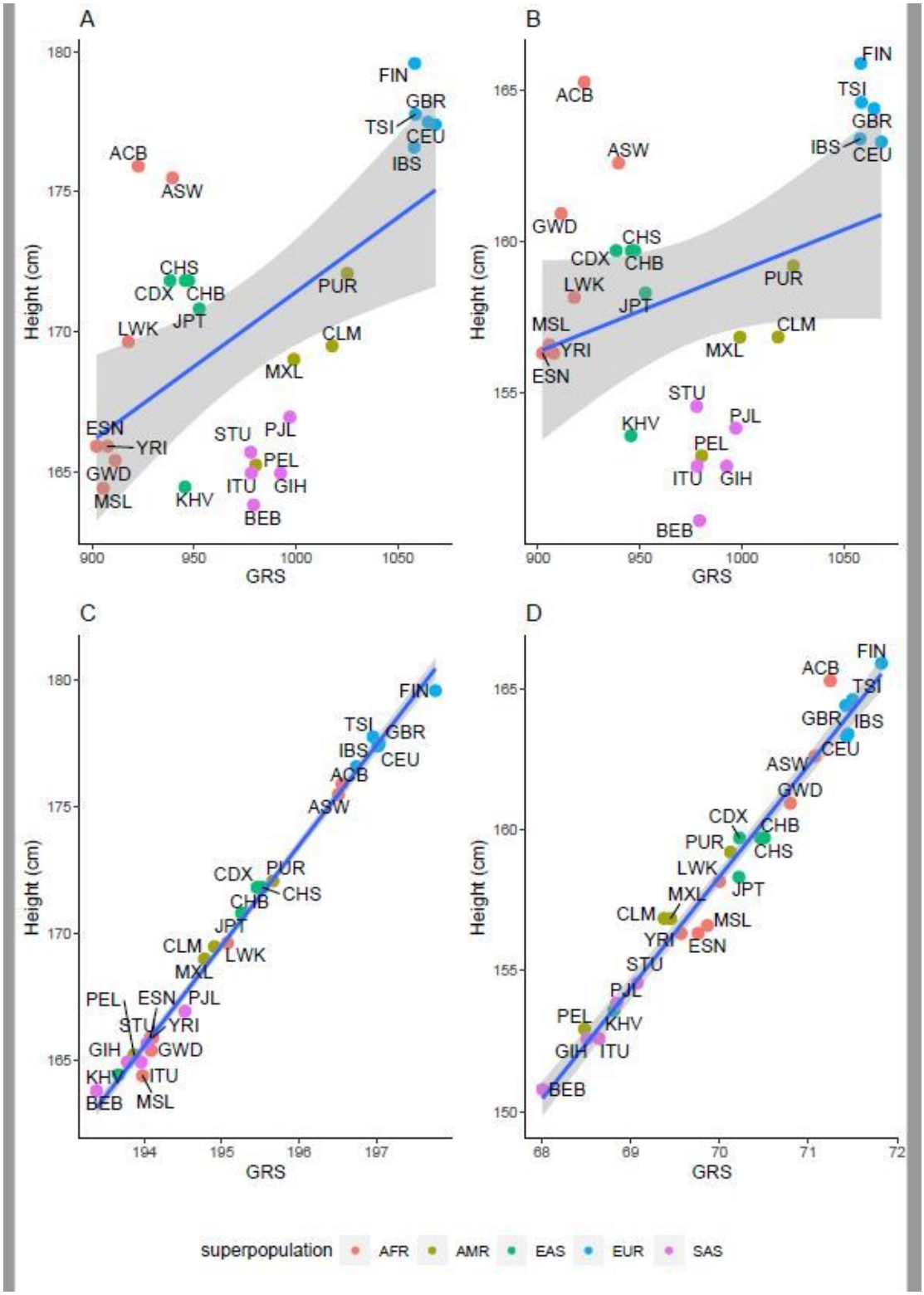

Figure 4

The correlation between height (cm) and psGRS. The data points are colored according to the super populations: AFR (orange), AMR (olive), EAS (green), EUR (blue) and SAS (purple). A) Full model male linear regression (r^2 = 0.32, P-value: 2.55 x 10-03). B) Full model female linear regression (r^2 = 0.11, p-value: 0.0992). C) Male linear regression after maximization (r^2 = 0.99, P-value: <2x10-16). D) Female linear regression after maximization (r2.98, P-value: <2x10-16).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupFig1.pdf
- Supfig2.pdf
- Supfig3.pdf
- Supfig4.pdf
- Supfig5.pdf
- SupplimentaryTables.xlsx
- suppfigs610.pdf