Predicting the direction of phenotypic difference

David Gokhman[1,*,†], Keith D. Harris[2,*], Shai Carmi[3], and Gili Greenbaum[2,†]

[1] *Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot 76100, Israel*
[2] *Department of Ecology, Evolution and Behavior, The Hebrew University of Jerusalem, Jerusalem 91904, Israel*
[3] *Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem 9112102, Israel*
[*] *Equal contribution*
[†] *Corresponding author*

Abstract

Predicting precise phenotypes from genomic data is a key goal in genetics, but it is often hampered by incomplete genotype-to-phenotype data. Here, we describe a more attainable approach than quantitative predictions, aimed at qualitatively predicting phenotypic differences, i.e., which individual has the higher phenotypic value. This approach could be useful in a wide variety of scenarios, e.g., estimating if an individual has an increased disease risk, or if genetically modifying a crop would increase yield. To investigate whether limited genotype-to-phenotype information can still be used to predict which individual has the higher phenotypic value, we developed an estimator of the ratio between known and unknown effects on the phenotype. We formalize a model to delineate the scenarios in which accurate predictions can be achieved and evaluate performance in real-world data from tens of thousands of individuals from either the same family, same population, different populations, or separate species. We find that even in phenotypes where only a small fraction of the genetic effects are known, our estimator can allow for the identification of the individual with the higher phenotypic value, often with over 90% accuracy. We also find that our approach circumvents some of the limitations in transferring association data across populations. Overall, our study introduces an approach for accurately predicting a key feature of phenotypes—their direction—and suggests that more phenotypic information can be extracted from genomes than previously appreciated.

Introduction

A key goal in genetics is to predict phenotypes from genomic data. Such predictions are pivotal for assessing disease risk (1; 2), understanding the genetics underlying adaptation (3; 4; 1), optimizing genetic engineering outcomes (5), reconstructing the traits of extinct species (6), and more. However, our current ability to predict phenotypic values from genetic information, for example by using polygenic scores (PGSs),

is restricted by several confounders. These confounders include environmental effects, the high polygenicity of many phenotypes, the limited ability to identify causal non-coding variants and quantify their effects, and the lack of power to detect small-effect loci (1; 2). Consequently, it is usually impossible to obtain the full information on genotype-environment-phenotype interactions required for precise phenotypic predictions.

Given the limitations associated with predicting precise phenotypes, we suggest here a more attainable objective: predicting only the direction of phenotypic difference. Namely, rather than striving to predict the precise phenotypic value encoded by a particular genome, we aim to predict whether this genome encodes for a larger or smaller phenotypic value relative to another genome. To illustrate, consider a scenario where one is interested in determining the probability that an offspring will be taller than their 170cm tall parent. Considering that a PGS predicts the offspring will be 180cm tall, what is the probability that the offspring will indeed be taller than their parent? We previously implemented a simplified version of this approach to reconstruct the anatomy of the Denisovan using gene regulatory data, validated it on Neanderthals and chimpanzees, and found that it reaches over 85% accuracy in predicting the direction of phenotypic differences (6).

Undoubtedly, predicting a precise phenotypic value is more informative than predicting only the direction of phenotypic difference. However, in studies where the precise phenotypic value cannot be accurately inferred (which is often the case), important insights could still be gained by inferring the phenotypic direction instead. Most importantly, the phenotypic direction is often the crux of phenotypic comparisons, for example, when estimating how likely it is that (i) an individual has an increased disease risk (2), (ii) a genetically modified crop will have increased yield (7), (iii) an offspring will be higher or lower in a certain trait compared to their parents (e.g., in preimplantation genetic diagnoses; (8; 9)), and (iv) the phenotypes of an extinct species differ from those of an extant species.

Here, we explored the feasibility of using currently available genotype-to-phenotype information to predict which individual has the higher phenotypic value. We explored this by deriving predictions from loci whose contribution to the phenotype is known, and then comparing their total effect to the range of the potential effects of unknown genetic and non-genetic contributors. We studied this ratio of known-to-unknown effects through two independent branches of investigation: (i) formalizing a model to delineate the scenarios in which accurate predictions can be achieved, and (ii) evaluating performance in real-world empirical data of individuals from either the same family, same population, different populations, or separate species. Our findings underscore the known-to-unknown ratio as a high-fidelity estimator of prediction accuracy. This allowed us to identify cases where we can reliably discern the individual with the higher phenotypic value. Importantly, this is possible even in cases

where only a small fraction of the genetic effects on the phenotype are known. Overall, our study introduces the known-to-unknown ratio as a valuable tool for predicting complex traits across diverse contexts and suggests that more phenotypic information can be reliably extracted from a genome than previously appreciated.

Results

Approach

We investigate what genomic information is needed to predict the direction of phenotypic difference between two individuals, and the conditions under which this prediction is accurate. We compare two individuals, one for which the phenotype is known (hereafter, *phenotyped* individual) and one for which the phenotype is predicted (hereafter, *unphenotyped* individual).

A phenotype is affected by loci whose contribution to the phenotype was measured (hereafter, *known effects*), as well as by loci or non-genetic factors whose contribution is unknown (hereafter, *unknown effects*, Fig. 1a). We make a prediction on the direction of phenotypic difference by summing up the contribution of the known effects and determining whether the unphenotyped individual has a larger or a smaller sum. We ignore loci where the two compared individuals have the same genotype, because only divergent loci could contribute to the phenotypic difference (Fig. 1b,c). This procedure is equivalent to computing the difference between the PGSs of the two genomes, and using the sign of this difference to predict the direction of the phenotypic difference (9). Finally, we investigate the conditions affecting the probability that a prediction based only on the known effects matches the true direction of phenotypic difference between two individuals (hereafter, *prediction accuracy* or P).

Modeling the conditions needed to predict the phenotypic direction

We can model this approach as a random walk. Each step is an effect on the phenotype in one or the other direction. We define the effect size of each locus as the difference between the contributions of the two genotypes to the phenotype. For example, if the phenotyped individual has genotype A, which increases height by 3mm (relative to a reference), and the unphenotyped individual has genotype B, which decreases height by 1mm, then we consider the effect size of that locus to be +4mm (Fig. 1a). The effect size of loci with the same genotype in the two individuals is 0, and is therefore ignored throughout this work. Our model makes the simplifying assumptions of additivity and no epistasis (10) (in the empirical section, where we test our approach,

simplifying assumptions in the model are evaluated). The sum of the known effects is the vertical displacement of the known effects (Fig. 1b). The direction of the sum of known effects (i.e., whether the displacement is above or below the x-axis in Figure 1b) is our prediction of the direction of phenotypic difference (Fig. 1c). If the sum (blue dot in Figure 1d) is above 0, then our prediction is that the unphenotyped individual has a larger phenotype than the phenotyped individual. If the remaining steps of the random walk (i.e., those of the unknown effects) do not take us below 0, then the final displacement (i.e., true phenotype, black dots in Fig. 1d) is still

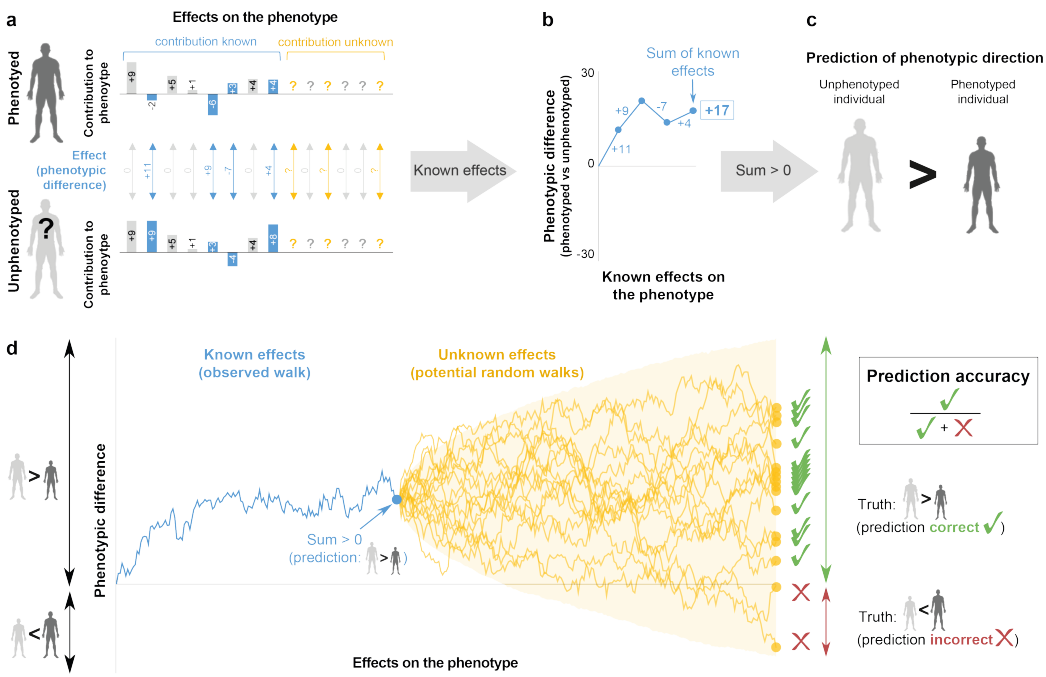

Figure 1: Schematic of the approach to predict the direction of phenotypic difference. (a) We start with a phenotyped individual and an unphenotyped individual. We consider the known and unknown effects contributing to the phenotype of interest. Known genetic effects on the phenotypic difference are in blue (measured in units of the phenotype), unknown genetic and non-genetic effects are in yellow. Cases where the contribution is identical between the two individuals (and therefore do not affect the phenotypic difference) are in gray. (b) Only the known divergent effects are used to predict the phenotypic difference between the individuals. The sum of the known effects can be thought of as the final position of a random walk with step sizes and directions corresponding to the effect sizes. (c) The direction of the total sum of the known effects is used to make a prediction of the direction of phenotypic difference between the phenotyped and unphenotyped individuals. If the the sum of the known effects between the individuals is positive, we predict that the phenotypic value of the unphenotyped individual is larger than the phenotyped individual (and the opposite prediction if the sum is negative). (d) Modeling prediction accuracy using random walks. The curves represent random walks where each step is an effect size. The blue curve shows the known effects of a specific random walk, and the sign (positive or negative) of the blue point at the end of the walk is the predicted direction of phenotypic difference. The yellow curves show potential random walks of the unknown effects (genetic and environmental). The yellow area behind these curves shows the 95% confidence interval of the random walks. In this example, effect sizes were drawn from a standard normal distribution. For a correct prediction of the direction of the phenotypic difference, the sum of the known effects (blue point) and the true phenotypic difference (black dot) need to be on the same side of the x-axis (both below or both above).

above 0, and thus our prediction is correct. Otherwise, if the remaining steps push the displacement below 0, our prediction based on the known effects is incorrect. The larger the vertical displacement of the sum of known effects is, the less likely it is for the final displacement to end on the opposite side of the x-axis, especially if there are many uncorrelated unknown effects, which would tend to cancel each other out.

We start by exploring the factors affecting prediction accuracy and the conditions required for high-accuracy predictions. Various factors have the potential to affect prediction accuracy: the total number of loci affecting a phenotype, the number of known effects, the characteristics of the unknown effects, and more. For example, it is likely that the larger the fraction of known effects is, the more accurate predictions based on these loci would be. However, our random walk perspective suggests that all of these factors amount to only two aspects of the walk that ultimately determine prediction accuracy. The first aspect is the vertical displacement of the sum of the known effects (blue dot in Fig. 1d). Namely, the further above or below 0 we "traveled", the less likely it is that the unknown effects would push the final position to the other side of the x-axis. The second aspect is the variation of the overall potential sums of the unknown effects (i.e., the variation in the displacements generated by the random walk of the unknown effects, yellow region in Fig. 1d). The smaller this variation is, the less likely the unknown effects are to push the final position of the walk to the other side. We propose here that prediction accuracy can in fact be characterized by the ratio between these two aspects. Namely, if the known effects position us on a certain side of the x-axis, and only 10% of the potential walks of the unknown effects end up pushing us to the other side of the x-axis, then the prediction based on the known effects is accurate in 90% of the cases (Fig. 1d). We refer to the ratio between the two aspects in terms of the sum of the known effects, Δ, and the standard deviation of the unknown effects, σ, as the *known-to-unknown ratio*, and denote it as κ:

$$\kappa = \frac{|\Delta|}{|\Delta| + \sigma}. \tag{1}$$

To understand the extent to which the known-to-unknown ratio (κ) reflects prediction accuracy, we mathematically derived a theoretical expectation by formulating prediction accuracy in terms of κ under idealized conditions (Eq. 4 in *Methods*). We have done so using two different approaches – one focuses on the viewpoint of random walks, and the other on the statistical components of the phenotypic variation; the second approach enables the incorporation of additional types of noise, such as shared genetic and environmental components in siblings, and errors in the measurement of effect sizes (see *Methods*). We then use simulations to investigate the robustness of this formulation and explore how different factors affect the distribution of κ values. We simulate pairs of individuals with random known and unknown effects; for each pair, we randomly assign genotypes and arbitrarily treat one individual as phenotyped

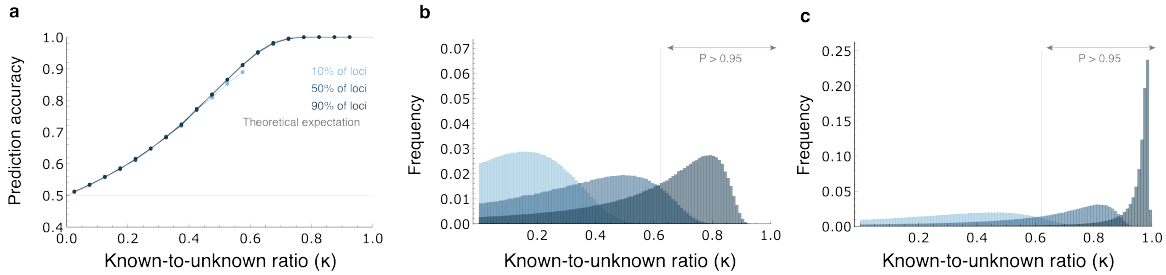

Figure 2: Evaluating prediction accuracy using the known-to-unknown ratio (κ). (a) Simulated prediction accuracies for various κ values, for different proportions of the known vs. unknown effects (10%, 50%, and 90% of effects known). In gray is the theoretical expectation from Eq. 4. (b) The distribution of κ values for the case where the known effects are randomly sampled from a normal distribution. The vertical lines denote the κ values required for prediction accuracy of $P > 0.95$ ($\kappa = 0.62$) (c). The distribution of κ values for the case where the known effects are those with the largest effect sizes. The vertical lines denote the κ values required for prediction accuracy of $P > 0.95$. In all panels, 10,000 effect sizes were drawn from a standard normal distribution to represent the effects on the phenotype, and one million simulation repeats were used for each set of parameters modeled.

and the other as unphenotyped (see *Methods*). Based on these simulated effects, we compute κ for each pair of individuals and determine whether the prediction is correct. Finally, we compute prediction accuracy by evaluating for each bin of κ values the proportion of correct predictions. We conduct these comparisons for different ratios of known to unknown effects, as well as for different effect-size distributions.

We find a strong agreement between the theoretical expectation and the simulated results, across the ratios of known to unknown effects (Fig. 2a), as well as across different effect-size distributions (Fig. S1a–b). This indicates that κ characterizes prediction accuracy well, with phenotypic predictions on pairs with higher κ values showing higher prediction accuracy. For example, for pairs of individuals with $\kappa > 0.62$, prediction accuracy was $P > 0.95$. We find that high values of κ are more common when the portion of known effects is larger (Fig. 2b) but that the underlying effect-size distribution does not substantially affect the κ distributions generated (Fig. S1c–d).

In our depiction of the approach, we implicitly assumed that there is no bias in choosing which effects are known and which effects are unknown. However, many detection methods (e.g., quantitative trait loci mapping or GWAS) have an ascertainment bias, where loci with larger effects are more readily detectable (2). We therefore analyzed scenarios with an ascertainment bias where the known effects are those with the largest effect sizes. As before, we find that κ is a precise descriptor of prediction accuracy (Fig. S2). However, κ values tend to be much higher than in the unbiased scenario (Fig. 2c). Therefore, if the known effects tend to be the largest effects, prediction accuracy tends to be very high. For example, with 50% of effects known in the unbiased scenario, only 10% of the simulated pairs of individuals had prediction accuracy > 0.95 ($\kappa > 0.62$), while in the scenario where the largest effects

are known, this increases from 10% to 65% (Fig. 2b–c, intermediate blue).

In sum, we find that the known-to-unknown ratio (κ) accurately captures the factors that affect the probability of predicting which individual has the higher phenotypic value. The κ estimator could thus be used to (i) evaluate prediction accuracy, and (ii) identify individuals for which high-accuracy predictions could be made, even when genotype-to-phenotype data is limited.

Identifying which individual has the higher phenotypic value in real-world data

The applicability of our approach to real-world data likely varies between phenotypes, conditions, and levels of divergence between the compared genomes. To investigate the relationship between κ and prediction accuracy in empirical data, we analyzed comparisons of increasingly divergent pairs of individuals. We analyzed within-family, within-population, and between-population comparisons of pairs of individuals from the UK Biobank (11). For each pair, we investigated six phenotypes: height, body mass index (BMI), metabolic rate, blood pressure, hip circumference, and bone density. For each phenotype, we selected the loci that significantly contribute to the phenotype, based on a GWAS that excluded the individuals we tested. The effect sizes generated in this GWAS were then used to ascertain Δ by computing the difference between the PGSs of the two individuals. The prediction for each pair was taken to be the sign of Δ, i.e., the individual with the higher PGS was predicted to have the higher phenotypic value. In each comparison, we also computed κ. For the within-family comparisons, we examined all 10,597 pairs of same-sex siblings in the dataset. For within-population comparisons, we randomly sampled 20,000 individuals (10,000 females and 10,000 males) who self-identified as White British with Northwestern European genetic ancestry (hereafter, labeled as European) and examined all pairwise same-sex comparisons among them. For between-population comparisons, we analyzed two genetic clusters in the UK Biobank that are distant from the European group (defined in Fig. S7), one with 1,794 genetically similar individuals, the majority of whom self-identified with ethnicity associated with the geographic region of East Asia (hereafter, labeled as East Asian), and one with 3,091 genetically similar individuals, the majority of whom self-identified with ethnicity associated with the geographic region of Africa (hereafter, labeled as African). For each of these two clusters, and for each sex, we sampled the same number of individuals from the European group; we then analyzed all same-sex comparisons between each non-European group and the equal-sized European group. We also analyzed all the same-sex pairwise comparisons between the two non-European groups. This setup allowed us to examine the ability to predict the direction of phenotypic difference in real-world data across

several phenotypes and varying levels of divergence.

We observed a remarkable agreement between the empirical results and the theoretical expectation (Eq. 4). Across the six phenotypes, higher κ values reflected higher prediction accuracy, with a relationship that tightly followed the theoretical expectation (with the exception of blood pressure at high κ values, see *Discussion*). Importantly, this is maintained across different levels of divergence between individuals, suggesting that κ captures the key aspects determining the ability to predict phenotypes (Fig. 3a–c). Empirical estimation of κ requires assessment of the heritability of the phenotypic difference in pairwise comparisons. We extracted these estimates from our analyses (Fig. S6, see *Methods*), but using external population-level estimations of heritability resulted in almost identical trends (Table S1), with the agreement between the theoretical expectation and the empirical results remaining strong (Fig. S6).

Our approach also allowed us to estimate the proportion of individuals for whom high-accuracy predictions can be achieved. For example, for 5% of pairs from the European group, κ values for bone mineral density are ≥ 0.4, and we can therefore predict which individual has higher bone mineral density with 75% accuracy (i.e., threefold more likely to predict correctly than incorrectly; Figs. 3e and S4a). For height, where a larger fraction of loci affecting the trait is known, the same prediction accuracy can be achieved for 25% of pairs. Notably, despite environmental effects and incomplete genotype-phenotype information, for 3% of pairs, we can predict with 90% certainty which individual is taller (Fig. S4a). Importantly, the percentage of pairs for which high-accuracy predictions can be attained increases with increasing genetic distance (κ distributions are shifted to the right with higher divergence between pairs in Fig. 3e–g). For example, in 3% of sibling pairs, we can predict which sibling is taller with 85% certainty; between unrelated individuals from the European group, this increases to 8% of pairs, and in pairs where one individual is from the European group and the other is from the East Asian group, this number increases to 36% (Fig. S4a–c). This is likely the result of the larger number of divergent sites between more divergent pairs, as well as the larger phenotypic difference.

A major concern in GWAS is its limited transferability across populations. PGSs computed using data from one population often perform substantially worse when applied to other populations (12). To test whether this phenomenon affects our approach, we evaluated the relationship between κ and prediction accuracy using GWAS conducted on individuals with European ancestry, but predicting phenotypes between pairs of individuals with East Asian or African Ancestry (populations defined in (13)). As expected, we observed lower κ values for these comparisons relative to the κ distribution without transfer (Fig. 3h). This is likely owing to the smaller fraction of the phenotypic variance being explained by the SNPs identified in the European-ancestry GWAS (12; 13). Nevertheless, here too, we observed a good agreement with

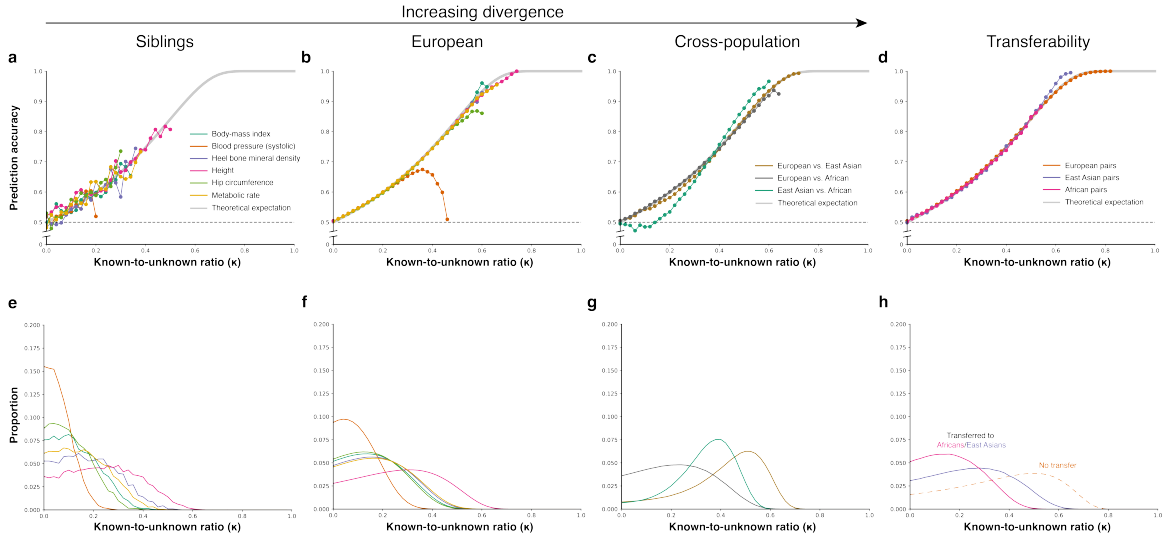

Figure 3: Predictions of the direction of phenotypic difference in humans. (a)–(d) The relationship between the known-to-unknown ratio (κ) and the proportion of correct predictions in different phenotypes and in comparisons with varying levels of divergence. The theoretical expectation (Eq. 4) is shown in gray). (a) Pairwise comparisons of siblings from the UK Biobank for six phenotypes. (b) Pairwise comparisons of individuals from the European group (self-identified White British with Northwestern European genetic ancestry) from the UK Biobank for the same six phenotypes. (c) Pairwise height comparisons of individuals with different ancestries in the UK Biobank. (d) Pairwise height comparisons of individuals from the same population (either European, East-Asian or African, as defined in Fig. S7), using GWAS generated from a European-ancestry group in Yengo *et al.* (14). (e)–(h) The distribution of κ values for all pairwise comparisons. Each panel corresponds to the panel above it. (g) See Methods for a discussion of differences between the curves.

the theoretical expectation for the relationship between κ and prediction accuracy (Fig. 3d). In fact, the fit between the empirical and theoretical predictions is as good as in the previous sections. Thus, while fewer usable SNPs and increased noise in effect size estimation lead to fewer pairs with high-accuracy predictions, the ability to robustly estimate prediction accuracy is maintained.

In summary, we found that: (i) given a pair of individuals, we are able to accurately evaluate our chances of correctly predicting which individual has the higher phenotypic value, and (ii) even for phenotypes with limited genotype-to-phenotype data, some pairs have sufficiently high known-to-unknown ratios (κ) to enable the identification of the individual with the higher phenotypic value. Two important implications of these findings are that we can (i) select the subset of pairs of individuals for which we can make high-confidence predictions, or (ii) given a pair of individuals, select the subset of phenotypes for which we can make high-confidence predictions.

Impact of directional selection on predictions between populations and species

In the model above, we have not addressed the role of selection. Directional selection most likely has little effect on the within-population UK Biobank comparisons, but may play a more central role when more divergent genomes are compared. In this section, we extend our model to include directional selection and examine predictions in divergent populations and species (see *Discussion* for the potential effects of negative and stabilizing selection).

Until now, our model assumed that the effects have an equal probability of increasing or decreasing the phenotypic difference. Under directional selection, the phenotype of a lineage is typically pushed towards a new optimal value. The directions of effects of that lineage relative to the ancestral lineage are more likely to be in the direction of this optimum. Thus, to model the case that directional selection shaped the divergence between the two compared genomes, we introduce correlated effects into our model. We consider the case where selection is stronger for larger effect sizes (15). In other words, each effect has a higher probability of being aligned with the direction of selection than with the opposite direction, and this probability increases with the size of the effect and the strength of selection. To model this, we introduce into the random walk a bias that favors one direction over the other and is stronger with larger effects.

In the neutral random walk, the direction of the sum of known effects tends to match the true direction of the phenotypic difference (Fig. 2a); in the biased random walk, this trend is stronger (Fig. 4a–b). This is because the sum of known effects and the sum of unknown effects are pushed by selection in the same direction throughout the walk (Fig. 4a). The bias we introduce violates the assumption of the model above, and therefore, we do not expect the relation between κ and prediction accuracy to be the same as the theoretical expectation in Eq. 4). Indeed, we find an improvement in the prediction accuracy relative to the neutral expectation, i.e., prediction accuracy is higher for any given value of κ (Fig. 4b). This improvement in prediction accuracy increases with stronger directional selection. Not only does the change in the relationship of κ and prediction accuracy improve prediction accuracy, but the proportion of pairs of individuals with high κ values also increases with stronger selection (Fig. S3). As a result, with directional selection, high-accuracy predictions can be achieved more often and with fewer known effects.

These results suggest that more divergent lineages, where directional selection might have played a more central role, would tend to show higher prediction accuracy. To investigate this, we explored genotype-to-phenotype datasets of more divergent lineages. We tested three quantitative trait loci (QTL) mapping datasets investigating stickleback (16), daisy (17), and mouse (18) phenotypes. The stickleback dataset

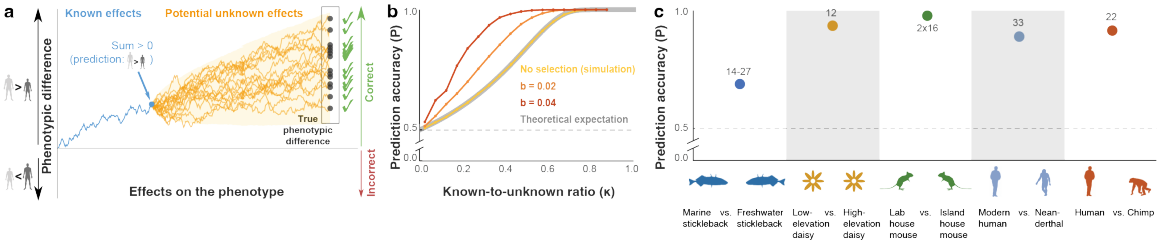

Figure 4: The effect of directional selection on predicting the direction of phenotypic difference. (a) Prediction accuracy under directional selection, modeled as a biased random walk. The random walks in this schematic are biased toward the positive direction, with larger effects having a stronger bias. The shaded yellow area shows the 95% confidence interval for the unknown effects, as in Fig. 1d. Biased random walks increase prediction accuracy. (b) Prediction accuracy for different κ values and different levels of bias (b), with 50% randomly selected known effects out of 10,000 overall. (c) Prediction accuracy based on known effects and the true direction. For each pair of populations or species, one group is considered phenotyped while the other is considered unphenotyped (although phenotypes are known for both groups). Each point represents the proportion of matches out of the number of phenotypes in the dataset, which is noted above each point. For sticklebacks, between 14 and 27 phenotypic predictions were made for four different freshwater populations. For mice, predictions were made for two phenotypes in 16 developmental stages.

included four freshwater populations that diverged from a common marine ancestor less than 12,000 years ago (16). We analyzed the 27 morphological phenotypes in the dataset, with 1–2 QTLs reported per phenotype, and found that, even with as few known loci as this, the proportion of matches was 63%-75% (depending on the pair of populations compared). In other words, using 1–2 loci, we were about 2–3 times more likely to predict the phenotypic direction correctly than incorrectly (Fig. 4c).

In the daisy dataset, we analyzed 1–5 QTLs for 12 phenotypes that differ between two species of daisy (17). We found a prediction accuracy of 92%, with 11 out of 12 phenotypes predicted correctly based on these known effects (Fig. 4c). The mouse dataset included growth rate and weight divergence of Gough Island vs. wild-type mice over 16 developmental stages (18), with 8–11 QTL per phenotype. The proportion of matches in this case was 100% (Fig. 4c). Interestingly, this perfect prediction accuracy is achieved despite the fact that in some developmental stages, the joint effect of all known effects explains as little as 6% of the variance in weight and 3% of the variance in growth rate. In addition, in all three datasets, the single largest-effect locus was sufficient to predict the direction of phenotypic difference with high accuracy (63%-75% for sticklebacks, 92% for daisies, and 75% for mice).

We also explored our previous study on more divergent lineages, where the phenotypic direction was predicted between Neanderthals and modern humans, and between chimpanzees and humans. These predictions were based on DNA methylation data and were made only for phenotypes where all known effects pointed in the same direction of phenotypic change (consequently, effect sizes were not required to generate predictions). Therefore, this study's setup filtered for predictions with higher κ

values. Prediction accuracy for the 33 Neanderthal phenotypes and 22 chimpanzee phenotypes was 88% and 91%, respectively (6). Interestingly, we observed similar patterns in our more recent study comparing human and chimpanzee gene expression in human-chimpanzee hybrid cells (19). In this study, the most differentially expressed genes were $\tilde{\text{f}}$our times more likely to predict the direction of phenotypic difference correctly than incorrectly (i.e., 81%) (19).

Overall, these datasets represent a diverse range of phenotypes, species, divergence times, and genotype-to-phenotype association methods. While we most often do not know the exact nature of the selection processes that shaped the genetics of organisms, our results suggest that when comparing divergent genomes, we can achieve relatively accurate prediction of the direction of the phenotypic difference with few (most likely large-effect) loci.

Discussion

Here, we investigated an alternative approach to inferring phenotypes from genetic data—instead of trying to predict a phenotype quantitatively, we explored a more modest approach, whereby only the direction of phenotypic difference is predicted. Our goal was to develop a model for the probability of accurately predicting the phenotypic difference under various conditions and to test it on empirical data. We found that prediction accuracy is affected by two main factors: the sum of known effects, and the variation in the potential sums of the unknown effects. We formulated the ratio between these factors as κ, which we found to be a high-fidelity estimator of prediction accuracy. The κ statistic allows us to identify pairs of individuals where the direction of phenotypic difference could be confidently predicted. This estimator is not affected by ascertainment bias, the level of divergence between individuals, or transferability problems with the data. Thus, even when there is limited information on the interaction between genotypes, environments, and phenotypes, one can still identify the pairs for whom accurate predictions can be made. Overall, we found that such pairs are more common when (i) more information is known about the genetic basis of the phenotypic variation, (ii) the phenotype was more strongly affected by positive selection, (iii) there is a stronger ascertainment bias in identifying large-effect loci.

Several factors have not been incorporated into our model. (i) Our model assumes additivity of effect sizes and does not incorporate epistasis or pleiotropy. Although previous studies have shown that variation in many complex traits is mostly additive (20; 21; 22), the assumption of additivity may not hold for some phenotypes (23; 24). (ii) In cases where loci affecting the phenotype are highly pleiotropic, the directions of effects may be driven by selection acting on other phenotypes. When this causes

the directions of effects to behave as Bernoulli random variables, our basic model would still be applicable. When this causes the effects to deviate from Bernoulli random variables and to be biased towards one direction, our selection model would likely be applicable, though without the correlation between effect size and strength of selection. (iii) Finally, we model environmental effects as part of the unknown effects, i.e., reflecting the same dynamics. However, in phenotypes that evolve under stabilizing selection in the face of shifting environments, genetic and environmental effects can have opposing trends (25). Thus, more complex modeling incorporating such scenarios, as well as epistatic and pleiotropic effects, would help elucidate their impact on prediction accuracy. Despite these limitations, testing our approach on real data, where no assumptions regarding these factors are made, suggests that our current model captures the main factors affecting predictions.

To model selection, we used an approach where loci are affected by selection in proportion to their effect sizes. While this is the general case, selection often follows more complex dynamics (3). For example, Hayward and Sella (26) investigated temporal evolutionary dynamics of a rapid adaptation phase followed by a prolonged stabilizing selection phase. This study showed that in the long term, phenotypic variation is dominated primarily by small and moderate effect sizes, but that the few remaining fixed large effects are almost always aligned in direction with the overall phenotypic difference. In other words, the larger the effect size of a locus that separates the two groups, the more likely it is to reflect the overall phenotypic difference between them (26). Interestingly, in the scenarios studied by Hayward and Sella, the few divergent large-effect loci are not particularly informative for predicting the phenotype quantitatively (because they tend to explain a small percentage of the phenotypic variation), but they are particularly informative for predicting its direction (because they are almost always aligned with the direction of phenotypic difference). This could further explain the high prediction accuracy reached in our between-species comparisons, where the few known large-effect loci explain a small percentage of the overall phenotypic difference, but are very predictive of the direction of phenotypic difference.

Other types of selection could also affect predictions. For example, negative or stabilizing selection is expected to reduce the number of divergent loci between two individuals, thus decreasing both the known and unknown effects. If it disproportionately affects larger-effect loci, it might reduce the relative contribution of the known effects, thus shifting κ values towards lower values, resulting in lower prediction accuracy. Unlike directional selection, this is not expected to affect the relation between κ and prediction accuracy.

One of the most intriguing uses of phenotypic inference is its potential to predict an individual's susceptibility to a particular disease. Since disease risk is not directly quantifiable per individual, we tested instead our ability to identify the in-

dividual with the disease in a pair of individuals where one is healthy and the other is reported to have the disease. Here too, the empirical results mostly align with the theoretical expectation. However, unlike all other analyses, at higher κ values ($\kappa >\sim 0.4$), the empirical results started to deviate from the theoretical expectation (Fig. S5a). We have not been able to pinpoint the underlying driver of this phenomenon. One plausible explanation is that in these comparisons, higher κ values reflect instances where one of the individuals is indeed more likely to develop the disease, but early signs of the disease or family history prompted treatment and thus exclusion from the disease group. Potential support for this can be seen in the context of the blood pressure phenotype. At higher κ values, predictions diverge from the theoretical expectation both in the within-population analysis of blood pressure (Fig. 3a), as well as in the disease analysis of hypertension (Fig. S5a), where for high κ values prediction accuracy approaches 0 and our predictions are not even random, but systematically wrong. This behavior may indicate a negative correlation between high kappa values and the disease, possibly reflecting medication-induced phenotypic changes that specifically affect individuals with a higher likelihood of elevated blood pressure, thereby altering the predictive outcome. Nevertheless, for most cases, where κ values are not extreme, it is possible to generate clinically relevant predictions, especially when the unphenotyped individual has a higher probability of having the disease compared to an individual known to have the disease.

The approach we presented evaluates the extent to which a key feature of a phenotype — its direction — can be predicted from genomic data. Given the currently limited ability to quantitatively predict phenotypes from genotypes (2), our approach opens a window to predicting phenotypes qualitatively. While there is still much to explore with regard to the applicability of this approach to various data, its capability to robustly estimate prediction accuracy and to identify individuals and phenotypes for which accurate predictions can be achieved, suggests that more phenotypic information can be extracted from genomes than previously appreciated.

Methods

Formal model for prediction accuracy

We consider a pair of individuals, one phenotyped and the other unphenotyped, with genomes that diverge in n loci that affect a certain phenotype. We denote the differential effect of these loci as e_i, which is the relative contribution of locus i to the difference between the phenotypes of the two individuals (Fig. 1a). Each effect of a divergent locus either increases the phenotypic difference in the direction of the phenotyped individual or in the direction of the unphenotyped individual; we (arbitrarily) denote this direction as $d_i = 1$ if it increases the phenotypic difference towards the phenotyped individual, and $d_i = -1$ if it increases it towards the unphenotyped individual. The sum of the known effects is, therefore, $\Delta = \sum_{i=1}^{n} d_i e_i$ (Fig. 1b). The sign of Δ is our prediction for the direction of the phenotypic difference (Fig. 1c).

We consider additional m unknown effects on the phenotype, and denote them as random variables $X_1, ..., X_m$. For the most part of this work (but see simulations with selection below), we consider these random variables to be independent random variables that attain one of two values, E_j or, $-E_j$, with equal probability, i.e. $X_j \sim 2E_j\Big(Bernoulli(\frac{1}{2}) - \frac{1}{2}\Big)$. This means that each divergent unknown effect has some contribution to the phenotype, and it can work to either increase the phenotype in the phenotyped or in the unphenotyped individuals. We denote the sum of the unknown effects as $\Omega = \sum_{i=1}^{m} X_j$. Following the definitions in Eq. 1, we denote the variance of Ω as σ^2.

The true phenotypic difference is $D = \Delta + \Omega$, the sum of both known and unknown effects. Our prediction is correct if the signs of Δ and D are the same; otherwise, our prediction is incorrect. We define 'prediction accuracy' P as the probability that the signs of Δ and D match, i.e., $P = Prob\Big(\text{Sign}[D] = \text{Sign}[\Delta]\Big)$.

Mathematical relationship between κ and P

Without loss of generality, let us assume that $\Delta > 0$. Prediction accuracy is therefore the probability that the true phenotypic difference is also positive, $P = Prob\Big(\Delta + \Omega > 0\Big)$. Reformulating this by plugging in Eq. 1 to replace Δ, we have

$$P = Prob\Big(\Omega < \frac{\kappa\sigma}{1-\kappa}\Big) \tag{2}$$

Notably, Ω is a sum of independent random variables that satisfy Lindberg's condition (for large n, the contribution of the variance of each random variable to the variance of Ω in negligible). Therefore, applying the generalized central limit theorem, Ω is approximately normally distributed. Ω has a mean of zero because each of the random variables has a mean of zero. We can now use the CDF of Ω, $F_\Omega(x) = \Phi(\frac{x}{\sigma})$ to explicitly compute prediction accuracy:

$$P = F_\Omega\Big(\frac{\kappa\sigma}{1-\kappa}\Big) = \Phi\Big(\frac{\kappa}{1-\kappa}\Big) \tag{3}$$

The CDF of a standard normal distribution can be formulated explicitly using the error function erf, $\Phi(x) = \frac{1}{2}\Big(1 + erf\big(\frac{x}{\sqrt{2}}\big)\Big)$, and therefore we can explicitly describe the expected relationship between P and κ:

$$P = \frac{1}{2}\left(1 + erf\Big(\frac{\kappa}{\sqrt{2}(1-\kappa)}\Big)\right) \tag{4}$$

Alternative derivation

We can also derive this result using standard notations in statistical genetics. As before, we consider that a phenotype is measured in normalized units, i.e., $y \sim N(0,1)$. The PGS of an individual p is then distributed $p \sim N(0, r^2)$, where r^2 is the proportion of the phenotypic variance explained by the PGS. We denote the combined non-measured genetic factors and non-genetic factors affecting the trait as e, which is also the residual of the regression of the trait on the PGS. We can thus write $y = p + e$. We assume p and e are independent and $e \sim N(0, 1 - r^2)$. Next, we consider two unrelated individuals with computed PGSs p_1 and p_2 such that $p_1 > p_2$, with residuals e_1 and e_2, respectively (we assume that e_1 and e_2 are independent because the individuals are unrelated). Denoting the difference in PGSs as $d = p_1 - p_2$, and using d to predict the direction of phenotypic difference, the prediction accuracy is therefore $P = Prob(y_1 > y_2)$, where y_1 and y_2 are the true phenotypic values of the two individuals. We can reformulate this probability as $P = Prob(e_2 - e_1 < p_1 - p_2)$, and therefore $P = Prob(e_2 - e_1 < d)$. We denote $e' = e_2 - e_1$, and because e_1 and e_2 are each normally distributed with variance $1 - r^2$ and zero mean, we have $e' \sim N\Big(0, 2(1-r^2)\Big)$. We can now observe that:

$$P = Prob(y_1 > y_2) = Prob(\Delta e' < d) = \Phi\Big(\frac{d}{\sqrt{2(1-r^2)}}\Big). \tag{5}$$

Reformulating Eq. 1 with the notation of this section (i.e., $|\Delta| = d$ and $\sigma = \sqrt{2(1-r^2)}$ because $2(1-r^2)$ is the variance of the differences of the unknown effects), we have $\kappa = \frac{d}{d+\sqrt{2(1-r^2)}}$, and therefore

$$\frac{\kappa}{1-\kappa} = \frac{\frac{d}{d+\sqrt{2(1-r^2)}}}{1-\frac{d}{d+\sqrt{2(1-r^2)}}} = \frac{d}{\sqrt{2(1-r^2)}}, \tag{6}$$

showing that equations 3 and 5 are equivalent.

We can use this formulation to also derive the prediction accuracy P for the case when the two compared individuals are siblings. In this case, we need to consider the shared genetic component and the shared environmental component. We therefore model the true phenotypic values of the two compared individuals as $y_1 = p_1 + \overline{g_1} + \overline{e_1}$ and $y_2 = p_2 + \overline{g_2} + \overline{e_2}$, where the PGSs are $p_i \sim N(0, r^2)$ as before, $\overline{g_i}$ is the genetic component that is not modeled in the score and is distributed as $\overline{g_i} \sim N(h^2 - r^2)$ with h^2 being the overall narrow-sense heritability of the phenotype in the population, and $\overline{e_i}$ is the non-genetic component which is distributed $e_i \sim N(0, 1-h^2)$. We then consider the shared genetic and environmental components by defining $\overline{g_i} = g_i + g_s$ and $\overline{e_i} = e_i + e_s$, where g_s and e_s are the shared genetic and environmental components, respectively, and g_i and e_i are unique to each sibling. In siblings, the correlation of the genetic components is 0.5, and therefore $g_i \sim N(0, \frac{h^2-r^2}{2})$. For the environment, we define the shared environmental variance as c^2, and therefore $e_i \sim N(0, 1-h^2-c^2)$ (27). We can now derive the prediction accuracy when comparing the two siblings given their PGS difference between d:

$$P = Prob(y_1 > y_2) = Prob(p_1 + g_1 + g_s + e_1 + e_s > p_2 + g_2 + g_s + e_2 + e_s) = Prob\Big(g_2 - g_1 + e_2 - e_1 < d\Big). \tag{7}$$

Assuming that the g_i's and e_i's are all pairwise independent, the variance of $(g_2 - g_1) + (e_2 - e_1)$ is $(h^2 - r^2) + 2(1-h^2-c^2) = 2 - h^2 - r^2 - 2c^2$. Therefore, the prediction accuracy of sibling comparisons can is given by

$$P = \Phi\Big(\frac{d}{\sqrt{2-h^2-r^2-2c^2}}\Big). \tag{8}$$

Note that, given that $r^2 < h^2$, as the PGS cannot explain more variance than the heritability, we have $2 - h^2 - r^2 - 2c^2 < 2 - r^2 - r^2 - 2c^2 < 2 - r^2$. Thus, $\Phi\Big(\frac{d}{\sqrt{2-h^2-r^2-2c^2}}\Big) > \Phi\Big(\frac{d}{\sqrt{2(1-r^2)}}\Big)$. This is expected, because a given PGS difference is more informative on phenotypic differences between siblings compared to unrelated pairs.

Simulations

To simulate a single pairwise comparison, assuming a pre-specified effect size distribution, we sample n effects $e_1, \ldots, e_n$ defined as 'known' and m effects $E_1, \ldots, E_m$ defined as 'unknown' from the distribution. We consider either symmetrical (around 0) distributions or non-negative distributions. For symmetrical distributions, the effect sizes can be either positive or negative. For non-negative distributions, we simulate the direction of the effects in the pairwise comparison by sampling $n+m$ values of 1's and -1's with equal probability, $d_i \sim 2\big(Bernoulli(\frac{1}{2}) - \frac{1}{2}\big)$; we now redefine the effect sizes such that e_i is replaced by $d_i e_i$ and E_j is replaced by $d_{j+n} E_j$. We then compute the sums $\Delta = \sum_{i=1}^{n} e_i$ and $D = \Delta + \sum_{j=1}^{m} e_j$, as in the formulation above. If $\text{Sign}[D] = \text{Sign}[\Delta]$ the simulation results in a match between the prediction and the true direction, otherwise, it results in a mismatch. By repeating the simulations 10^6 times for a given scenario, we compute prediction accuracy P as the proportion of matches out of all simulation repeats.

We evaluated different portions of known effects out of all effects (10%, 50%, and 90%). Effect size distributions can be shaped by various evolutionary processes, such as mutation, selection, and genetic drift (3; 28); therefore, we simulated effect size distributions of various types (normal, gamma, and Orr's negative exponential model distributions; (29; 3)). We also considered the case where the known effects tend to be the larger effects. To simulate this, we start by sampling $n+m$ effect sizes from the pre-defined ES distribution,

$\tilde{e}_1, \ldots, \tilde{e}_{n+m}$. We then reorder the effect sizes by their size and re-index them such that $\tilde{e}_i \geq \tilde{e}_{i+1}$ for all $i = 1, \ldots, n+m$. We define the known effects to be $e_i = \tilde{e}_i$ for $i = 1, \ldots, n$ and the unknown effects as $E_j = \tilde{e}_{j+n}$ for $j = 1, \ldots, m$. This way, the n known effects are taken to be the larger effects from those drawn. We then simulate the $n+m$ directions d_i and continue with the rest of the simulation as described above.

Modeling and simulating directional selection

To model directional selection, we adjust the random variables representing the unknown effects, $X_1, \ldots, X_m$, to tend to be directed towards the direction in which selection is operating. For simplicity, we arbitrarily take this direction to be the positive direction. For each random variable, we adjust the probability for attaining a positive direction for the effect size according to a parameter denoting the strength of selection, s, and the absolute size of the effect. For a sampled effect size e_i, we do this by defining a probability $p_i = 1 - \frac{1}{2}e^{-s|e_i|}$ for attaining e_i, and $-e_i$ is attained at probability $1 - p_i$. Under this formulation, $X_i \sim 2e_i\Big(Bernoulli(p_i) - \frac{1}{2}\Big)$. Note that s is not a selection coefficient in units of fitness, but is rather a unitless parameter that is proportional to the impact of selection on the alignment of the directions. The motivation for this particular formulation is based on the Ornstein-Uhlenbeck model, which is used to model the evolution of quantitative traits subject to both drift and selection by considering random walks with some pull toward a particular state (30; 31; 32). Under our model, when $s = 0$ or e_i is very small, then $p_i = \frac{1}{2}$, as in the model above, which does not consider selection; as s and e_i increase, p_i approaches 1, meaning that the direction of the effect is likely to be in the positive direction.

The simulation with directional selection with a parameter s is identical to the simulation procedure described above, except that the effect sizes $e_1, \ldots, e_{n+m}$ are simulated differently. Here, for a symmetrical effect size distribution, we draw $n+m$ effect sizes $e_1^*, \ldots, e_{n+m}^*$, and take their absolute values $e_i = |e_i^*|$; for positive effect size distributions we draw the values as described above for the no-selection model. We then proceed with simulating the direction of the effects (for both symmetrical and positive effect size distributions) as $d_i \sim 2\big(Bernoulli(p_i) - \frac{1}{2}\big)$, using the probability p_i defined above. The computation of Δ and D is as described above for positive distributions.

Analysis of pairwise comparisons in humans

Computing κ from empirical data

Computing κ for a given pair of individuals using Eq. 1 requires (i) effect sizes for known loci contributing to the phenotypic difference to compute Δ, and (ii) the variance of unknown effects σ^2. The effect sizes can be ascertained from summary statistics of large genotype-phenotype datasets (see next section), but computing the variance of the unknown effects could be challenging. In order to compute this parameter, we first normalize the effect sizes by standardizing them (i.e., transforming them to z-scores), denoting the new standardized effect sizes as e_i^*, and denoting $\Delta^* = \sum_{i=1}^{n} e_i^*$, where n is the number of known effects that are divergent between the two individuals. We introduce a new parameter, $\overline{h}$, which denotes the overall contribution of known effects to the phenotypic difference between the pairs. $\overline{h}$ is, in many respects, similar to the narrow-sense heritability of the phenotype, but refers only to those loci that are divergent between the two compared individuals; it is, therefore, expected to have similar values to population heritability estimates computed using other means. Next, we assume that the measured phenotypic values have been normalized and transformed to z-scores. Under this normalization, the sum of the known effects in units of the true phenotypic difference is $\Delta = \overline{h}\Delta^*$, and the variance of the true phenotypic difference is 1. The variance of the true phenotypic difference is composed of the sum of the portion of the variance explained by the known effects, $\overline{h}$, and the variance of unknown effects σ^2; therefore $\sigma^2 = 1 - \overline{h}$. Using these standardized units, we can reformulate Eq. 1:

$$\kappa = \frac{\overline{h}\Delta^*}{\overline{h}\Delta^* + \sqrt{1 - \overline{h}}} \tag{9}$$

To use this formulation with empirical data, we require an estimation of $\overline{h}$. Below we explore the option of estimating $\overline{h}$ from the data by considering the fit to the theoretical expected relationship of κ and P (Eq. 4). We also used, as an alternative method for estimating $\overline{h}$, heritability values derived from population-level analysis of the UK Biobank.

Analysis of the UK Biobank

To test our approach on empirical data, we used the UK Biobank (UKB), a large dataset containing almost 500 thousand genotyped individuals with associated phenotype data (11). We generated subsets of comparisons that have different levels of divergence: (i) sibling pairs (within-family), (ii) pairs of individuals with Northwestern European ancestry (within-population), and (iii) pairs of individuals with different ancestry groups, either pairs from the European–East Asian, European–African, or East Asian–African groups (cross-population). Northwestern European ancestry was determined using the UKB Data-Field 22006. Our non-European groups where defined by demarcating clusters of genetically similar individuals that are distant from the European group on the PC1 and PC2 of the UKB PCA results from UKB Data-Field 22009 (Fig. S7). The two clusters were labeled as East Asian and African based on the majority of self-identifications of these groups as reported in UKB Data-Field 21000, associating them with these two geographic regions, and consisted of 1,794 and 3,091 individuals, respectively.

To compute κ values, we first generated GWAS results for a number of continuous traits. We included variants with high-quality imputation scores (imputation INFO scores ≥ 0.8) from the UKB imputed genotype release version 3 (11); this yielded roughly 30 million variants. The discovery dataset included individuals with NorthWestern European ancestry, excluding 20,000 (10,000 female, 10,000 male) individuals as a validation subset. We generated single-variant association results using SAIGE v1.1.6.3 (33). We used 280,628 markers to fit the null linear mixed model, and age, sex, and the first ten genetic PCs as covariates. To generate PGSs, GWAS results were filtered with a fixed P-value threshold of P-value ≤ 0.01 and minor allele count threshold of MAC ≥ 20. We used PRSice-2 to compute PGSs for all individuals (34). κ values were computed for all same-sex pairs from our validation subset as detailed above in *Computing κ from empirical data*. The match for each pair was computed by comparing the direction of the PGS prediction (sign of the PGS difference for the pair) and the true direction of phenotype difference as reported in the UKB.

To compare our results to the theoretical expectation of the relationship between κ and prediction accuracy, we binned comparisons according to their κ values, and computed the proportion of matching comparisons in each bin. To estimate $\overline{h}$ from the data, we computed κ values for a range of $\overline{h}$ values, and selected the value of $\overline{h}$ that yielded the least sum of absolute distances between the proportion of matches for each bin and the theoretical expectation of prediction accuracy for this κ value, weighted by the number of comparisons per bin (Table S2). We also examined the relationship between κ and prediction accuracy when using heritability estimates derived from the UK Biobank (35) (Fig. S6 and Table S1).

To apply our model to cross-population predictions, we used several PGSs for height published by Yengo *et al.* (13). This meta-analysis of biobank data produced several PGS that were developed using specific populations. We examined three cross-population comparisons: Northwestern European–East Asian ancestries (EUR–EAS), Northwestern European–African ancestries (EUR–AFR) and East Asian–African ancestries (EAS–AFR). Due to the bias towards European ancestry samples in the all-ancestries Yengo PGS (roughly 76% of the samples), we used population-specific GWAS for each comparison (detailed in (13)): East Asian ancestry for EUR–EAS, African ancestry for EUR–AFR, and East Asian ancestry for EAS–AFR (the African ancestry GWAS underperformed for this comparison). Because of the underperformance of the African GWAS, resulting in fewer and noisier loci associated with the phenotype, the EUR-AFR curve is shifted to the left. For the East Asian vs. African analysis, we compared all pairs of individuals from these two groups. For the comparisons of these groups with the European group, we selected random subsets of the European group of the same size as the non-European groups and considered all the same-sex pairwise comparisons between the equal-sized subsets. For all of these cross-population pairwise comparisons, we computed κ and prediction accuracy as explained above.

PGSs are known to have poor transferability between genetically distinct populations. To test the effect

of PGS transferability on our model fit, we used the PGSs from the European ancestry group in Yengo *et al.* (13) to evaluate our predictions in non-European pair comparisons, using the same subsets as indicated above (1,794 individuals with East Asian ancestry for EAS–EAS comparisons, and 3,091 individuals with African ancestry for AFR–AFR comparisons), relative to our pairwise predictions in the European group with the same PGSs (20,000 individuals for EUR–EUR comparisons). Note that in the EUR–EUR comparisons the Yengo *et al.* (13) PGSs do included the tested individuals, but these individuals constitute a very small portion of the overall european population analyzed in this study.

We also generated predictions for a number of common diseases reported in the UKBB. For each disease, we generated single-variant association results using SAIGE2 (33) for binary traits with default parameters. The discovery dataset included individuals with Northwestern European ancestry (as defined by UKB Datafield 22006), excluding 10,000 samples, 5,000 controls and 5,000 cases, as a validation subset. To generate PGSs, GWAS results were filtered with a fixed P-value threshold of P-value ≤ 0.01 and minor allele count threshold of MAC ≥ 20. We used PRSice to compute PGSs for all individuals (34). κ values were computed for random same-sex pairs from our validation subset (only case-control comparisons) as detailed above in *Computing κ from empirical data.* The match between the prediction and the true direction of phenotypic difference for each pair was computed by comparing the direction of the PGS difference and the direction of the phenotype difference (i.e., a correct prediction was one in which the PGS for the disease risk was higher in the case individual).

Analysis of population and species datasets

To evaluate our approach in cases where the compared genomes are more diverged, we examined several datasets. The phenotypes we tested were the phenotypes reported in the different studies.

In the stickleback QTL mapping dataset (16), the comparison was done between a marine population (treated in our analysis as the *phenotyped* population) and four freshwater populations (treated as *unphenotyped*). The compared populations diverged after the end of the last ice age, within the last 20,000 years (16). We investigated 27 morphological phenotypes (measurements of shape landmark coordinates), resulting in four pairwise comparisons of 27 phenotypes. Because not all the phenotypes had significant QTLs in each population, some of the comparisons (three out of four) included fewer than 27 predictions (Fig. 4c).

In the mouse QTL mapping dataset (18), the comparison was done between a wild-derived inbred laboratory house mouse strain (treated in our analysis as the *phenotyped* population) and the Gough island house mouse subpopulation (*unphenotyped* population), which have diverged in the 19th century. Two phenotypes (weight and growth rate) measured across 16 weeks were examined, resulting in a pairwise comparison of 2×16 phenotypes. We then averaged predictions for each of the two phenotypes across the 16 time points.

In the daisy QTL mapping dataset (17), the comparison was done between two daisy species (*Senecio aethnensis*, treated as the *phenotyped* population, and *Senecio chrysanthemifolius*, *unphenotyped*) that have likely diverged within the last 176,000 years (Brennan et al., 2016). For one phenotype out of 13, a prediction could not be made because the sum of the known effects was 0.

The Neanderthal and chimpanzee datasets (6) included comparisons of DNA methylation maps between (i) modern humans (treated as the *phenotyped* population) and Neanderthals (*unphenotyped*), and (ii) between humans (*phenotyped*) and chimpanzees (*unphenotyped*). Because these analyses do not contain effect sizes, they were limited to phenotypes for which the loci with the largest differences in methylation levels showed unidirectionality (likely resulting in high Δ values, and therefore higher κ values). These analyses included predictions of phenotypic direction for 33 Neanderthal phenotypes and 22 chimpanzee phenotypes. We used the prediction accuracy reported in this study.

Acknowledgments

We thank David Reich for the original idea to test this approach with a model, and Dmitri Petrov, Hunter Fraser, Noah Rosenberg, Arbel Harpak, Yuval Simons, Liran Carmel, Jaehee Kim, John (Tony) Capra, Moi

Exposito-Alonso, and members of the Fraser, Petrov, Rosenberg, Greenbaum, and Gokhman labs for input. This research was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center, a research grant from the Center for New Scientists at the Weizmann Institute of Science, and the Kahn Family Research Center for Systems Biology of the Human Cell. This research has been conducted using data from UK Biobank, a major biomedical database, UK Biobank project ID 26664.

References

[1] Rosenberg, N. A., Edge, M. D., Pritchard, J. K. & Feldman, M. W. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evolution, Medicine, and Public Health* **2019**, 26–34 (2019).

[2] Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).

[3] Dittmar, E. L., Oakley, C. G., Conner, J. K., Gould, B. A. & Schemske, D. W. Factors influencing the effect size distribution of adaptive substitutions. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20153065 (2016).

[4] Orr, H. A. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* **6**, 119–127 (2005).

[5] Scheben, A. & Edwards, D. Towards a more predictable plant breeding pipeline with CRISPR/Cas-induced allelic series to optimize quantitative and qualitative traits. *Current Opinion in Plant Biology* **45**, 218–225 (2018).

[6] Gokhman, D. *et al.* Reconstructing Denisovan Anatomy Using DNA Methylation Maps. *Cell* **179**, 180–192.e10 (2019).

[7] Thudi, M. *et al.* Genomic resources in plant breeding for sustainable agriculture. *Journal of Plant Physiology* **257**, 153351 (2021).

[8] Karavani, E. *et al.* Screening Human Embryos for Polygenic Traits Has Limited Utility. *Cell* **179**, 1424–1435.e8 (2019).

[9] Lello, L., Raben, T. G. & Hsu, S. D. Sibling validation of polygenic risk scores and complex trait prediction. *Scientific Reports* **10**, 13190 (2020).

[10] Palmer, D. S. *et al.* Analysis of genetic dominance in the UK Biobank. *Science* **379**, 1341–1349 (2023).

[11] Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[12] Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584–591 (2019).

[13] Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

[14] Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700 000 individuals of European ancestry. *Human Molecular Genetics* **27**, 3641–3649 (2018).

[15] Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).

[16] Rogers, S. M. *et al.* Genetic signature of adaptive peak shift in threespine stickleback. *Evolution* **66**, 2439–2450 (2012).

[17] Brennan, A. C., Hiscock, S. J. & Abbott, R. J. Genomic architecture of phenotypic divergence between two hybridizing plant species along an elevational gradient. *AoB Plants* **8**, plw022 (2016).

[18] Gray, M. M. *et al.* Genetics of rapid and extreme size evolution in Island mice. *Genetics* **201**, 213–228 (2015).

[19] Gokhman, D. *et al.* Human–chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nature Genetics* **53**, 467–476 (2021).

[20] Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* **4**, e1000008 (2008).

[21] Hivert, V. *et al.* Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *bioRxiv* 2020.11.09.375501 (2020).

[22] Pazokitoroudi, A., Chiu, A. M., Burch, K. S., Pasaniuc, B. & Sankararaman, S. Quantifying the contribution of dominance effects to complex trait variation in biobank-scale data. *bioRxiv* 2020.11.10.376897 (2020).

[23] Exposito-Alonso, M., Wilton, P. & Nielsen, R. Non-additive polygenic models improve predictions of fitness traits in three eukaryote model species. *bioRxiv* 2020.07.14.194407 (2020).

[24] Mackay, T. F. Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nature Reviews Genetics* **15**, 22–33 (2014).

[25] Harpak, A. & Przeworski, M. The evolution of group differences in changing environments. *PLoS Biology* **19**, e3001072 (2021).

[26] Hayward, L. K. & Sella, G. Polygenic adaptation after a sudden change in environment. *eLife* **11**, e66697 (2022).

[27] Lakhani, C. M. *et al.* Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nature Genetics* **51**, 327–334 (2019).

[28] Simons, Y. B., Mostafavi, H., Smith, C. J., Pritchard, J. K. & Sella, G. Simple scaling laws control the genetic architectures of human complex traits. *bioRxiv* 2022.10.04.509926 (2022).

[29] Orr, H. A. The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).

[30] Bedford, T. & Hartl, D. L. Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences* **106**, 1133–1138 (2009).

[31] Butler, M. A. & King, A. A. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist* **164**, 683–695 (2004).

[32] Hansen, T. F. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51**, 1341–1351 (1997).

[33] Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018).

[34] Choi, S. W. & O'Reilly, P. F. Prsice-2: Polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).

[35] Walters, R. & Palmer, D. Nealelab/UKBB_ldsc: v2.0.0 (Round 2 GWAS update). *Zenodo* (2022). URL `https://doi.org/10.5281/zenodo.7186871`.