

Predicting the Direction of Phenotypic Difference

Corresponding Author: Dr David Gokhman

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

It's a worthwhile topic to explore, but the authors' focus on predicting whether or not $A > B$, rather than separately predicting the values of A and B, does not appeal greatly to me. If A only slightly exceeds B the condition $A > B$ is satisfied but the difference may not be of practical interest. The authors' given examples of when their approach may be preferred, I don't disagree entirely but am not strongly persuaded. What would be much better is a probabilistic joint prediction: i.e. the output is a joint probability distribution for A and B, from which $\Pr(A > B)$ is easily derived along with many other useful quantities such as $\Pr(A > B + c)$ for any constant c.

Another problem I had was understanding what is meant by a "known" effect/locus which wasn't clearly explained. For complex traits we are used to the idea that relatively few loci are genome-wide significant, and PRS typically consist mainly of loci that are not individually confirmed to be associated with the phenotype so in that sense are not "known". Yet at L135 it seems the authors regard all the SNPs in a PRS as "known", which seems confusing to me as inclusion in a PRS is somewhat arbitrary.

Specific points:

L12 "predictions are hampered by incomplete genotype-to-phenotype mapping" this seems an unnecessarily obscure way to make a simple point. What does "complete" mean here? Guessing its meaning, I'd say it is obviously impossible for mappings to be complete so it's not a useful point. Better to say that predictions are rarely accurate due to limited heritability and widely-dispersed small effects for many traits.

L44 "whether this genome encodes for a larger" another obscure and I think inaccurate way to make a simple point. I think you are predicting whether $A > B$ for two specific individuals, which is often verifiable at a future time. Whether or not a genome "encodes for a larger ..." is impossible to verify and not of much interest.

L53 "... is more informative than predicting only the direction of phenotypic difference. However, the precise phenotypic value can rarely be accurately inferred." Both sentences are correct, but the second sentence is irrelevant to the authors' proposal: there is no gain in accuracy by changing the question to something simpler and less useful.

L228 "disease risk is not directly quantifiable per individual" yes it is! "risk" here can only mean probability which is a quantity. Possibly what the authors mean is that risk is derived from data and models and there is no objective "true" value - but that's a different point (and holds for most statistical modelling).

The manuscript is generally well written. Most minor point ever made by a reviewer: I prefer not to try to pluralise acronyms, so PGS can be singular or plural (it's the same initials in both cases). A more substantial point is that "divergent" is over-used and in my opinion sometimes misused ("different" would be better). L 183 "the probability of predicting ..." is a confusing term: given the data a prediction is deterministic, what is the basis of the probability? This term highlights a failure to clarify whether by "prediction" the authors mean a point value or a probability distribution - the latter is much more useful.

Reviewer #2

(Remarks to the Author)

This is an interesting paper elucidating a clever new (as applied to modern humans) use of polygenic scores. The authors

developed a new test statistic relating PGS to the number of correctly predicted phenotypic difference directions. To support this approach the authors model this statistic under simple conditions and apply their model to several different species. Importantly the authors evaluate how this prediction accuracy is affected by limiting the number of known effects (e.g., applying a PGS developed on one genotyping technology to another), and they helpfully examine some of their simplifying assumptions. The authors also provided some sample code that was extremely helpful in understanding what was being described in the paper. However I think there are a few outstanding questions which could be answered and make this into an even nicer publication. I have separated my thoughts into major concerns, minor concerns, and suggestions.

****major concerns****

(1) There should be some sort of uncertainty/error bar associated with the measurements, plots, and data table entries. When computing κ , $\bar{r}^2$, etc. This can easily be done by resampling data (e.g., a bootstrap sampling). I think this would be of great help to identifying when your results are differing from your theoretically expected curves (e.g., in Fig. 3). It would also be helpful for others to compare to your estimates of $\bar{r}^2$ as this is a similar quantity to a "standard" r^2 or even an approximate linear/SNP heritability.

(2) Paragraph @ line 129: If the model is accurate I think this suggests that the effects (known and unknown) are likely linear. I.e., any non-linear GXG or GxE effects should be small. I agree that there is a large body of evidence supporting that many phenotypes are well-approximated by linear effects, but I wonder if your model has lower accuracy for phenotypes where there are GxE effects (e.g., CAD, obesity, and others in "Canalization of the Polygenic Risk for Common Diseases and Traits in the UK Biobank Cohort" by Nagpal et al., "Dual exposure-by-polygenic score interactions highlight disparities across social groups in the proportion needed to benefit" by Nagpal & Gibson, and "GxE PRS: Genotype-environment interaction in polygenic risk score models for quantitative and binary traits" by Jayasinghe et al.). You eventually address this with one line in the Discussion about epistasis, but I think some more discussion about when one should expect deviations from this model would be appropriate. It could also be addressed in the introduction when you discuss the limitation of PGS (i.e., around line 38).

(3) The language about how κ and/or $\bar{r}^2$ is determined from real data can be confusing and even appears circular at times. For example, in the description in "Analysis of pairwise comparisons in humans", the language reads as if: "(a) to compute κ we need to estimate $\bar{r}^2$. (b) Next we compute κ and use this to estimate $\bar{r}^2$ ".

The code example was a great help in clearing this up, but I think the language in the draft could be changed to make things clearer.

(4) Fig 3: I'm not sure it's explicitly described in the text (I might have missed it), but do the different data points in (a)-(c) correspond to only keeping the top x% of loci? I thought that once you have a PGS, you compute the PGS differences and phenotype differences and then fit to an expected/theoretical percentage correct to get a single data point? The text seems to hint that these data points might correspond just to the pairs of subjects with phenotypic/pgs differences corresponding to that κ value? I think this is more clear when discussing non-human results ala Fig 4.

****minor concerns****

General: Do I understand the general proposed use case for this research: when applying a PGS to a new population you can compute Δ , then take an estimate of $\bar{r}^2$ from a similar population to estimate κ , which in turn gives you a probability of knowing the phenotypic difference direction?

line 94: does "unknown effects" contain the environment and covariates?

Figure 1: This is a fantastic figure! Any chance that you can increase the font sizes? It's very hard to read some of the numbers. (This is true for some of the other figures as well)

Fig 2 and results section: I'm a little confused about how the known effect sizes were determined. What does "those with the largest effect sizes" mean? Sometimes the largest feature weight (e.g. PGS beta value for a SNP) can be different from the weights which contribute the most to the variance of a phenotype. (E.g., this can happen because of MAF differences.) In the main text you seem to indicate you are speaking about largest contributions to the phenotype (e.g., ¶ at line 168), and I just want to double check this is the correct interpretation.

Fig 2: it's very hard to see the difference between the curves in (a). Can you compute the other light/medium blue points above $k=0.6$ as well or is this the limit because of the limited number of loci?

¶ at line 168: you talk about "prediction accuracy" but it seems you are only comparing it to simulated data. Would this be more properly described as an "expected prediction accuracy" or something along those lines? (I.e., because you use both real and simulated data in this paper I think it would be good to have more clearly distinct labels.)

Fig 3: Panel d seems to show κ values that go above the κ values in a. Why the discrepancy?

The UKB does include some medication data. Can you check if any of the individuals that contributed to the computation of $\kappa = 0$ for hypertension are using any of the common medications for hypertension: ace inhibitors, ARBs, beta blockers, diuretics, and calcium channel blockers. While not definitive this would bolster the conclusion that medication is playing a role.

In derivation 2 (alternative derivation) of distribution: can e_i be correlated because of shared environments? I.e. you are assuming the e_i don't have environmental correlations either.

line 507 "subsets of comparisons" -> "subsets for comparisons"?

Methods: ¶ at 518: why was PRSice-2 chosen to build PGS? Why were these particular parameter values selected? I think a comparison of different input PGS is probably outside the feasible scope of this paper, but I think some indication of why the method was chosen should be included. (Even if it's as simple as: PRSice-2 is easy to (re)run which lends itself to building a new method like this.) This might lead to a larger discussion, but I ran your $\bar{r}^2$ estimation pipeline on my own height phenotype with a LASSO based PGS in the UKB. I used 10,000 samples (5k male, 5k female) and I found $\bar{r}^2 \sim 0.41$. You might want to mention that these estimates are highly dependent on the method used to generate the PGS. (It is of course much more tricky and time resource intensive to train many different types of PGS. And if you just download from the PGS catalog UKB results the results can be biased due to possible overtraining/contamination from the PGS training samples.) This is related to the first "minor concern" mentioned above.

Line 543: do you know how many (or approximately) individuals overlap between these sets? We have found (not sure if there is a good reference for this) that even having a small number of people in the GWAS step appear in the final testing set can significantly inflate typical PGS metrics (e.g. AUC, r^2 , etc)

¶ at line 545: there are many additional cases in ICD9 and self report (should look up some examples). Is there a reason these were excluded?

It would be interesting to see phenotypes with $\bar{r} \sim 0.1$ and $\bar{r} \sim 0.9$ as examples of highly environmental and highly genetic traits (i.e., a stand in for heritability). Did you test any traits that were in the extremes?

****Stylistic suggestions: The following are only suggestions and do not need to be addressed****

In the reviewer's experience many other readers and reviewers like less concrete and more mathematical conclusions, i.e. they don't like broad statements. E.g. in the summary of section "Identifying which individual has the higher phenotypic value in real-world data" starting at line 266 you can give an example instead of just stating "we are able to accurately estimate the chances ...". I think this information is clearly obtainable by reading your plots and data, but I know some readers like to see this in the text.

Line 129: I think the opening sentence of this paragraph is superfluous.

Eq. 1: maybe this is pedantic, but isn't your metric really "known-to-total" and not "known-to-unknown"?

Reviewer #3

(Remarks to the Author)

The authors describe an approach to estimate the probability that one individual will have a greater phenotypic value than another. The method relies on assumptions of normally distributed phenotypes, and an estimate of the variance in the phenotype that is unaccounted for by a genetic model. They show using individuals from the UK biobank that the method is well calibrated, and that it is calibrated even in cross-population or cross-species comparisons.

Overall, the method is closely related to polygenic risk score calculations (which this method draws on), or genomic predictions from plant and animal breeding. Most applications of those methods do not explicitly compute an uncertainty window around their predictions, so the probability of two predictions being different in a specific direction can't be easily calculated from standard results. However, note that methods for calculating prediction error variances or posterior distributions are straightforward e.g. using GBLUP or BayesCpi for predictions, and that would give a more accurate estimate than the method here because it would account for any correlations in predictions between the two compared individuals. The method also seems related to the idea of a local false sign rate (Stephens et al 2017), which might be a useful comparison.

My main concern with this paper is that I'm still not convinced that most of the application ideas discussed here are actually useful in any way. Why would a patient care if their probability of a disease was 0.001% higher vs 0.001% lower than another's? This method would treat those two cases as completely different, but they're really practically the same. In medicine, I think only the actual phenotypic value (probability of a disease) is useful. In medical insurance also, it is the magnitude of the difference between groups, not the probability that one is greater than another that matters. I think the method could be adjusted to give the probability that a trait will be greater than a pre-specified threshold in a straightforward way which might be more useful. The same arguments apply in breeding (if a line has a 95% chance of giving a higher yield, but the magnitude of that effect is 0.001 mg, I won't be very excited about selecting it for crosses). Same with embryo selection. I think this topic in particular is highly controversial, and so without careful consideration of ethical implications, discussing this as a potential application seems unsafe.

From the statistical side, I am glad the authors considered possible sources of correlation between the predicted and

unpredictable sources of variation, as those are clearly possible. One case that I didn't see discussed is where loci are discovered using a model that includes PCs or a kinship matrix to control for effects of population structure. These terms will bias the model towards loci that are more diverged within populations than between populations. I would guess that if loci are discovered using such a model, the calibration for comparisons between individuals in distinct populations will be different than for individuals within populations.

A minor note: the derivation of $k = \Delta/(\Delta + \sigma)$ seems unnecessarily complex. Since only the ratio $k/(1-k)$ is used, the method would be the same if they simply used Δ/σ , ie the predicted effect measured in units of the residual standard deviation. Phrased as such, the probability is simply the cdf of a standard normal evaluated at Δ/σ .

I found the section on $\bar{r}^2$ hard to follow. Why not just use the s^2_e estimate from the model used to fit the genetic model (ie make the PGS)?

I found the descriptions of the non-human analyses incomplete. It looks like QTL mapping results were used, ie biparental populations created by crossing individuals from each population. Were the tests evaluated on pairs of individuals from these biparental populations, or individuals from the original (non-admixed) populations?

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I thank the authors for their responses to my first set of comments.

I appreciate that you have made an attempt to explain what you mean by "known" effects (L101-102), but it's still not satisfactory. Since there are no rules for what is "known" ("assumed to be associated with the phenotype" adds nothing to a definition, it does help with intuition). The "knowns" can be any arbitrary set of genetic variants but for the method to work well they should be a good predictors of phenotype. I'd suggest something like "We distinguish the contribution to the phenotype of a chosen set of genotyped variants that we will refer to as 'known', from all other genetic and environmental effects". Or maybe since in practice the knowns are a PGS maybe just call it that from the start as

It seems to me essential to have a discussion here (with perhaps more later) of what makes a good choice for the "known" variants. You later consider "... known effects are those with the largest effects" but wouldn't it be better (and more common) to use the variants that pass a p-value threshold? Further, with modern PRS methods that simultaneously estimate effect sizes for all SNPs using shrinkage why wouldn't you include all genotyped SNPs as known? It's unsatisfactory that all knowns are treated the same irrespective of the uncertainty in effect size estimates, but maybe shrinkage estimates can deal adequately with this: the less information (due to low MAF) the greater the shrinkage. PRSice-2 may still be widely used but it is not methodologically state-of-art: advanced PRS like DBSLMM or MegaPRS can do much better

You now clarify (L114) that "the same approach could be applied to test if the difference between the phenotypes exceeds a predetermined threshold of interest" and you also say that it can be applied to compare an individual with a fixed value, without reference to a comparison individual. But doing that for every possible threshold corresponds to specifying a probability distribution for the unknown phenotype, which I previously said would be much more useful than just comparison with a reference individual (another reviewer said something similar). It seems worth highlighting this. In practice it's not feasible to compute probabilities for every threshold, but you can approximate the probability distribution function to arbitrary accuracy by increasing the grid of thresholds considered.

L158 "probability describing prediction accuracy" this is too vague, there are multiple probabilities that can measure prediction accuracy, better to be explicit.

Methods: why is this section in a tiny font? Why bother defining the E as binary and then invoke the CLT? Better to just assume Gaussian error as is standard.

Minor comments:

L62 "estimate the likelihood" in a statistics-heavy paper you should only use "likelihood" in its technical sense. Here, you mean "probability". A bit more subtle, but "estimate" is not really correct here either - there is no true parameter that you are trying to estimate. You compute or approximate a probability under given assumptions.

L487 typo "an used"

Reviewer #2

(Remarks to the Author)
(Reviewer #2 Tim Raben)

I thank the authors for the extensive and careful work they put into revising this manuscript and addressing all the previous comments: their efforts are greatly appreciated. I only have a couple of minor details to bring up:

* Although I believe this to be a relatively minor detail, I do think it would be helpful to have some sort of uncertainty measurement for the r^2 ($\bar{r^2}$) values in table S1 and S2. I raise this issue because while it seems likely that, for example, the r^2 and $\bar{r^2}$ values agree for EUR BMI, it is not clear if the difference between r^2 and $\bar{r^2}$ for standing height is meaningful (vs a statistical fluctuation). Similarly for hip circumference (EUR), metabolic rate (siblings), and BMI (siblings). I think this can be quickly done by a simple bootstrap (e.g. reselecting your inference and validation sets) or even by sub-sampling the groups you already have.

* Missing period at the end of line 186.

Very minor suggestion that doesn't need to be addressed for this manuscript to proceed to publication:

In Methods "Alternate Derivation" you use e_i as the residual (phenotype minus PGS) which are used to determine the scale σ . However, in the following section "Simulation" you used e_i for both genetic and non-genetic contributions (used to determine both Δ and D), and in "Analysis of pairwise comparisons in humans" you use e_i again to mean something slightly different (additive effect of single alleles to the phenotype) which is used to determine Δ . All the while, much of the other notation the same. This is all clear in context, but can get confusing when jumping around.

Reviewer #3

(Remarks to the Author)
The authors have addressed my technical concerns.

I remain largely unconvinced that predicting the direction of change as described here is useful. Even with the +c threshold, a value 0.01% above this new threshold would still be treated completely different than a value 0.01% below that new threshold, so it's just pushing the goalposts, but not addressing the criticism that thresholds themselves like this are not that useful for individual-level decision-making. Only effect size estimates are.

Also, I remain concerned about the discussion about embryo selection. The authors added a caveat saying this approach isn't recommended, but the rest of the paper makes it sound like it is likely to be very useful to "sufficient to substantially reduce the chances of exceeding the disease threshold.". Given all the uncertainty in GxE, additivity, etc discussed above, is this really a statement that the authors or the journal want to make in a paper when it can have such major ethical implications if mis-used?

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed my previous set of comments. Like Reviewer 3, I remain a little sceptical about the importance of the question addressed but I think that hurdle was passed with the journal's decisions to invite revisions and I am happy for the paper to proceed to publication. I have only two minor comments:

I note from the response to Rev 3 that a paragraph has been deleted but this is not shown as a deletion in the version that supposedly marks up changes made to the last version. In fact no deletions are indicated only additions to the text making it difficult for me to fully assess the changes made. I didn't support the comments of Rev 3 that led to this deletion, but of course if the authors are happy to delete then that's fine.

Apparently there has been an improvement in font size since my previous comment, but I don't understand why in a methods paper the Methods are in a small font and with narrow line spacing, suggesting that these are unimportant minor details. At least they are now legible.

Reviewer #2

(Remarks to the Author)

Thank you for the considerate revisions. I have no further comments. --Tim Raben

Reviewer #3

(Remarks to the Author)

Thank you for your responses. I have no further comments.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer #1 (Remarks to the Author):

It's a worthwhile topic to explore, but the authors' focus on predicting whether or not $A > B$, rather than separately predicting the values of A and B, does not appeal greatly to me. If A only slightly exceeds B the condition $A > B$ is satisfied but the difference may not be of practical interest. The authors' given examples of when their approach may be preferred, I don't disagree entirely but am not strongly persuaded. What would be much better is a probabilistic joint prediction: i.e. the output is a joint probability distribution for A and B, from which $\Pr(A > B)$ is easily derived along with many other useful quantities such as $\Pr(A > B + c)$ for any constant c.

We thank the reviewer for the helpful, constructive, and specific suggestions. We agree that adjusting the method to give the probability that a trait exceeds a certain threshold is helpful and illuminating, and that incorporating this adjustment significantly enhances the paper's applicability and practical relevance. Following the reviewer's feedback, we derived the probability $P(A > B + c)$ as a function of the sum of known effects (or PRS difference) between individuals A and B (new Eq. (7)). To illustrate it, we included a visualization in the new Fig. 2c, where the probability is shown for various values of c. This figure highlights how increasing values of c reduce prediction accuracy, demonstrating the trade-offs involved.

Furthermore, we revised the abstract and introduction to emphasize that predicting the direction of phenotypic differences is more helpful in scenarios where even small differences are informative. For example, comparing an unphenotyped individual to one already known to have a high risk can be actionable, as even a slightly higher risk indicates that the unphenotyped individual is also likely at high risk. We clarified this point, as well as the addition of the threshold approach the reviewer suggested with the following additions: "Importantly, the same approach could also be applied to examine if the tested individual has a higher phenotypic value than an individual confirmed to have the phenotype (e.g., an individual with the disease), or a threshold of interest (e.g., the unphenotyped crop will produce at least 10% more yield than another crop)." (L52-56), "or that the phenotype of two individuals differs by at least a predetermined constant" (L63-64), "predict if an individual's disease risk exceeds a baseline (e.g., the population average or a set clinical threshold) or that of other individuals with the disease." (L71-73), "We provide two derivations - one from the viewpoint of random walks and the other from the viewpoint of statistical genetics, which also enabled us to model shared genetic and environmental components in siblings, and to formulate the probability that the difference between two individuals exceeds a certain threshold of interest" (L160-164), "As expected, adding a threshold c for the phenotypic difference in our prediction reduces prediction accuracy (Fig.2c). However, the reduction in prediction accuracy compared to the no-threshold prediction becomes substantial only when the threshold c is very large. For example, for a threshold $c=0.1$ (0.1 standard deviations of the phenotypic values), the reduction in prediction accuracy is 1.5-5.8% for the delta values we simulated, whereas this reduction is 21-53% for the threshold $c=1$ (a full standard deviation)" (L180-186), and "In sum, we found that the known-to-total ratio (κ) captures the factors that affect the probability of correctly predicting which individual has

the higher phenotypic value (and that they can be used to predict if the difference is higher than any defined threshold)." (L201-204).

Another problem I had was understanding what is meant by a "known" effect/locus which wasn't clearly explained. For complex traits we are used to the idea that relatively few loci are genome-wide significant, and PRS typically consist mainly of loci that are not individually confirmed to be associated with the phenotype so in that sense are not "known". Yet at L135 it seems the authors regard all the SNPs in a PRS as "known", which seems confusing to me as inclusion in a PRS is somewhat arbitrary.

We agree that it is important to clarify early on what we mean by "known effects". We now added an explicit definition to clarify this at the beginning of the results section L100-104). Because we want to keep the description of the approach as general as possible, allowing it to be relevant to many different types of data (such as the human population genetic application using PGS in Fig. 3 and application to between-species comparisons in Fig. 4), we don't specify at this stage of the manuscript how we come about to the conclusion that a locus is "known" to affect a phenotype. The reviewer is correct to point out that in the context of complex traits studied in population genetic datasets, such as in the analysis in Fig. 3, the notion of "known" is not straightforward, and it typically requires a procedure for evaluation of significance of effects to decide which loci are included in the PGS. This means that different parameterizations and methods can lead to different sets of loci being assigned to the "known" loci group. In the section that the reviewer is referring to in this comment, we discuss simulations, in which case we simply define a set of loci that is "known", without the need to consider problems that are specific to inclusions of SNPs in PGS.

Specific points:

L12 "predictions are hampered by incomplete genotype-to-phenotype mapping" this seems an unnecessarily obscure way to make a simple point. What does "complete" mean here? Guessing its meaning, I'd say it is obviously impossible for mappings to be complete so it's not a useful point. Better to say that predictions are rarely accurate due to limited heritability and widely-dispersed small effects for many traits.

Following the reviewer's suggestion, we rephrased this sentence to "for most complex phenotypes, predictions are rarely accurate due to limited heritability, widely-dispersed small effects, and more".

L44 "whether this genome encodes for a larger" another obscure and I think inaccurate way to make a simple point. I think you are predicting whether $A > B$ for two specific individuals, which is often verifiable at a future time. Whether or not a genome "encodes for a larger ..." is impossible to verify and not of much interest.

We rephrased the sentence to "we aim to predict whether this individual has/will have a larger or smaller phenotype than another individual".

L53 "... is more informative than predicting only the direction of phenotypic difference. However, the precise phenotypic value can rarely be accurately inferred." Both sentences are correct, but the second sentence is irrelevant to the authors' proposal: there is no gain in accuracy by changing the question to something simpler and less useful.

We agree that predicting the direction of phenotypic difference does not improve accuracy, and that the current phrasing of this part was confusing. The goal of these sentences was to emphasize that one can still learn important information about the phenotype even when precise predictions are unavailable. To clarify the text, we rephrased the sentences to: "Despite being less informative than a precise phenotypic value, the direction of phenotypic difference is often the crux of phenotypic comparisons (Young2019, Thudi 2021, Karavani 2019, Lello 2020), and perhaps could be predicted with high accuracy."

L228 "disease risk is not directly quantifiable per individual" yes it is! "risk" here can only mean probability which is a quantity. Possibly what the authors mean is that risk is derived from data and models and there is no objective "true" value - but that's a different point (and holds for most statistical modelling).

We agree. We removed this sentence.

The manuscript is generally well written. Most minor point ever made by a reviewer: I prefer not to try to pluralise acronyms, so PGS can be singular or plural (it's the same initials in both cases). A more substantial point is that "divergent" is over-used and in my opinion sometimes misused ("different" would be better). L 183 "the probability of predicting ..." is a confusing term: given the data a prediction is deterministic, what is the basis of the probability? This term highlights a failure to clarify whether by "prediction" the authors mean a point value or a probability distribution - the latter is much more useful.

We appreciate both major and minor comments! We (a) changed all occurrences of PGSs to PGS, (b) replaced all occurrences of "divergent loci" with "loci that differ in their genotype", and (c) we clarified the terminology in L202 (which was indeed imprecise, as pointed out by the reviewer), to "the probability of correctly predicting which individual has the higher phenotypic value." Throughout the paper, we discuss the probability of a prediction being correct, not a probability distribution.

Reviewer #2 (Remarks to the Author):

This is an interesting paper elucidating a clever new (as applied to modern humans) use of polygenic scores. The authors developed a new test statistic relating PGS to the number of correctly predicted phenotypic difference directions. To support this approach the authors model this statistic under simple conditions and apply their model to several

different species. Importantly the authors evaluate how this prediction accuracy is affected by limiting the number of known effects (e.g., applying a PGS developed on one genotyping technology to another), and they helpfully examine some of their simplifying assumptions.

The authors also provided some sample code that was extremely helpful in understanding what was being described in the paper.

However I think there are a few outstanding questions which could be answered and make this into an even nicer publication. I have separated my thoughts into major concerns, minor concerns, and suggestions.

We thank the reviewer for the positive review and very helpful suggestions, which we addressed as described below.

****major concerns****

(1) There should be some sort of uncertainty/error bar associated with the measurements, plots, and data table entries. When computing κ , $\bar{r}^2$, etc. This can easily be done by resampling data (e.g., a bootstrap sampling). I think this would be of great help to identifying when your results are differing from your theoretically expected curves (e.g., in Fig. 3). It would also be helpful for others to compare to your estimates of $\bar{r}^2$ as this is a similar quantity to a "standard" r^2 or even an approximate linear/SNP heritability.

We agree. The main direction in which we evaluated the variability in our estimation of the probability is by comparing the theory to several different phenotypes (e.g., Fig. 3b). However, following this comment, we also added confidence intervals to all relevant figures (Fig.2a, Fig.3a-c, Fig.4b, Fig.S2), as suggested.

(2) Paragraph @ line 129: If the model is accurate I think this suggests that the effects (known and unknown) are likely linear. I.e., any non-linear GXG or GxE effects should be small. I agree that there is a large body of evidence supporting that many phenotypes are well-approximated by linear effects, but I wonder if your model has lower accuracy for phenotypes where there are GxE effects (e.g., CAD, obesity, and others in "Canalization of the Polygenic Risk for Common Diseases and Traits in the UK Biobank Cohort" by Nagpal et al., "Dual exposure-by-polygenic score interactions highlight disparities across social groups in the proportion needed to benefit" by Nagpal & Gibson, and "GxE PRS: Genotype-environment interaction in polygenic risk score models for quantitative and binary traits" by Jayasinghe et al.). You eventually address this with one line in the Discussion about epistasis, but I think some more discussion about when one should expect deviations from this model would be appropriate. It could also be addressed in the introduction when you discuss the limitation of PGS (i.e., around line 38).

We have now extended the discussion on the effects of interactions on our results (L381-389). In some sense, at least in the context of using population-level datasets, the limitations induced by assuming linearity in PGS also apply to our approach.

However, although prediction accuracy is expected to be affected by non-linearity in our approach, we do not observe that different phenotypes have a markedly different fit to the theoretical expectation of the relationship between κ and P (Fig.3b). We now clarify this point, and specifically highlight the potential difficulties that may arise in estimating σ when $G \times E$ effects are substantial. We also added this point at an early stage in the introduction (L40–41), as suggested.

(3) The language about how κ and/or $\bar{r}^2$ is determined from real data can be confusing and even appears circular at times. For example, in the description in "Analysis of pairwise comparisons in humans", the language reads as if: "(a) to compute κ we need to estimate $\bar{r}^2$. (b) Next we compute κ and use this to estimate $\bar{r}^2$ ".

The code example was a great help in clearing this up, but I think the language in the draft could be changed to make things clearer.

We agree that this introduction of $\bar{r}^2$ was confusing. We have now revised the text to better explain and clarify what r^2 and $\bar{r}^2$ denote, and how they relate to each other (L541-548).

(4) Fig 3: I'm not sure it's explicitly described in the text (I might have missed it), but do the different data points in (a)-(c) correspond to only keeping the top x% of loci? I thought that once you have a PGS, you compute the PGS differences and phenotype differences and then fit to an expected/theoretical percentage correct to get a single data point? The text seems to hint that these data points might correspond just to the pairs of subjects with phenotypic/pgs differences corresponding to that κ value? I think this is more clear when discussing non-human results ala Fig 4.

The reviewer's comment helped us realize that the text was not clear enough regarding this point. The reviewer's description of the approach is correct in that each data point shows the prediction accuracy (proportion of correct predictions) for the set of individual pairs with κ values in a certain bin. In other words, for each pair of individuals, we compute a PGS difference and use an r^2 value (computed on a different set of individuals than our test set) to compute the κ value; we then bin the pairs based on these κ values and compute the proportion of correct predictions. To clarify this, we added the following sentence to the legend: "Each data point shows the proportion of correct predictions for pairs of individuals in a certain κ bin.". We also added relevant clarifications in L545-546 and L587-593.

****minor concerns****

General: Do I understand the general proposed use case for this research: when applying a PGS to a new population you can compute Δ , then take an estimate of $\bar{r}^2$ from a similar population to estimate κ , which in turn gives you a probability of knowing the phenotypic difference direction?

Indeed, this is an accurate description of the approach.

line 94: does "unknown effects" contain the environment and covariates?

Indeed, unknown effects include environmental effects (as well as unknown/unsampled genetic effects). We changed the text to reflect this (L100-104). The question of covariates/epistasis is slightly more complex, and we thus elaborated on it in the Discussion (L381-389).

Figure 1: This is a fantastic figure! Any chance that you can increase the font sizes? It's very hard to read some of the numbers. (This is true for some of the other figures as well)

Following the reviewer's suggestion, we increased font sizes.

Fig 2 and results section: I'm a little confused about how the known effect sizes were determined. What does "those with the largest effect sizes" mean? Sometimes the largest feature weight (e.g. PGS beta value for a SNP) can be different from the weights which contribute the most to the variance of a phenotype. (E.g., this can happen because of MAF differences.) In the main text you seem to indicate you are speaking about largest contributions to the phenotype (e.g., at line 168), and I just want to double check this is the correct interpretation.

Indeed, "largest effect" could have more than one interpretation, and we realize that this requires further clarification in the text. In our case, we meant the largest feature weight (e.g., PGS beta value). We therefore changed the text to "largest effects (e.g., largest beta values in GWAS data)" to exemplify this.

Fig 2: it's very hard to see the difference between the curves in (a). Can you compute the other light/medium blue points above $k=0.6$ as well or is this the limit because of the limited number of loci?

Indeed, the curves are entirely overlapping in this panel, even when we increase panel size, supporting the fit between the theoretical expectation and simulated data. As the reviewer conjectured, there are no points beyond $k=0.6$ in the light blue line because there were no data points reaching this value. We now state this in the legend: "Only data points with > 100 pairwise comparisons are shown, and in (d)–(f) all pairwise comparisons are shown".

at line 168: you talk about "prediction accuracy" but it seems you are only comparing it to simulated data. Would this be more properly described as an "expected prediction accuracy" or something along those lines? (I.e., because you use both real and simulated data in this paper I think it would be good to have more clearly distinct labels.)

Yes, thank you for highlighting this point. The more accurate term would be "the theoretical expectation for prediction accuracy", and we have now rephrased this sentence accordingly (L172).

Fig 3: Panel d seems to show κ values that go above the κ values in a. Why the discrepancy?

Indeed. In (a)-(c) we computed prediction accuracy when there were sufficient pairwise comparisons for that kappa value bin (>100 pairwise comparisons). In (d)-(f), however, we used all available data to present the full distribution of kappa values. This was not explained in the legend, which created a confusing discrepancy. Therefore, we added this information at the end of the caption of figure 3 for clarification.

The UKB does include some medication data. Can you check if any of the individuals that contributed to the computation of $\kappa = 0$ for hypertension are using any of the common medications for hypertension: ace inhibitors, ARBs, beta blockers, diuretics, and calcium channel blockers. While not definitive this would bolster the conclusion that medication is playing a role.

Thank you for this suggestion. We completely agree that it would be informative to examine our hypothesis regarding medication as an explanation for the interesting results we have for hypertension. Our results indicate that in pairs assigned to the high-kappa bins, the control individual was more likely to be the one with the high PGS, rather than the disease individual (prediction accuracy approaches 0 for high kappa values, whereas “random noise” should result in prediction accuracy 0.5). For the medication to be driving this, we would expect that the medication use in the control individuals in the high-kappa bins would be higher than expected (i.e. higher than the medication use in control individuals in low- and mid- kappa bins). We tested this using data from UKB Data Field 6177, which reports antihypertensive use (without specifying which medications). We find that the overall fraction of antihypertensive use in the dataset is 6.3% (315 out of 5000 control individuals). However, in contradiction with the hypothesis above, we find that in high kappa bins the fraction of medication use in control individuals is substantially lower than 6.3%. Therefore, we cannot use this result to provide support for our hypothesis that medication use is the explanation for the low P values at high kappa values. However, this does not rule out other interventions (such as diet/life-style alterations) as potential explanations. Following this comment, we have broadened the scope of the hypothesis to include all types of interventions, and described the negative result regarding antihypertensive medication (L267-271, L618-620).

In derivation 2 (alternative derivation) of distribution: can e_i be correlated because of shared environments? I.e. you are assuming the e_i don't have environmental correlations either.

In the alternative derivation in the main text, we assumed that e_1 and e_2 are independent and uncorrelated, but in the "Siblings" section in the Supporting Information, we relax this assumption and show a derivation where the shared environment component is explicitly modeled.

line 507 "subsets of comparisons" -> "subsets for comparisons"?

This wording was unclear. We changed the text to “We compared pairs of individuals with various levels of genetic divergence”.

Methods: at 518: why was PRSice-2 chosen to build PGS? Why were these particular parameter values selected? I think a comparison of different input PGS is probably outside the feasible scope of this paper, but I think some indication of why the method was chosen should be included. (Even if it's as simple as: PRSice-2 is easy to (re)run which lends itself to building a new method like this.) This might lead to a larger discussion, but I ran your $\bar{r}^2$ estimation pipeline on my own height phenotype with a LASSO based PGS in the UKB. I used 10,000 samples (5k male, 5k female) and I found $\bar{r}^2 \sim 0.41$. You might want to mention that these estimates are highly dependent on the method used to generate the PGS. (It is of course much more tricky and time resource intensive to train many different types of PGS. And if you just download from the PGS catalog UKB results the results can be biased due to possible overtraining/contamination from the PGS training samples.) This is related to the first “minor concern” mentioned above.

As noted by the reviewer, our general approach in this manuscript was to use standard tools with standard parameters, rather than attempt to optimize prediction accuracy. We fully agree that the estimates for prediction accuracy and r^2 depend on the methods and pipelines used, and that better results might be achieved by tailoring the method to the specific data types and phenotypes analyzed. Our main objective was to demonstrate that PGS across various traits, applied to populations with differing levels of divergence, exhibit a consistent alignment with our theoretical model describing the relationship between kappa and prediction accuracy. To achieve this, we used the widely adopted PRSice-2 and adhered to its standard parameters to minimize concerns about whether parameter selection influenced our results and to ensure the robustness of our findings. Following the reviewer's suggestion, we now make this choice clear in L578-583.

Line 543: do you know how many (or approximately) individuals overlap between these sets? We have found (not sure if there is a good reference for this) that even having a small number of people in the GWAS step appear in the final testing set can significantly inflate typical PGS metrics (e.g. AUC, r^2 , etc)

We agree that this could be informative but, unfortunately, we cannot check how many individuals overlap in the two subsets because UKB sample IDs are not reported with summary statistics in the Yengo et al. data. We agree with the reviewer that this overlap may generate issues in inflating r^2 , as we mention in L601-610. However, our goal was to determine the fit of the empirical prediction accuracy to our theory, and this is not expected to be substantially affected by the inflation of r^2 . This can be seen when comparing the Yengo et al. result in Fig.3c (dark brown) to the non-inflated result in Fig.3b (magenta), both for European individuals (L601-610).

at line 545: there are many additional cases in ICD9 and self report (should look up some examples). Is there a reason these were excluded?

The vast majority of cases in the UK Biobank for the diseases we investigated were coded with ICD10, for example 162,262 (ICD10 code I10) vs. 204 (ICD9 code 401) for hypertension, 51,715 (ICD10 code J45) vs. 208 (ICD9 code 493) for asthma, 53,211 (ICD10 code E11; type II) vs. 229 (ICD9 code 250) for diabetes, and 31,806 (ICD10 code E03) vs. 35 (ICD9 code 244) for hypothyroidism. In general, only 4% of participants in the UKB have ICD9 codes, and of these 80% also have ICD10 codes, leaving just ~3000 participants in the UK Biobank that have exclusively ICD9 codes. In addition, ICD10 codes are more detailed, allowing us to differentiate, for example, between cases of diabetes types I and II. Because the ICD9 codes are older and less specific, and their appearance in the UKB appears to be sporadic, we decided to consider only the ICD10 codes.

It would be interesting to see phenotypes with $\bar{r} \sim 0.1$ and $\bar{r} \sim 0.9$ as examples of highly environmental and highly genetic traits (i.e., a stand in for heritability). Did you test any traits that were in the extremes?

We agree with the reviewer's suggestion regarding the selection of phenotypes. In this study, we prioritized phenotypes that are easily measured, widely available in the UKB, and well-studied within the field, ensuring their robustness and relevance. The phenotypes we used exhibit r values ranging from $r=0.17-0.5$ ($r^2=0.03-0.25$), which represent a relatively broad range. Following the reviewer's comment, we explored the possibility of including phenotypes with more extreme r values. Since r values are not directly reported in the UKB, we instead examined h^2 values as an upper bound. However, we were unable to identify a well-studied polygenic trait with sufficient sample size and such extreme values within the UKB dataset.

****Stylistic suggestions: The following are only suggestions and do not need to be addressed****

In the reviewer's experience many other readers and reviewers like less concrete and more mathematical conclusions, i.e, they don't like broad statements. E.g. in the summary of section "Identifying which individual has the higher phenotypic value in real-world data" starting at line 266 you can give an example instead of just stating "we are able to accurately estimate the chances ...". I think this information is clearly obtainable by reading your plots and data, but I know some readers like to see this in the text.

We agree with the reviewer's suggestion regarding adding numeric statements to the manuscript, which we also tend to find useful. We therefore tried, throughout the manuscript, to balance between providing specific examples and referring to the plots, and broader statements that convey to the readers the generalized conclusions. Following the reviewer's comment, we added numeric examples to the following lines:

L183-186 – “For example, for a threshold $c=0.1$ (0.1 standard deviations of the phenotypic values), the reduction in prediction accuracy is 1.5-5.8% for the delta values we simulated, whereas this reduction is 21-53% for the threshold $c=1$ (a full standard deviation)”

L282-283 – “we observed lower kappa values for these comparisons relative to the kappa distribution in Europeans (between East Asian pairs: 33% lower on average, and between African pairs: 57% lower on average”.

Line 129: I think the opening sentence of this paragraph is superfluous.

Indeed. We removed this sentence.

Eq. 1: maybe this is pedantic, but isn't your metric really "known-to-total" and not "known-to-unknown"?

Following the reviewer's suggestion, we changed all occurrences of this term to “known-to-total”.

Reviewer #3 (Remarks to the Author):

The authors describe an approach to estimate the probability that one individual will have a greater phenotypic value than another. The method relies on assumptions of normally distributed phenotypes, and an estimate of the variance in the phenotype that is unaccounted for by a genetic model. They show using individuals from the UK biobank that the method is well calibrated, and that it is calibrated even in cross-population or cross-species comparisons.

We thank the reviewer for the helpful feedback and suggestions, which have contributed substantially to improving the manuscript.

Overall, the method is closely related to polygenic risk score calculations (which this method draws on), or genomic predictions from plant and animal breeding. Most applications of those methods do not explicitly compute an uncertainty window around their predictions, so the probability of two predictions being different in a specific direction can't be easily calculated from standard results. However, note that methods for calculating prediction error variances or posterior distributions are straightforward e.g. using GBLUP or BayesCpi for predictions, and that would give a more accurate estimate than the method here because it would account for any correlations in predictions between the two compared individuals. The method also seems related to the idea of a local false sign rate (Stephens et al 2017), which might be a useful comparison.

Indeed, while our paper is focused on PGS, our research question would equally apply to other prediction methods. In other words, we would still be concerned with the probability that the traits of two individuals are sorted in the same order as their

predicted values. To the best of our understanding, methods such as GBLUP or BayesCpi provide the posterior distribution for the genomic estimated breeding values, not the trait itself. To obtain posterior samples for the trait, some assumptions will have to be made about the distribution of the residuals, which would essentially lead to similar theoretical predictions as we already have in the paper. Performing simulations with these tools would be beyond the scope of this paper, but we agree that this would be of interest and should be addressed in the future. Thus, following the reviewer's comment, we added the following text in L399-407: "The prediction accuracy can also be theoretically computed via methods from animal breeding, such as GBLUP or BayesCpi, which provide the posterior variance of the genetic component of the trait in the individuals considered, given the genomic markers. Such approaches would also account for relatedness between the pair of individuals. However, the posterior variances of the genetic component would not directly provide an estimate for prediction accuracy when considering predictions of the direction of phenotypic difference.". We now also cite the Stephens et al 2017 paper as an example of a similar calculation that arises in the field of statistics (L65).

My main concern with this paper is that I'm still not convinced that most of the application ideas discussed here are actually useful in any way. Why would a patient care if their probability of a disease was 0.001% higher vs 0.001% lower than another's? This method would treat those two cases as completely different, but they're really practically the same. In medicine, I think only the actual phenotypic value (probability of a disease) is useful. In medical insurance also, it is the magnitude of the difference between groups, not the probability that one is greater than another that matters. I think the method could be adjusted to give the probability that a trait will be greater than a pre-specified threshold in a straightforward way which might be more useful. The same arguments apply in breeding (if a line has a 95% chance of giving a higher yield, but the magnitude of that effect is 0.001 mg, I won't be very excited about selecting it for crosses).

We agree that adjusting the method to give the probability that a trait exceeds a certain threshold is helpful and illuminating, and that incorporating this adjustment significantly enhances the paper's applicability and practical relevance. Following the reviewer's feedback, we derived the probability $P(A > B + c)$ as a function of the sum of known effects (or PRS difference) between individuals A and B (new Eq. (7)). To illustrate it, we included a visualization in the new Fig. 2c, where the probability is shown for various values of c . This figure highlights how increasing values of c reduce prediction accuracy, demonstrating the trade-offs involved.

Furthermore, we revised the abstract and introduction to emphasize that predicting the direction of phenotypic differences is more helpful in scenarios where even small differences are informative. For example, comparing an unphenotyped individual to one already known to have a high risk can be actionable, as even a slightly higher risk indicates that the unphenotyped individual is also likely at high risk. We clarified this point, as well as the addition of the threshold approach the reviewer suggested with the following additions: "Importantly, the same approach could also be applied to examine if the tested individual has a higher phenotypic value than an individual confirmed to have

the phenotype (e.g., an individual with the disease), or a threshold of interest (e.g., the unphenotyped crop will produce at least 10% more yield than another crop).” (L52-56), “or that the phenotype of two individuals differs by at least a predetermined constant” (L63-64), “predict if an individual’s disease risk exceeds a baseline (e.g., the population average or a set clinical threshold) or that of other individuals with the disease.” (L71-73), “We provide two derivations - one from the viewpoint of random walks and the other from the viewpoint of statistical genetics, which also enabled us to model shared genetic and environmental components in siblings, and to formulate the probability that the difference between two individuals exceeds a certain threshold of interest” (L160-164), “As expected, adding a threshold c for the phenotypic difference in our prediction reduces prediction accuracy (Fig.2c). However, the reduction in prediction accuracy compared to the no-threshold prediction becomes substantial only when the threshold c is very large. For example, for a threshold $c=0.1$ (0.1 standard deviations of the phenotypic values), the reduction in prediction accuracy is 1.5-5.8% for the delta values we simulated, whereas this reduction is 21-53% for the threshold $c=1$ (a full standard deviation)” (L180-186), and “In sum, we found that the known-to-total ratio (κ) captures the factors that affect the probability of correctly predicting which individual has the higher phenotypic value (and that they can be used to predict if the difference is higher than any defined threshold).” (L201-204).

Same with embryo selection. I think this topic in particular is highly controversial, and so without careful consideration of ethical implications, discussing this as a potential application seems unsafe.

We fully agree that embryo screening is controversial and that it raises several ethical concerns. In the revised manuscript, we cite (PMID 38805697), which is a comprehensive review of ethical (and other) concerns, as well as PMID 38244362, 34916614, 38393332, 33106616, 38519038, which cover various ethical issues with polygenic embryo screening. Given that embryo screening is a technology already available to patients, and given that our results have direct relevance for the evaluation of the utility of this technology, we would prefer to keep the paragraph on embryo screening in the manuscript. At the same time, we clearly emphasize that embryo screening is not intended (or endorsed) as an application of our method, but rather that our method could be useful for the evaluation of this technology: “This technology is considered by many as unethical (Polyakov 2022, Treff 2022, Lencz 2022), and thus, we do not describe embryo screening as an application of our method but rather suggest that our method could be used to evaluate this technology.” (L432-434).

From the statistical side, I am glad the authors considered possible sources of correlation between the predicted and unpredictable sources of variation, as those are clearly possible. One case that I didn’t see discussed is where loci are discovered using a model that includes PCs or a kinship matrix to control for effects of population structure. These terms will bias the model towards loci that are more diverged within populations than between populations. I would guess that if loci are discovered using such a model, the calibration for comparisons between individuals in distinct populations

will be different than for individuals within populations.

We agree that controlling for population structure is very important in loci discovery. In our analysis, we used the first 10 PCs as covariates in the discovery process for this purpose (L576). We agree that if the comparison of individuals from distinct populations was to be done, special consideration would need to be made for population structure correction during the discovery stage. One possibility is to repeat the analysis, once using genotype-phenotype associations discovered in one population, and once with genotype-phenotype associations discovered in the other population, and compare how this affects the results. Another is to correct for population structure in a joint dataset that includes both populations using linear mixed models or other methods. However, we are wary of presenting between-population comparisons in our analysis of the UK Biobank, because of issues of bias towards certain variants highlighted by the reviewer, which still need to be worked out, but also because we want to be careful with the potential misapplication of such comparisons to make unsubstantiated claims about phenotypic differences between human groups. We therefore think this topic should be dealt with in more detail and with the proper care in future investigations of this approach (which we are planning to do).

A minor note: the derivation of $k = \Delta/(\Delta + \sigma)$ seems unnecessarily complex. Since only the ratio $k/(1-k)$ is used, the method would be the same if they simply used Δ/σ , ie the predicted effect measured in units of the residual standard deviation. Phrased as such, the probability is simply the cdf of a standard normal evaluated at Δ/σ .

We agree with the reviewer, this is a nice addition. We now describe this in L487-489.

I found the section on $\bar{r}^2$ hard to follow. Why not just use the s^2_e estimate from the model used to fit the genetic model (ie make the PGS)?

We agree that the section explaining $\bar{r}^2$ was a bit unclear. Following this comment and a comment from reviewer 2 we have revised our explanation of $\bar{r}^2$ in L541-548. The reviewer is correct that one way to estimate $\bar{r}^2$ is through evaluation of the variance explained by the PGS (we present these results in Tables S1). In general, there are different ways in which r^2 can be estimated, and in the paper we explore two: (i) estimating r^2 as the proportion of variance explained by the PGS, and (ii) using the theoretical expectation we developed here. Method (ii) capitalizes on the mathematical formulations we developed in this paper and is used in the main text, and we compare the r^2 estimates of method (ii) with those attained using method (i), which is more conventional, in Table S1, showing that they are comparable.

I found the descriptions of the non-human analyses incomplete. It looks like QTL mapping results were used, ie biparental populations created by crossing individuals from each population. Were the tests evaluated on pairs of individuals from these biparental populations, or individuals from the original (non-admixed) populations?

We agree with the reviewer that the description in this section was incomplete. As the reviewer correctly understood, QTL mapping results were used from biparental populations created by crossing individuals from each population. Then, these data were used to evaluate the original (non-admixed) populations. In the modern human vs Neanderthal, and human vs chimp comparisons, we show our previously published results from Gokhman et al., *Cell*, 2019. In these analyses, the association between genetics and phenotypes is conducted differently than in the stickleback, daisy and mouse analyses. In Gokhman et al, 2019, the direction of phenotypic change was predicted based on DNA methylation changes separating the lineages, rather than sequence variants. We have now expanded the text describing these results and methods. L350 - “These predictions were based on DNA methylation changes separating the lineages”. L329-332 - “We tested three datasets that mapped quantitative trait loci (QTL) separating pairs of populations of sticklebacks (Rogers et al., 2012), daisies (Brennan et al., 2016), and mice (Gray et al., 2015). Then, we tested to what extent these QTLs predict the true direction of phenotypic change between population pairs.”. L623-626 - “In all of these analyses, we took genotype-to-phenotype data as reported in the original studies. The first three comparisons were based on QTLs, which were detected by analyzing admixed populations. Then, these QTLs were used to predict the direction of phenotypic change between the original (non-admixed) populations.”. Finally, we added the following text to the figure legend: “For the modern human vs Neanderthal, and human vs chimp comparisons, we show results from Gokhman et al., 2019”.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer for the constructive suggestions, and we hope the reviewer finds the new version satisfactory.

Reviewer #1:

I thank the authors for their responses to my first set of comments.

COMMENT 1: I appreciate that you have made an attempt to explain what you mean by "known" effects (L101-102), but it's still not satisfactory. Since there are no rules for what is "known" ("assumed to be associated with the phenotype" adds nothing to a definition, it does help with intuition). The "knowns" can be any arbitrary set of genetic variants but for the method to work well they should be a good predictors of phenotype. I'd suggest something like "We distinguish the contribution to the phenotype of a chosen set of genotyped variants that we will refer to as 'known', from all other genetic and environmental effects". Or maybe since in practice the knowns are a PGS maybe just call it that from the start as

RESPONSE 1: We agree with the reviewer that the current wording was unclear. We rephrased the definition by merging the two suggestions made by the reviewer to: "We distinguish between two groups of contributions to the phenotype. We refer to the first group as known effects, representing a chosen set of genotyped variants predictive of the phenotype (e.g., variants contributing to a PGS). The second group is the unknown effects, which include loci or environmental factors not yet discovered." (lines 100-104).

COMMENT 2: It seems to me essential to have a discussion here (with perhaps more later) of what makes a good choice for the "known" variants. You later consider "... known effects are those with the largest effects" but wouldn't it be better (and more common) to use the variants that pass a p-value threshold? Further, with modern PRS methods that simultaneously estimate effect sizes for all SNPs using shrinkage why wouldn't you include all genotyped SNPs as known? It's unsatisfactory that all knowns are treated the same irrespective of the uncertainty in effect size estimates, but maybe shrinkage estimates can deal adequately with this: the less information (due to low MAF) the greater the shrinkage. PRSice-2 may still be widely used but it is not

methodologically state-of-art: advanced PRS like DBSLMM or MegaPRS can do much better

RESPONSE 2: Following this comment, we realized that our wording was not sufficiently clear. The comment refers to both our simulation analysis and the analysis of the UK biobank data. In our simulations, we wanted to understand the effect of ascertainment bias for large effect (in PGS but also, perhaps even more so, in between-species studies). Therefore, in one of the analyses, we simulated the effect of this bias by treating the largest effect sizes as those that are discovered, and are therefore “known”. In the simulations, we did not model any uncertainty of the measurements (i.e., the simulated effect sizes are treated as known), and therefore there is no need for p-value thresholds. In the analysis of the UK biobank data, we used standard procedures for selecting a set of variants to be included in the PGS, including p-value thresholds. Indeed, as mentioned in lines 571-576, our aim is not to optimize these choices, but rather to keep them as standard as possible. In this case, we also filter based on p-values, as described by the reviewer as the appropriate approach. We acknowledge that our analysis of the UK biobank data does not use the strongest available methods (e.g., DBSLMM or MegaPRS) and parameterizations, but we chose this approach to demonstrate that our results do not depend on methodological choices, but are rather inherent to the general approach of making predictions of phenotypic direction. We now clarify this point in lines 570-571.

COMMENT 3: You now clarify (L114) that "the same approach could be applied to test if the difference between the phenotypes exceeds a predetermined threshold of interest" and you also say that it can be applied to compare an individual with a fixed value, without reference to a comparison individual. But doing that for every possible threshold corresponds to specifying a probability distribution for the unknown phenotype, which I previously said would be much more useful than just comparison with a reference individual (another reviewer said something similar). It seems worth highlighting this. In practice it's not feasible to compute probabilities for every threshold, but you can approximate the probability distribution function to arbitrary accuracy by increasing the grid of thresholds considered.

RESPONSE 3: We agree. In fact, we are currently developing this approach in a follow-up study. We now highlight this direction in the Discussion (lines 409-411), as suggested.

COMMENT 4: L158 "probability describing prediction accuracy" this is too vague, there are multiple probabilities that can measure prediction accuracy, better to be explicit.

RESPONSE 4: We apologize for the confusion. We rephrased this sentence to more precisely define the probability described in Eq. 2 as “the probability that a prediction for the direction of the phenotypic difference is indeed correct (i.e., the prediction accuracy)” (lines 159-160)

Methods:

COMMENT 5: why is this section in a tiny font?

RESPONSE 5: We have now increased the font size for the method section (but we believe the final typesetting will be addressed by the journal).

COMMENT 6: Why bother defining the E as binary and then invoke the CLT? Better to just assume Gaussian error as is standard.

RESPONSE 6: We find that this approach has several advantages. First, we believe that when presenting the argument in terms of discrete effect sizes, it is easier to relate the theory to the evolutionary perspective we present in the main text (e.g., Fig. 1). Second, this discrete formulation is the one we used in our simulations. While for the neutral case this could have been done using normal distributions, for the simulations involving selection this would not have been straightforward. Lastly, formulating the model starting from discrete effect sizes rather than normal distributions can be more flexible for extensions, for example to model different effect size distributions and selection regimes.

Minor comments:

COMMENT 7: L62 "estimate the likelihood" in a statistics-heavy paper you should only use "likelihood" in its technical sense. Here, you mean "probability". A bit more subtle, but "estimate" is not really correct here either - there is no true parameter that you are trying to estimate. You compute or approximate a probability under given assumptions.

RESPONSE 7: We agree, thank you for the attention to the details of the phrasing here. We changed the phrasing to "evaluate the probability" (line 62).

COMMENT 8: L487 typo "an used"

RESPONSE 8: Thank you, fixed.

Reviewer #2

I thank the authors for the extensive and careful work they put into revising this manuscript and addressing all the previous comments: their efforts are greatly appreciated. I only have a couple of minor details to bring up:

COMMENT 9: * Although I believe this to be a relatively minor detail, I do think it would be helpful to have some sort of uncertainty measurement for the $\bar{r^2}$ values in

table S1 and S2. I raise this issue because while it seems likely that, for example, the r^2 and $\bar{r}^2$ values agree for EUR BMI, it is not clear if the difference between r^2 and $\bar{r}^2$ for standing height is meaningful (vs a statistical fluctuation). Similarly for hip circumference (EUR), metabolic rate (siblings), and BMI (siblings). I think this can be quickly done by a simple bootstrap (e.g. reselecting your inference and validation sets) or even by sub-sampling the groups you already have.

RESPONSE 9: We agree. We have now added confidence interval information, obtained by bootstrapping, to Tables S1, as suggested.

COMMENT 10: Missing period at the end of line 186.

RESPONSE 10: We did not notice a missing period at the end of the sentence in line 186, but we made sure there are no missing periods.

COMMENT 11: Very minor suggestion that doesn't need to be addressed for this manuscript to proceed to publication: In Methods "Alternate Derivation" you use e_i as the residual (phenotype minus PGS) which are used to determine the scale σ . However, in the following section "Simulation" you used e_i for both genetic and non-genetic contributions (used to determine both Δ and D), and in "Analysis of pairwise comparisons in humans" you use e_i again to mean something slightly different (additive effect of single alleles to the phenotype) which is used to determine Δ . All the while, much of the other notation the same. This is all clear in context, but can get confusing when jumping around.

RESPONSE 11: We agree, thank you for noticing this abuse of notations. To resolve the dual use of e for effect sizes and residuals, we have now replaced e in the "Alternate Derivation" section with η . In the "simulations" section and "Analysis of pairwise comparisons in humans" we use e_i in slightly different meanings that relate to the initial formal definition of e_i in the "Formal model for prediction accuracy" section. In this case, the meaning and usage of e_i is similar, and it is used to determine Δ in a similar way. Therefore, here we think that it is better to leave the notation as-is, rather than add more notation that could confuse the reader.

Reviewer #3

The authors have addressed my technical concerns.

COMMENT 12: I remain largely unconvinced that predicting the direction of change as described here is useful. Even with the +c threshold, a value 0.01% above this new threshold would still be treated completely different than a value 0.01% below that new threshold, so it's just pushing the goalposts, but not addressing the criticism that thresholds themselves like this are not that useful for individual-level decision-making. Only effect size estimates are.

RESPONSE 12: We agree with the reviewer that an individual with trait value just above a threshold should not necessarily be treated differently than an individual just below this

threshold. We also agree that accurate effect sizes are the most useful information, but, unfortunately, accurate effect sizes are often unavailable. Instead, we suggest, as Reviewer 1 and 2 suggested too, to pre-determine a threshold of interest (e.g., having a disease risk at least 2 times higher than the population average) and filter for individuals passing this threshold. This can also be done to filter for individuals substantially below a threshold, e.g., creating one group of individuals with a disease risk at least 2 times higher than the population average, and another group with a disease risk that is at least 5 times lower than the population average (i.e., these two groups are not just above and just below the threshold, but represent two groups with sufficiently large difference in their phenotypes, and with high accuracy of the partition). In some applications, like prioritizing treatment of patients or selecting crops, we believe this can be a useful approach when more precise information on effect sizes is unavailable or less reliable. Generally, medical practice often involves hard thresholds for interventions, including prescribing medications, initiating treatments, and funding. Having a better understanding of probabilities that a certain phenotype will cross a threshold is thus desired, despite the limitations as stated by the reviewer.

COMMENT 13: Also, I remain concerned about the discussion about embryo selection. The authors added a caveat saying this approach isn't recommended, but the rest of the paper makes it sound like it is likely to be very useful to "sufficient to substantially reduce the chances of exceeding the disease threshold.". Given all the uncertainty in GxE, additivity, etc discussed above, is this really a statement that the authors or the journal want to make in a paper when it can have such major ethical implications if mis-used?

RESPONSE 13: Based on this comment and the request of the editor, we have now removed this paragraph from the discussion.

Reviewer #4

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.