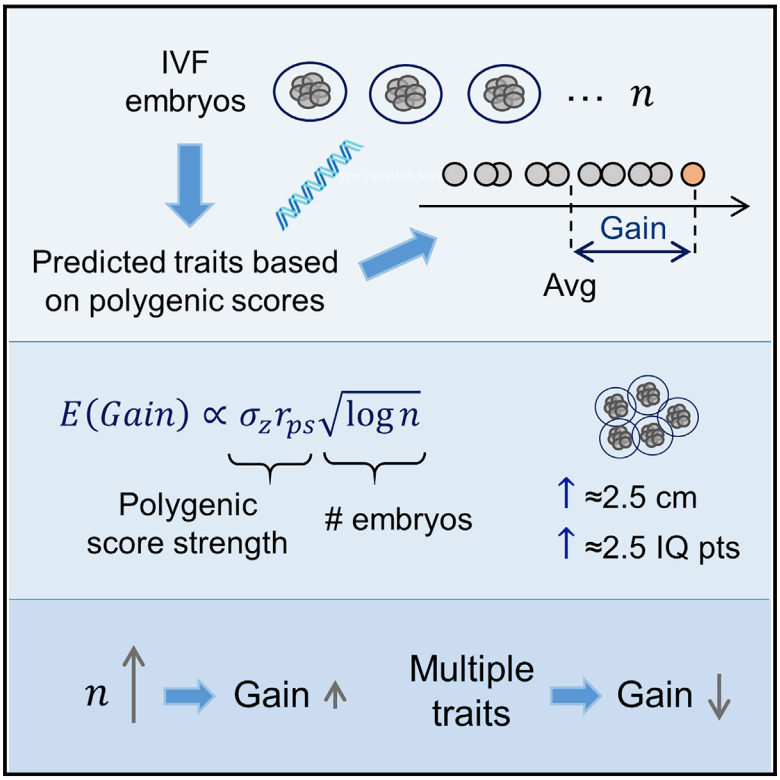
Cell

Screening Human Embryos for Polygenic Traits Has Limited Utility

Graphical Abstract



Highlights

- IVF embryos could be profiled with polygenic scores for traits such as height or IQ
- The top-scoring embryo is expected to be ≈2.5 cm or ≈2.5 IQ points above the average
- The adult trait value of the top-scoring embryo would remain widely distributed
- Multiple ethical and other factors impose practical limits on the actual gain

Authors

Ehud Karavani, Or Zuk, Danny Zeevi, ..., Max Lam, Todd Lencz, Shai Carmi

Correspondence

tlencz@northwell.edu (T.L.), shai.carmi@huji.ac.il (S.C.)

In Brief

Recent progress in genetic testing of embryos has made it technically feasible to profile IVF embryos for polygenic traits such as height or IQ, but simulations, models, and empirical data show that the gain in trait value when selecting the top-scoring embryo is currently limited and uncertain.

Karavani et al., 2019, Cell *179*, 1–12
November 27, 2019 © 2019 Elsevier Inc.
https://doi.org/10.1016/j.cell.2019.10.033

CellPress

Please cite this article in press as: Karavani et al., Screening Human Embryos for Polygenic Traits Has Limited Utility, Cell (2019), https://doi.org/10.1016/j.cell.2019.10.033

Cell

Screening Human Embryos for Polygenic Traits Has Limited Utility

Ehud Karavani,[1,18] Or Zuk,[2,18] Danny Zeevi,[3] Nir Barzilai,[4,5] Nikos C. Stefanis,[6,7,8] Alex Hatzimanolis,[6,8] Nikolaos Smyrnis,[6,7] Dimitrios Avramopoulos,[9,10] Leonid Kruglyak,[3,11,12] Gil Atzmon,[4,5,13] Max Lam,[14,15,16] Todd Lencz,[14,15,17,*] and Shai Carmi[1,19,*]

[1]Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem 9112102, Israel
[2]Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 9190501, Israel
[3]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA
[4]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[5]Department of Genetics, Institute for Aging Research, Albert Einstein College of Medicine, Bronx, NY 10461, USA
[6]Department of Psychiatry, National and Kapodistrian University of Athens Medical School, Eginition Hospital, 115 28 Athens, Greece
[7]University Mental Health Research Institute, 115 27 Athens, Greece
[8]Neurobiology Research Institute, Theodor-Theohari Cozzika Foundation, 115 21 Athens, Greece
[9]Department of Psychiatry, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA
[10]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[11]Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA
[12]Howard Hughes Medical Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA
[13]Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa 3498838, Israel
[14]Division of Psychiatry Research, Zucker Hillside Hospital, Glen Oaks, NY 11004, USA
[15]Institute of Behavioral Science, Feinstein Institutes of Medical Research, Manhasset, NY 11030, USA
[16]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[17]Department of Psychiatry, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY 11549, USA
[18]These authors contributed equally
[19]Lead Contact
*Correspondence: tlencz@northwell.edu (T.L.), shai.carmi@huji.ac.il (S.C.)
https://doi.org/10.1016/j.cell.2019.10.033

SUMMARY

The increasing proportion of variance in human complex traits explained by polygenic scores, along with progress in preimplantation genetic diagnosis, suggests the possibility of screening embryos for traits such as height or cognitive ability. However, the expected outcomes of embryo screening are unclear, which undermines discussion of associated ethical concerns. Here, we use theory, simulations, and real data to evaluate the potential gain of embryo screening, defined as the difference in trait value between the top-scoring embryo and the average embryo. The gain increases very slowly with the number of embryos but more rapidly with the variance explained by the score. Given current technology, the average gain due to screening would be ≈2.5 cm for height and ≈2.5 IQ points for cognitive ability. These mean values are accompanied by wide prediction intervals, and indeed, in large nuclear families, the majority of children top-scoring for height are not the tallest.

INTRODUCTION

The use of biotechnology to influence the genetic composition of human embryos in the absence of specific disease risk raises many ethical concerns, and the recent live births resulting from human embryonic CRISPR editing have heightened global attention to these issues (Coller, 2019; National Academies of Sciences Engineering and Medicine, 2017). Currently, the most practical approach to genetic "enhancement" of embryos is preimplantation genetic screening of in vitro fertilization (IVF) embryos. Preimplantation genetic diagnosis and screening (Sullivan-Pyke and Dokras, 2018) have been utilized for years to avoid implantation of embryos harboring monogenic disease-causing alleles or aneuploidies. Recently, it also became technically feasible to generate accurate genome-wide genotypes from single-cell input (Kumar et al., 2015; Treff et al. 2019). This development, coupled to recent progress in complex traits genetics, has made it possible to genetically screen embryos for polygenic traits, and has raised the prospect of "designer babies" (The Economist, 2018).

Perhaps the most controversial potential application of polygenic embryo selection would be selection for intelligence, especially given the abhorrent history of the early 20th century eugenics movement (Tabery, 2015). While most ethicists are deeply troubled by such prospects, at least one scholar has suggested that there is an ethical obligation for parents to "select the best children" (Savulescu, 2001). In our view, any discussion of the ethics of embryo selection would ideally be informed by quantification of the expected utility of polygenic selection, either as of today, or as reasonably projected into the future. In this report, we thus utilize statistical and empirical methods to evaluate the potential effects of human embryo selection for polygenic traits.

Please cite this article in press as: Karavani et al., Screening Human Embryos for Polygenic Traits Has Limited Utility, Cell (2019), https://doi.org/10.1016/j.cell.2019.10.033

Polygenic scores (PS) are derived from large-scale genome-wide association studies (GWAS) of complex traits, which can be quantitative (such as intelligence or height) or categorical (such as disease status, in which case they are often referred to as "polygenic risk scores" [Wray et al., 2013]). A PS is the count of effect alleles in an individual's genome, weighted by each allele's strength of association with the trait of interest in an independent GWAS (Purcell et al., 2009). The predictive power of a PS is usually represented by r^2_{ps}, or the proportion of variance of the quantitative trait explained by the PS. GWASs of intelligence (Davies et al., 2018; Savage et al., 2018) have demonstrated a relatively modest out-of-sample r^2_{ps} ($\approx$5%), despite large sample sizes ($n \approx$ 300,000 individuals). By contrast, recent large-scale GWASs of height have attained r^2_{ps} of ~25%, demonstrating a highly polygenic genetic architecture similar to intelligence (Yengo et al., 2018a). Consequently, in the present report, we analyze height in addition to cognitive ability, which also allows us to exploit several datasets in which height data, but not intelligence data, are available.

PSs are typically evaluated on a cohort basis and are not used to differentiate one individual from another (although a recent report has demonstrated that, for an extraordinarily tall NBA player, the PS for height was >4 SDs above the population mean [Sexton et al., 2018]). In order for polygenic embryo selection to hold potential utility (independent of ethical considerations), PSs must provide sufficient predictive power to differentiate between embryos within the restricted range of genetic variance available in a single family, and with a finite number of embryos. Two reports utilizing only mathematical modeling have suggested that substantial effect sizes for embryonic selection are possible (Branwen, 2016; Shulman and Bostrom, 2014). However, to our knowledge, despite the widespread application of polygenic scores to complex traits and precision medicine in the research literature (Torkamani et al., 2018), no published studies have empirically examined the possibilities and limitations of a polygenic approach to embryo selection.

We consider here embryo selection in the context of a hypothetical IVF cycle. Our quantity of interest is the difference between the predicted value of the selected trait (i.e., height or intelligence) when the embryo with the highest PS is selected, compared with the mean across embryos. We term this difference the "gain," and we further differentiate between the "predicted" gain, as determined by the PS, and the "realized" gain, as observed in the fully grown offspring. Because no study can be performed in actual embryos, we utilize three sources of data: (1) a quantitative genetic model, (2) simulated embryo genomes generated using realistic parameters from existing genotyped adults with known phenotypic values, and (3) a unique pedigree dataset of nuclear families with large numbers of offspring (10 on average), now fully grown adults, with available genotype and phenotype data. In our simulated data, we examine the gain as a function of varying predictive strengths (r^2_{ps}) of the PS, as well as of the number of embryos (n) available; these results are compared against a theoretical model derived for average gain. Although a typical IVF cycle may produce 3–8 viable embryos (median = 5) (Sunkara et al., 2011), we examine the gain across a broad range of values of n, given the possibility of future advances in IVF technology. Particular emphasis is placed on n = 10, representing a plausible upper bound within the foreseeable future.

RESULTS

We first developed a simple quantitative genetic model for the expected gain. The model assumes a polygenic additive trait with no assortative mating, and hence no correlation between the scores of SNPs from homologous chromosomes or chromosomes of spouses. We recognize that statistically significant assortative mating has been demonstrated for genetic variants associated with polygenic traits such as height and educational attainment (Conley et al., 2016); however, the overall magnitude of this effect accounts for <5% of the variance in spousal phenotype (Robinson et al., 2017; Tenesa et al., 2016). Assortative mating would tend to reduce the efficacy of embryo selection due to reduced variance available from which to select and lower within-family score accuracy (Mostafavi et al., 2019), and thus our results described below represent an upper bound on the potential gain.

We assumed a couple has generated n embryos, and computed the distribution of the polygenic scores of these embryos for a trait with phenotypic variance σ^2_z, of which a proportion r^2_{ps} is explained by the PS. The set of n polygenic scores can be modeled as having a multivariate normal distribution with zero means, all variances equal to $\sigma^2_z r^2_{ps}$, and all covariances equal to $(1/2)\sigma^2_z r^2_{ps}$. The predicted gain is formally defined as the difference between the maximal and average PSs among the n embryos. Based on properties of multivariate normal distributions, the mean gain can be shown to be approximately (for details see Methods S1, sections 1–3)

$$E[gain] \propto \sigma_z r_{ps} \sqrt{\log n}, \qquad \text{(Equation 1)}$$

where the coefficient of proportion is $\approx$0.77. A more accurate formula based on extreme value theory can also be derived (Methods S1; Equation 33). Most notably for our purposes, the mean gain increases with the square root of the variance explained (or linearly with the correlation coefficient between the PS and the trait), but the effect of n is considerably attenuated, as denoted by the square root and log transformation in Equation 1.

Next, for our simulations, we used genotypic and phenotypic data from two cohorts. The Longevity cohort contained 102 couples of Ashkenazi Jewish origin with genome-wide genotypes and information on height, drawn from a larger longevity study (Atzmon et al., 2009). The ASPIS cohort (Stefanis et al., 2004) contained 919 young Greek males with genome-wide genotypes and information on general cognitive function. To simulate embryos, we used either actual couples (for the Longevity cohort) or randomly matched couples (for both cohorts), and generated n = 10 or 50 synthetic offspring per couple based on a standard model of recombination (see STAR Methods for details).

To predict the height or IQ of each embryo, we used polygenic scores based on summary statistics derived from recent large-scale GWAS meta-analyses. For height, the most recent meta-analysis contained $\approx$700,000 individuals (Yengo et al., 2018a)

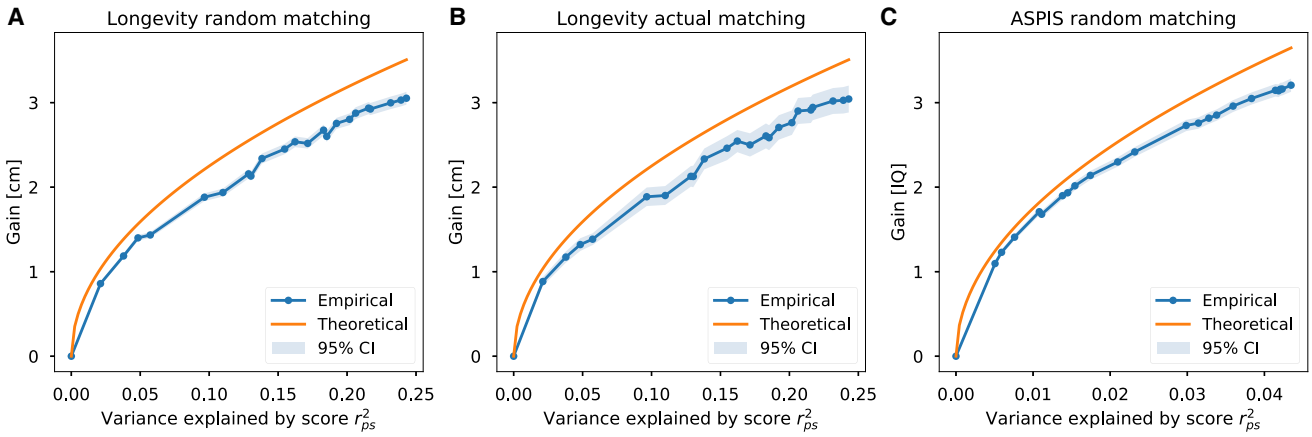


Figure 1. The Mean Gain versus the Proportion of the Variance Explained by the PS

Blue dots and the 95% confidence intervals (light blue bands) represent simulations with 10 embryos per couple. To generate scores with increasing proportions of variance explained, we gradually added chromosomes 1 to 22 to the computed PS. The orange line corresponds to the theoretical model derived in Methods S1 and described in Equation 1. For each value of r^2_{ps}, dots are averages and 95% confidence intervals are based on ±1.96 the SEM over the simulated families.

(A) Gain in height for random couples: 500 simulated pairings drawn from the Longevity cohort.

(B) Gain in height for actual couples: 102 couples from the Longevity cohort.

(C) Gain in IQ for random couples: 500 simulated pairings drawn from the ASPIS cohort.

See also Figure S3.

and did not include the subjects in our test (Longevity) cohort. For IQ, we utilized the most recent published meta-analysis (Savage et al., 2018), from which the COGENT set of cohorts (including the ASPIS cohort) had been removed, resulting in a discovery sample size of $n = 234,569$. We optimized the polygenic scores with respect to imputation, LD-pruning, and the p value threshold (STAR Methods). Our scores predicted height in the Longevity cohort with $r^2_{ps} = 24.3\%$ and IQ in the ASPIS cohort with $r^2_{ps} = 4.3\%$, both within one percentage point of the out-of-sample predictive power reported in the original GWASs. Using linear regression of the phenotype (age- and sex-corrected for height) on the polygenic scores in each cohort, we predicted the height or IQ of each simulated embryo.

Having calculated the predicted height of each simulated embryo from the Longevity cohort and the predicted IQ of each simulated embryo from the ASPIS cohort, we sought to test the predictions of the mathematical model in Equation 1. To examine the relationship between predicted gain and the variance accounted for by the PS, we fixed the number of embryos to $n = 10$ and plotted the mean gain for height against increasing r^2_{ps}. Because polygenic contributions to most complex traits (including height and IQ) are evenly distributed throughout the genome (Shi et al., 2016), we generated PSs that were progressively stronger using PSs derived from growing subsets of the 22 autosomes (e.g., chromosome 1 SNPs only, chromosome 1 + chromosome 2 SNPs only, etc.). As shown in Figure 1, the average gain reaches ≈3 cm or ≈3 IQ points when the full genome-wide PS is used (corresponding to ≈0.5 and ≈0.2 SDs of the trait, respectively). The average gains obtained from varying r^2_{ps} are close to the values predicted by the theoretical model (Equation 1). Our results did not differ when the actual couples were used as the source of the simulated embryos (Figure 1B), compared to couples randomly matched from the Longevity cohort (Figure 1A), indicating that effects of any assortative mating in this dataset are *de minimis*.

The PSs used so far are based on current GWAS results and on a simple LD-pruning and p value-thresholding strategy. However, GWASs are expected to increase in size (in particular given the rapid growth of the direct to consumer genetic industry [Khan and Mittelman, 2018]), and statistical prediction methods are constantly improving (Chung et al., 2019; Lello et al., 2018; Mak et al., 2017; Vilhjálmsson et al., 2015). Given that the theoretically predicted relationship of the gain with r_{ps} was supported by the data in Figure 1, we can forecast the prospects of embryo selection as predictors become increasingly accurate. For example, doubling the proportion of explained variance of height from ≈25% to 50% is expected to increase the mean gain from ≈3 to ≈4.24 cm, with a maximum possible gain of ≈5.5 cm for $r^2_{ps} \approx 80\%$ (the upper bound of the heritability of the trait, as derived from twin studies [Jelenkovic et al., 2016]). Similarly, quadrupling the variance explained for IQ would lead to a doubling of the gain, to ≈6 IQ points (given $n = 10$ embryos).

Next, we tested the relationship between the gain and the number of embryos, holding r^2_{ps} constant. In Figure 2, we show the expected gain versus the number of embryos, for up to 50 embryos. Comparison to the theoretical model again shows good agreement, with an even better fit demonstrated in Figure S1 based on a more accurate approximation (Methods S1; Equation 33). Two implications are immediately apparent from Figure 2. First, current reproductive technologies are in the most sensitive area of the curve. With a typical IVF cycle yielding 5 testable, viable embryos (Sunkara et al., 2011), the predicted gain is reduced from ≈3 to ≈2.5 (cm or IQ points); below 5 embryos, the gain drops precipitously. Second, there is a rather

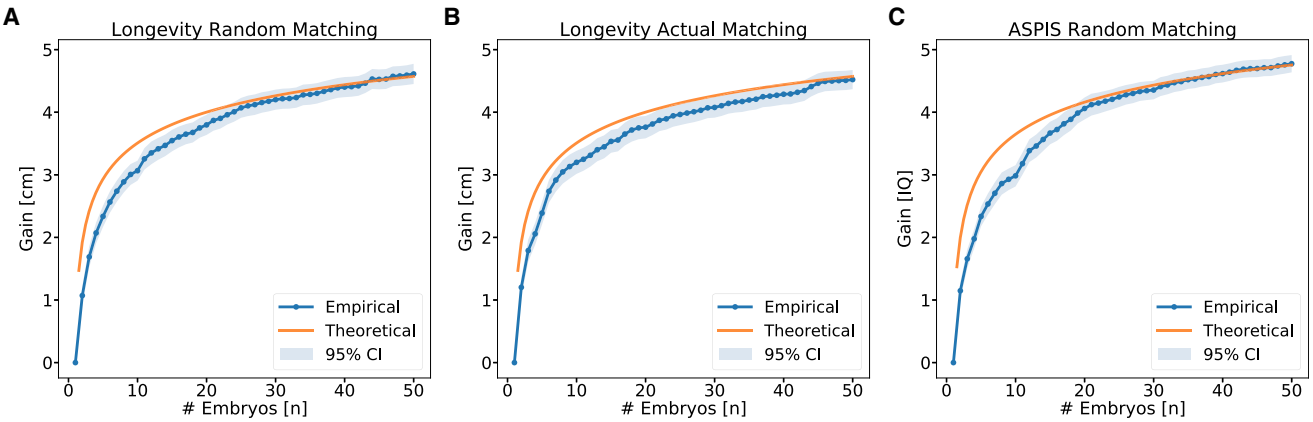


Figure 2. The Mean Gain versus the Number of Embryos

Blue dots are from simulations, and orange lines are for the theoretical prediction (Equation 1). All details are as in Figure 1. See also Figures S1 and S2.

slow increase of the mean gain as the number of embryos increases beyond 10. Thus, even with 1,000 embryos, the mean gain would be only ≈1.7 times higher compared to selection with 10 embryos. Again, no differences were observed between randomly paired and actually married couples (Figures 2A and 2B). The pattern for intelligence was roughly equivalent to that observed for height (Figure 2C).

Both of the results above demonstrate the average gain expected under varying levels of r_{ps}^2 and n across 102 real couples or 500 simulated couples. However, for any given couple, the predicted gain will further vary around this mean. The distribution of the gain, when choosing out of 10 embryos, is shown in Figure 3 for height (for both random and actual couples) and IQ. The gain in height is typically between 1–6 cm, with a median of 2.85 cm for random couples (SD: 1.03; IQR (interquartile range): 2.34–3.80) and 3.02 cm (SD: 0.98; IQR: 2.30–3.60) for actual couples. The gain in IQ was between ≈1–7 points (SD: 1.06; IQR: 2.43–3.84), with a median of 3.02 IQ points. Thus, the predicted gain for a given couple may be somewhat higher or lower than suggested by the mean results of our simulations, due to variation across couples and the random assortment of SNPs in the offspring (see Methods S1, section 4 for a derivation of the variance of the gain). The mean gain itself is affected by the genotypes of the parents, but not by their total scores (Methods S1, section 5).

Figure 3 demonstrates the variability of the predicted gain across couples, but environmental variance leads to additional and substantial variability in the realized gain, as observed in the phenotype of the offspring. Naively (Methods S1, section 6), given PS_{max}, the score of the top-scoring embryo, the 95% prediction interval for the (zero-centered) trait value is

$$\left[PS_{max} - 1.96\sigma_z\sqrt{1-r_{ps}^2}\,, PS_{max} + 1.96\sigma_z\sqrt{1-r_{ps}^2}\right]. \quad \text{(Equation 2)}$$

Equation 2 can be compared to a 95% prediction interval of $[-1.96\sigma_z, 1.96\sigma_z]$ without selection. However, prediction intervals can be narrowed based on the parental phenotypic values, which are usually known. For example, it has been long known that mid-parental height can explain ≈40% of the variance in height of the offspring (Aulchenko et al., 2009) or theoretically $h^4/2 \approx 32\%$ (Visscher et al., 2010). These ≈32% of the variance overlap with the ≈25% explained by the PS, and the combination of both sources of information can never explain more than the heritability. As shown in Figure 4A, even under the extreme scenario where the combination of the PS and the parental values explain the entire heritability of height (≈80%), there would still be a ±5-cm interval around any predicted gain due to environmental and stochastic factors. Based on either the current PS alone, or based on the parents alone, the interval would be as large as ±9–10 cm. For IQ, the 95% prediction interval would be ±13–19 points in case the entire heritability is explained (assuming $h^2 \in [0.6, 0.8]$), or ±24–27 points based on the parents (Figure 4B). Thus, the unexplained variance yields a wide confidence interval around any predicted value for an offspring's trait and therefore a considerable uncertainty in the realized gain that any given couple can expect from embryo selection. This would need to be combined with the variability in the predicted gain itself, as depicted in Figure 3, thereby substantially attenuating any guarantees on the potential benefit.

To give another example, assume there is no variability in the gain, the entire heritability is explained by the combination of the score and the parental phenotypes, and the proportion of variance explained by the PS is 40% for height and 15% for IQ. Selecting out of 10 embryos, a 95% prediction interval for the height of a male child (assuming average parents, 176 cm for the population average, and an SD of 6 cm) would be ∼180 ± 5 cm (i.e., 175–185 cm). This is compared to 176 ± 10 cm (166–186 cm) without selection (Methods S1, section 6). For IQ (mean 100 and SD 15, assuming $h^2 = 0.6$), the 95% prediction interval would be ∼106 ± 19 (88–125), compared to 100 ± 27 (73–127) without selection. The future child has a non-negligible probability (≈0.25, assuming a normal distribution) to have an IQ below the population average.

To evaluate the utility of embryo selection in a real-world setting, we examined a unique cohort of 28 large families

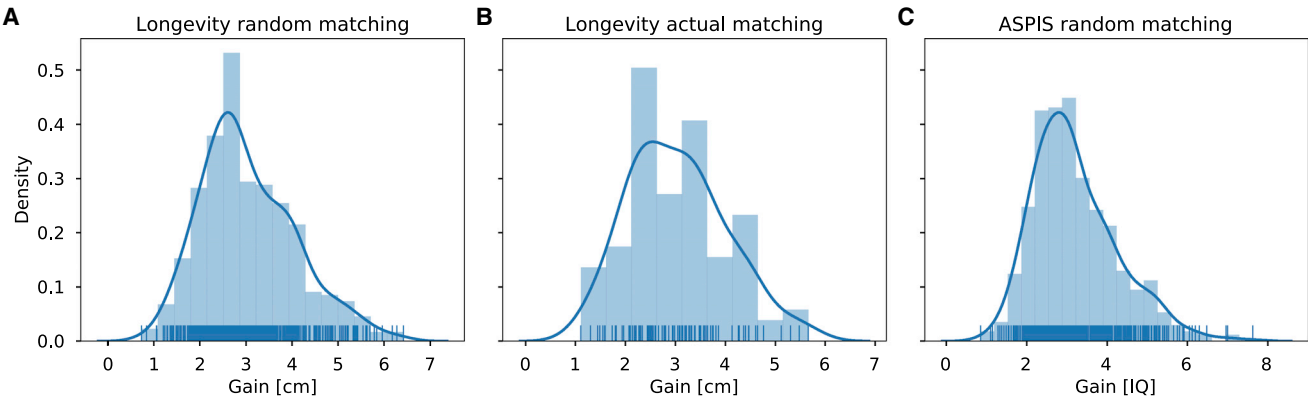


Figure 3. The Distribution of the Predicted Gain from Embryo Selection with 10 Embryos per Couple

(A) The gain in height by simulating 500 random couples from the Longevity cohort.

(B) Same as (A), but with actual spouses ($n = 102$).

(C) The gain in IQ by simulating 500 random couples from the ASPIS cohort. Lines are estimated densities.

with up to 20 offspring each (range: 3–20; mean = 9.6), now grown to adulthood and phenotyped for height. While all these families were the result of traditional means of procreation, we treated the offspring data as if all offspring were simultaneously generated embryos available for selection based on their PSs. Figure 5A depicts the actual difference in height between the offspring with the highest PS, compared to the average height of all the offspring in each family, i.e., the realized gain. (All heights were corrected for age and sex.) While the observed values average around the mean gain predicted by the theory, there was substantial variability in the realized gain. Some families realized a gain of up to 10 cm, while for 5 of the 28 families, choosing the embryo with the highest PS would have resulted in an offspring with height below the average (i.e., gain <0).

The inherent uncertainty in PS-based selection is also demonstrated in Figure 5B, which displays the actual height for each family member. It is notable that the offspring with the highest PS (red squares) is the tallest actual offspring in only 7 of the 28 families. Moreover, when repeatedly downsampled to $n = 7$ children, the offspring with the highest PS was the tallest in $\approx$31.5% of the families, close to the theoretical prediction ($\approx$33.4%; Methods S1, section 7). Across all families, the tallest child was on average $\approx$3.0 cm taller than the child with the tallest predicted height, again very close to the theoretical prediction (3.1 cm; Methods S1, section 7).

Finally, embryo selection could be desired or attempted on the basis of scores for multiple traits, some of which may be positively or negatively correlated. We extended our quantitative model to predict the outcome of this selection scheme (Methods S1, section 8). Specifically, we assumed selection for a weighted average of the scores for T traits, with correlation $\rho_{ps,ij}$ between the scores of traits i and j. We defined the weight of trait i as $\lambda_i/\sigma_{ps,i}$, where $\sigma_{ps,i} = \sigma_{z,i} r_{ps,i}$ is the standard deviation of the score of trait i ($\sigma^2_{z,i}$ is the variance of trait i and $r^2_{ps,i}$ is the proportion of variance in trait i explained by the PS). The mean gain in trait i (i.e., the predicted value of trait i of the embryo with the maximal combined score; denoted G_i), is

$$E(G_i) \propto \sigma_{z,i} r_{ps,i} \sqrt{\log n}\ \frac{\lambda_i + \sum_{j=1, j\neq i}^{T} \lambda_j \rho_{ps,ij}}{\sqrt{\sum_{j=1}^{T} \lambda_j^2 + \sum_{j=1}^{T} \sum_{k=1, k\neq j}^{T} \lambda_j \lambda_k \rho_{ps,jk}}}. \quad \text{(Equation 3)}$$

We demonstrate the application of this formula when jointly selecting for height and BMI in the Longevity cohort (Figure S2).

To gain more insight into Equation 3, consider the case when all trait-trait correlations are equal to ρ, and all weights are equal to $\lambda/\sigma_{ps,i}$. This corresponds to giving each trait an equal weight, after accounting for the different variance explained by each score. The mean gain per trait is

$$E(G_i) \propto \sigma_{z,i} r_{ps,i} \sqrt{\log n}\ \frac{\sqrt{1 + (T-1)\rho}}{\sqrt{T}}. \quad \text{(Equation 4)}$$

If $\rho = 1$, i.e., all scores are equal after normalization, the gain per trait is the same as the gain achieved when selecting for a single trait, as expected. When $\rho = 0$ (i.e., when selecting for T independent traits), the mean gain per trait is $1/\sqrt{T}$ smaller compared to selecting for a single trait. When all traits are maximally anti-correlated ($\rho = -1/(T-1)$), the mean gain per trait completely vanishes. Thus, when selecting for multiple traits simultaneously, the gain per trait can be much smaller compared to selection for a single trait, in particular if PSs of traits are anti-correlated.

DISCUSSION

In this paper, we explored the expected gain in trait value due to selection of human embryos for height and IQ. We showed that the average gain, with current predictors and with five viable embryos, is around $\approx$2.5 cm and $\approx$2.5 IQ points. We predicted and confirmed by simulations that the gain will increase proportionally to the square root of the proportion of the variance explained by the predictor, but much more slowly with the number of

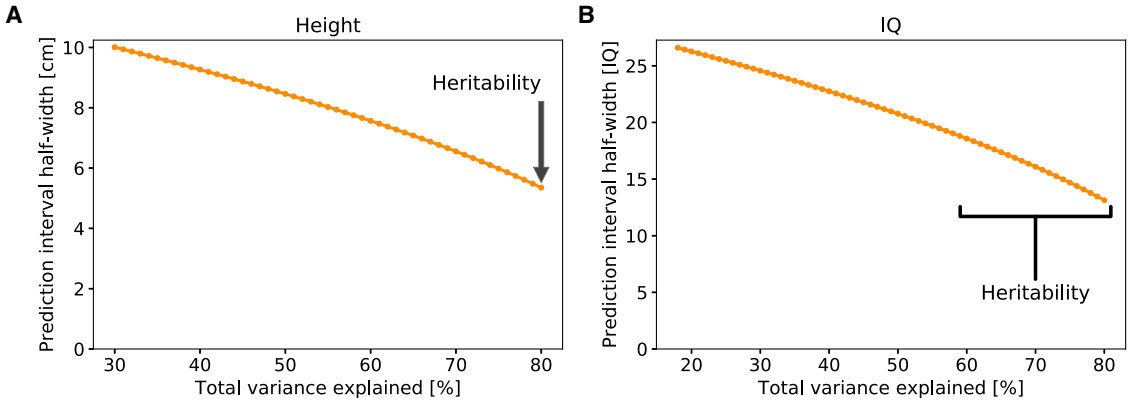


Figure 4. The Prediction Interval Half-Width as a Function of the Proportion of Variance Explained by the Combination of Parental Phenotypes and the PS of the Child

If the proportion of variance explained is p, the half-interval width is $1.96\sigma_z\sqrt{1-p}$.

(A) The prediction interval for height, assuming $\sigma_z = 6$cm. The proportion p is unknown, but cannot exceed the heritability, which we assume to be $h^2 \approx 0.8$, and cannot fall under $h^4/2 \approx 0.32$, which is the theoretical variance explained by the mid-parental height.

(B) The prediction interval for IQ, with $\sigma_z = 15$ points. We assume the heritability is in the range [0.6,0.8], with a minimal variance explained of $0.6^2/2 = 0.18$.

embryos. Only two previous studies have addressed this question to date, both of which employed only mathematical modeling. One study has assumed the entire heritability can be explained by the genetic predictor, leading to larger effect sizes than possible with currently available scores (Shulman and Bostrom, 2014). The second study (a blog post) used a model similar to ours, but focused on futuristic approaches to increasing the number of available embryos (Branwen, 2016).

In animal breeding, genomics-based selection is usually performed not by selecting embryos but by genotyping young males and using top-scoring animals as sires for the next generation. The recent success of genomic selection is mostly attributed to the shortening of the generation time (García-Ruiz et al., 2016), as the genetic value of an animal can already be determined at birth (Meuwissen et al., 2016; van der Werf, 2013). Beyond generation time, genomic selection is expected to be more powerful than embryo selection, because first, the population variance is double the variance between siblings, increasing the gain by a factor of $\sqrt{2}$, and second, the number of individuals to select from (n) is not limited as in IVF cycles. Indeed, we have identified only one study in animal genetics that has suggested and empirically examined embryo selection (Mullaart and Wells, 2018).

Given that r^2_{ps} holds the strongest effect on the potential gain from embryo selection, it is worthwhile to consider the potential for increasing r^2_{ps} in the foreseeable future. Increasing sample sizes of discovery GWASs is the most straightforward means of increasing r^2_{ps} (Chatterjee et al., 2013). For educational attainment, a trait strongly correlated with IQ ($r_g \approx 0.70$) (Hagenaars et al., 2016), increasing GWAS sample size from ≈300 K (Okbay et al., 2016) to ≈1.1 M (Lee et al., 2018) resulted in an increase in out-of-sample variance explained from 3.2% to 11%. For height, the out-of-sample r^2_{ps} increased more modestly, from 17% to 24.4% when GWAS sample size increased from ≈250 k (Wood et al., 2014) to ≈700 k individuals (Yengo et al., 2018a). The variance explained by the predictor should approximately satisfy $r^2_{ps} = h^2_{snp}(1 + M/(Nh^2_{snp}))^{-1}$, where N is the (discovery) GWAS sample size, M is the effective number of markers, and h^2_{snp} is the SNP-based heritability (Pasaniuc and Price, 2017; Wray et al., 2019). The dependence of the gain on N has an empirical S shape (Figure S3). For IQ, increasing GWAS sample sizes to $N \approx 10^7$ is expected to double the gain, up to ≈7 IQ points (for $n = 10$ embryos). For height, we are closer to saturation, and using $N \approx 10^7$ will only increase the gain to ≈4.5 cm. These limitations are to some extent due to the strict upper bound $r^2_{ps} \leq h^2_{snp}$.

Further improvement is expected with the use of whole-genome sequencing (WGS), because it was recently shown that WGS data explain the entire heritability of height and BMI (Wainschtein et al., 2019). For cognitive ability, a recent family-based study (Hill et al., 2018) has demonstrated that more than half of the variation is attributable to rare variation not captured by current GWASs. However, as the effective number of markers in WGS is much larger compared to microarrays and the sample sizes much smaller, the current predictive power is very low (expected gain for height <1 cm; Figure S3). Once sample sizes reach $N = 10^7 - 10^8$, the gain for height can reach ≈5.5 cm, nearly double the current gain (Figure S3). To incorporate rare variation while overcoming the problem of small WGS sample sizes, imputation is a promising approach (Yang et al., 2015), and as reference panels grow in size and diversity, imputation is expected to accurately assess variants with frequencies down to 0.1% or even lower (Lencz et al., 2018; Taliun et al., 2019).

Finally, statistical approaches to calculating PSs from GWASs are becoming increasingly sophisticated (Khera et al., 2018; Privé et al., 2019a; Torkamani et al., 2018). Most notably, the application of penalized regression methods to the generation of PSs holds a potential for rapid gains in r^2_{ps} without requiring any additional data collection in either GWAS datasets or imputation reference panels (Mak et al., 2017; Privé et al., 2019b). For example, initial evidence suggests that currently available

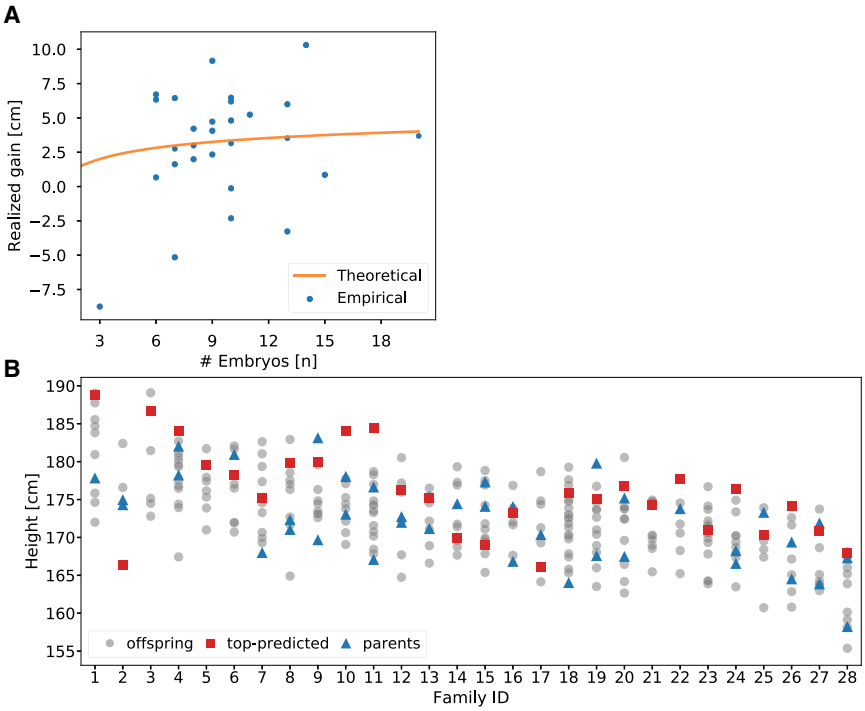


Figure 5. An Analysis of Selection for Height in 28 Real Families with up to 20 Adult Offspring Each

(A) The realized gain in each family, defined as the difference between the actual (age- and sex-corrected) height of the offspring with the highest PS and the average height of all offspring in the family. The theoretical prediction is based on Equation 1.

(B) The actual height (age- and sex-corrected) of all members of all families. The figure demonstrates the effect of the current low-accuracy prediction models, as the tallest-predicted sibling (red squares) is usually not the actual-tallest sibling (only 7/28 times). Siblings are depicted as gray dots and the parents of each family as blue triangles. In some families, only one parent was available.

datasets might be able to explain up to 40% of the variance in height by using LASSO (Lello et al., 2018). Additionally, the use of multiple related phenotypes has been demonstrated to enhance the predictive power of PS (Krapohl et al., 2018); for example, the combination of educational attainment and intelligence GWAS may permit a doubling of cognitive r^2_{ps} (Allegrini et al., 2019). Finally, it has recently been suggested that enrichment of certain subcategories of functional variation (e.g., coding, conserved, regulatory, and LD-related genomic annotations) in GWAS results can be leveraged to further enhance prediction accuracy (Kichaev et al., 2019; Marquez-Luna et al., 2018).

While it is likely that some combination of the above factors will increase the accuracy of PSs in the near future, substantial limitations to PSs must also be acknowledged (Loos and Janssens, 2017). First, PSs do not account for extremely rare Mendelian variants associated with extreme phenotypes such as short stature (Grunauer and Jorge, 2018) or intellectual disability (Vissers et al., 2016). More broadly, the lower end of the phenotypic distribution is less well predicted from common variant PS than the middle and upper percentiles (Chan et al., 2011); this fact limits the utility of PSs for "reverse" embryonic selection (i.e., to avoid extreme low values). Second, it is well known that PSs lose substantial power, or may even be invalid, when applied across different populations (Coram et al., 2017; Kim et al., 2018; Martin et al., 2017). Moreover, even within a single population, subtle ethnic and geographic stratification effects may result in inflated estimates of r^2_{ps} (Barton et al., 2019; Haworth et al., 2019), and prediction accuracy may also vary by age and sex (Mostafavi et al., 2019). Third, polygenic scores are correlated with parental genotypes and hence with the environment induced by the parents, in particular for education. This leads to lower prediction accuracy within families (Cheesman et al., 2019; Domingue and Fletcher, 2019; Morris et al., 2019; Mostafavi et al., 2019; Selzam et al., 2019), which further limits the utility of embryo screening. Fourth, SNP effects may be environmentally sensitive, and may not be consistent across time and place (Keyes et al., 2015).

Beyond these limitations in PS power and accuracy, several additional constraints on the expected utility of embryo selection are notable. First, we did not explicitly model assortative mating, which likely exists to some extent for traits such as height and cognitive ability (Conley et al., 2016; Yengo et al., 2018b), and is expected to further reduce the potential available variance for embryo selection. While there was no detectable effect of assortative mating in our Longevity cohort, these subjects represented an older birth cohort, and assortative mating on phenotypic traits may be increasing. Second, the number of embryos per IVF cycle is usually less than 10 (Sunkara et al., 2011), and, as can be seen in Figure 2, in this regime the utility drops sharply with a decreasing number of embryos. Third, with the increasing age of childbearing, the proportion of aneuploid embryos increases. For example, the proportion of aneuploid embryos is 35% for women aged 35 and 60% at age 40 (Franasiak et al., 2014). Relatedly, embryos with particularly high polygenic scores are not guaranteed to implant and lead to a live birth. While it is theoretically possible to perform multiple IVF cycles to generate more embryos, IVF is invasive, involves a substantial discomfort to the prospective mother, and requires significant financial means (Teoh and Maheshwari, 2014) (that would often also imply an older age of the prospective parents and fewer viable embryos per cycle). To the best of our knowledge, no upcoming technology is expected to significantly increase the number of oocytes extracted per IVF cycle (Casper et al., 2017; Lin et al., 2013). While it has been suggested that induced pluripotent stem cells may greatly increase the potential number of available embryos (Hikabe et al., 2016; Yamashiro et al., 2018), such technologies are not close to implementation for human

reproduction. Either way, even with tens of viable embryos, our simulations show that the gain in trait value would be relatively small (Figure 2). Finally, once IVF and genotyping/sequencing have been performed, couples may attempt to select for multiple phenotypes, and as we have shown, this will lead to smaller gains per each individual trait.

Perhaps more importantly, we have also demonstrated that two sources of variability result in wide confidence intervals for the prediction of final observed phenotypic values: (1) random assortment of SNPs will result in variability of the predicted gain around its mean value; and (2) environmental variation will produce considerable additional uncertainty around the predicted gain. In our empirical dataset, the majority of offspring who were the tallest among their siblings were not those with the highest PS, and a substantial fraction of the top-scoring offspring had lower than average phenotypic values. Regardless of the future accuracy of r^2_{ps} or the number of available embryos, these uncontrollable sources of variability will limit the appeal of selection for any individual couple.

A final reason for caution over the utility of embryo selection is the widespread pleiotropy across most traits (Bulik-Sullivan et al., 2015; Pickrell et al., 2016; Visscher et al., 2017). For example, while IQ is negatively correlated with most psychiatric disorders (Zheng et al., 2017), it is genetically positively correlated with autism and anorexia (Hill et al., 2019; Savage et al., 2018). Therefore, selecting an embryo on the basis of higher predicted IQ will increase the risk for autism or anorexia in the offspring. In animal breeding, selection for production and growth traits has resulted in serious health issues in dairy cattle (Oltenacu and Algers, 2005), broiler chickens (Bessei, 2006), and other animals (Rauw et al., 1998; Rodenburg and Turner, 2012), and in plants, it was recently demonstrated that a flavor allele was lost due to human selection (Gao et al., 2019). Thus, negative effects on correlated health traits should be seriously considered.

In addition to practical limitations, there are major ethical and societal concerns with embryo screening, mostly due to associations with ideas of eugenics. Eugenics was originally developed by Galton, who envisioned breeding of humans for higher intelligence (Tabery, 2015). In short order, Galton's concept was extended in some countries to the forced sterilization of those possessing mental traits deemed as "undesirable" (Hoge and Appelbaum, 2012; Wikler, 1999). The specter of eugenics has accompanied the development of modern reproductive technologies since the development of IVF and preimplantation genetic diagnosis of monogenic diseases (Bonnicksen, 1992). At the same time, application of the term "eugenics" to modern reproductive practices can lead to terminological and conceptual ambiguities that require careful delineation (Cavaliere, 2018). However, even when completely removed from the context of state coercion, embryo selection raises ethical concerns of equity and justice in the availability of expensive reproductive technologies (President's Council on Bioethics (US), 2003), as well as potential conflicts between individual benefits and societal costs (Anomaly et al., 2019). More broadly, embryo selection for non-disease traits raises the possibility of fundamentally altering "the meaning of childbearing" (President's Council on Bioethics (US), 2003).

The legal and regulatory framework for preimplantation genetic diagnosis (PGD) remains unsettled, especially in the United States. While PGD is legal in most countries, its use is often restricted (European Society of Human Reproduction and Embryology, 2017; Jones and Cohen, 2007; Knoppers et al., 2006). Across much of Europe, PGD is legally allowed only when risk for a serious medical condition is high (Dondorp and de Wert, 2019). In this context, high risk generally refers to highly penetrant (dominant or recessive) alleles for clearly defined diseases; thus, polygenic scores for quantitative traits would fail to meet these requirements. In the United Kingdom, the set of permitted conditions is determined by a designated body, which issues explicit guidelines as to which diseases and genes are included (Bayefsky, 2016). In Israel, such decisions are made by institutional review boards, and PGD is not permitted for traits (Israel Ministry of Health, 2013). In China, PGD is regulated and social sex selection and selection for traits are not permitted (Cyranoski, 2017). In contrast, in the United States, the targeted use of PGD is not regulated, and hence, to the best of our knowledge, embryo selection for polygenic traits can be offered to consumers (Bayefsky, 2018, 2016). In such an environment, given the results of this work, the concerns over pleiotropic effects, and the invasive nature of PGD, it may be desired to introduce oversight over at least the advertised outcomes.

Beyond legal restrictions, an additional concern involves the principle of informed consent (Katz, 1994), which suggests that embryo screening should be offered in the context of appropriate genetic counseling. It is our hope that the present work provides an initial evidence base for professionals and regulators to consider the risks and benefits that are at the heart of the informed consent process.

Finally, in this paper, we did not consider the prospects, nor the ethics, of "population-scale" embryo selection for IQ or other traits. While claims were made that population-scale selection could lead to a dramatic increase in trait values at the population level (e.g., the popular article [Hsu, 2016]), we leave a rigorous evaluation of this prediction to future studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cohorts for simulating offspring
- METHOD DETAILS
  - Phasing
  - Polygenic score calculation
  - Simulating embryos
  - Multiple traits
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - The quantitative genetic model
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.10.033.

ACKNOWLEDGMENTS

We thank Yaniv Erlich and Peter Visscher for discussions. S.C. thanks the Abisch-Frenkel Foundation for financial support. We thank the NIH (GM112625) for supporting the CGSI workshop in which one of the collaborations was initiated. T.L. and M.L. were supported, in part, by a grant from the NIH (R01MH117646). The study of the nuclear families was supported by the James S. McDonnell Centennial Fellowship in Human Genetics to L.K.

AUTHOR CONTRIBUTIONS

Conceptualization, T.L. and S.C.; Methodology, O.Z., T.L., and S.C.; Investigation, E.K., O.Z., D.Z., T.L., and S.C.; Resources, N.B., N.C.S., A.H., N.S., D.A., L.K., and G.A.; Formal Analysis, E.K. and S.C.; Data Curation, D.Z. and M.L.; Writing – Original Draft, E.K., O.Z., T.L., and S.C.; Writing – Review & Editing, E.K., O.Z., D.Z., N.B., L.K., G.A., T.L., and S.C.; Visualization, E.K. and S.C.; Funding Acquisition, T.L. and S.C.; Supervision, T.L. and S.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 21, 2019
Revised: September 11, 2019
Accepted: October 25, 2019
Published: November 21, 2019

REFERENCES

Allegrini, A.G., Selzam, S., Rimfeld, K., von Stumm, S., Pingault, J.-B., and Plomin, R. (2019). Genomic prediction of cognitive traits in childhood and adolescence. Mol. Psychiatry *24*, 819–827.

Anomaly, J., Gyngell, C., and Savulescu, J. (2019). Great minds think different: Preserving cognitive diversity in an age of gene editing. Bioethics. https://doi.org/10.1111/bioe.12585.

Atzmon, G., Barzilai, N., Surks, M.I., and Gabriely, I. (2009). Genetic predisposition to elevated serum thyrotropin is associated with exceptional longevity. J. Clin. Endocrinol. Metab. *94*, 4768–4775.

Aulchenko, Y.S., Struchalin, M.V., Belonogova, N.M., Axenovich, T.I., Weedon, M.N., Hofman, A., Uitterlinden, A.G., Kayser, M., Oostra, B.A., van Duijn, C.M., et al. (2009). Predicting human height by Victorian and genomic methods. Eur. J. Hum. Genet. *17*, 1070–1075.

Barton, N., Hermisson, J., and Nordborg, M. (2019). Why structure matters. eLife *8*, e45380.

Bayefsky, M.J. (2016). Comparative preimplantation genetic diagnosis policy in Europe and the USA and its implications for reproductive tourism. Reprod. Biomed. Soc. Online *3*, 41–47.

Bayefsky, M. (2018). Who Should Regulate Preimplantation Genetic Diagnosis in the United States? AMA J. Ethics *20*, E1160–E1167.

Bessei, W. (2006). Welfare of broilers: a review. Worlds Poult. Sci. J. *62*, 455–466.

Bonnicksen, A. (1992). Genetic Diagnosis of Human Embryos. Hastings Cent. Rep. *22*, S5–S11.

Branwen, G. (2016). Embryo selection for intelligence. https://www.gwern.net/Embryo-selection.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Carmi, S., Hui, K.Y., Kochav, E., Liu, X., Xue, J., Grady, F., Guha, S., Upadhyay, K., Ben-Avraham, D., Mukherjee, S., et al. (2014). Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat. Commun. *5*, 4835.

Casper, R., Haas, J., Hsieh, T.-B., Bassil, R., and Mehta, C. (2017). Recent advances in *in vitro* fertilization. F1000Res. *6*, 1616.

Cavaliere, G. (2018). Looking into the shadow: the eugenics argument in debates on reproductive technologies and practices. Monash Bioeth. Rev. *36*, 1–22.

Chan, Y., Holmen, O.L., Dauber, A., Vatten, L., Havulinna, A.S., Skorpen, F., Kvaløy, K., Silander, K., Nguyen, T.T., Willer, C., et al. (2011). Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. PLoS Genet. *7*, e1002439.

Chang, A.L.S., Atzmon, G., Bergman, A., Brugmann, S., Atwood, S.X., Chang, H.Y., and Barzilai, N. (2014). Identification of genes promoting skin youthfulness by genome-wide association study. J. Invest. Dermatol. *134*, 651–657.

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. *45*, 400–405, e1–e3.

Cheesman, R., Hunjan, A., Coleman, J.R.I., Ahmadzadeh, Y., Plomin, R., McAdams, T.A., Eley, T.C., and Breen, G. (2019). Comparison of adopted and non-adopted individuals reveals gene-environment interplay for education in the UK Biobank. bioRxiv. https://doi.org/10.1101/707695.

Choi, Y., Chan, A.P., Kirkness, E., Telenti, A., and Schork, N.J. (2018). Comparison of phasing strategies for whole human genomes. PLoS Genet. *14*, e1007308.

Chung, W., Chen, J., Turman, C., Lindstrom, S., Zhu, Z., Loh, P.-R., Kraft, P., and Liang, L. (2019). Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. Nat. Commun. *10*, 569.

Coller, B.S. (2019). Ethics of Human Genome Editing. Annu. Rev. Med. *70*, 289–305.

Conley, D., Laidley, T., Belsky, D.W., Fletcher, J.M., Boardman, J.D., and Domingue, B.W. (2016). Assortative mating and differential fertility by phenotype and genotype across the 20th century. Proc. Natl. Acad. Sci. USA *113*, 6647–6652.

Coram, M.A., Fang, H., Candille, S.I., Assimes, T.L., and Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. Am. J. Hum. Genet. *101*, 218–226.

Cyranoski, D. (2017). China's embrace of embryo selection raises thorny questions. Nature *548*, 272–274.

Davies, G., Lam, M., Harris, S.E., Trampush, J.W., Luciano, M., Hill, W.D., Hagenaars, S.P., Ritchie, S.J., Marioni, R.E., Fawns-Ritchie, C., et al. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. Nat. Commun. *9*, 2098.

Domingue, B., and Fletcher, J. (2019). Separating Measured Genetic and Environmental Effects: Evidence Linking Parental Genotype and Adopted Child Outcomes. bioRxiv. https://doi.org/10.1101/698464.

Dondorp, W., and de Wert, G. (2019). Refining the ethics of preimplantation genetic diagnosis: A plea for contextualized proportionality. Bioethics *33*, 294–301.

Eny, K.M., Lutgers, H.L., Maynard, J., Klein, B.E.K., Lee, K.E., Atzmon, G., Monnier, V.M., van Vliet-Ostaptchouk, J.V., Graaff, R., van der Harst, P., et al.; LifeLines Cohort Study Group; DCCT/EDIC Research Group (2014).

GWAS identifies an NAT2 acetylator status tag single nucleotide polymorphism to be a major locus for skin fluorescence. Diabetologia *57*, 1623–1634.

European Society of Human Reproduction and Embryology (2017). Fact sheet 2: Regulation and legislation in assisted reproduction (ESHRE).

Franasiak, J.M., Forman, E.J., Hong, K.H., Werner, M.D., Upham, K.M., Treff, N.R., and Scott, R.T., Jr. (2014). The nature of aneuploidy with increasing age of the female partner: a review of 15,169 consecutive trophectoderm biopsies evaluated with comprehensive chromosomal screening. Fertil. Steril. *101*, 656–663.e1.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. *51*, 1044–1051.

García-Ruiz, A., Cole, J.B., VanRaden, P.M., Wiggans, G.R., Ruiz-López, F.J., and Van Tassell, C.P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. USA *113*, E3995–E4004.

Grunauer, M., and Jorge, A.A.L. (2018). Genetic short stature. Growth Horm. IGF Res. *38*, 29–33.

Hagenaars, S.P., Harris, S.E., Davies, G., Hill, W.D., Liewald, D.C.M., Ritchie, S.J., Marioni, R.E., Fawns-Ritchie, C., Cullen, B., Malik, R., et al.; METASTROKE Consortium, International Consortium for Blood Pressure GWAS; SpiroMeta Consortium; CHARGE Consortium Pulmonary Group, CHARGE Consortium Aging and Longevity Group (2016). Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112-151) and 24 GWAS consortia. Mol. Psychiatry *21*, 1624–1632.

Hatzimanolis, A., Bhatnagar, P., Moes, A., Wang, R., Roussos, P., Bitsios, P., Stefanis, C.N., Pulver, A.E., Arking, D.E., Smyrnis, N., et al. (2015). Common genetic variation and schizophrenia polygenic risk influence neurocognitive performance in young adulthood. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. *168*, 392.

Haworth, S., Mitchell, R., Corbin, L., Wade, K.H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani, G., Paternoster, L., Smith, G.D., et al. (2019). Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat. Commun. *10*, 333.

Hikabe, O., Hamazaki, N., Nagamatsu, G., Obata, Y., Hirao, Y., Hamada, N., Shimamoto, S., Imamura, T., Nakashima, K., Saitou, M., and Hayashi, K. (2016). Reconstitution in vitro of the entire cycle of the mouse female germ line. Nature *539*, 299–303.

Hill, W.D., Arslan, R.C., Xia, C., Luciano, M., Amador, C., Navarro, P., Hayward, C., Nagy, R., Porteous, D.J., McIntosh, A.M., et al. (2018). Genomic analysis of family data reveals additional genetic effects on intelligence and personality. Mol. Psychiatry *23*, 2347–2362.

Hill, W.D., Harris, S.E., and Deary, I.J. (2019). What genome-wide association studies reveal about the association between intelligence and mental health. Curr. Opin. Psychol. *27*, 25–30.

Hoge, S.K., and Appelbaum, P.S. (2012). Ethics and neuropsychiatric genetics: a review of major issues. Int. J. Neuropsychopharmacol. *15*, 1547–1557.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Hsu, S. (2016). Super-intelligent humans are coming (Nautilus).

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

Israel Ministry of Health (2013). Preimplantation Genetic Diagnosis: Administrative Circular. https://www.health.gov.il/hozer/mr29_2013.pdf.

Jelenkovic, A., Sund, R., Hur, Y.-M., Yokoyama, Y., Hjelmborg, J.V., Möller, S., Honda, C., Magnusson, P.K., Pedersen, N.L., Ooki, S., et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. Sci. Rep. *6*, 28496.

Jones, H.W.J., and Cohen, J. (2007). Chapter 14: Preimplantation genetic diagnosis. Fertil. Steril. *87*, S47–S49.

Katz, J. (1994). Informed consent–must it remain a fairy tale? J. Contemp. Health Law Policy *10*, 69–91.

Keyes, K.M., Smith, G.D., Koenen, K.C., and Galea, S. (2015). The mathematical limits of genetic prediction for complex chronic disease. J. Epidemiol. Community Health *69*, 574–579.

Khan, R., and Mittelman, D. (2018). Consumer genomics will change your life, whether you get tested or not. Genome Biol. *19*, 120.

Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. Am. J. Hum. Genet. *104*, 65–75.

Kim, M.S., Patel, K.P., Teng, A.K., Berens, A.J., and Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. Genome Biol. *19*, 179.

Knoppers, B.M., Bordet, S., and Isasi, R.M. (2006). Preimplantation genetic diagnosis: an overview of socio-ethical and legal considerations. Annu. Rev. Genomics Hum. Genet. *7*, 201–221.

Krapohl, E., Patel, H., Newhouse, S., Curtis, C.J., von Stumm, S., Dale, P.S., Zabaneh, D., Breen, G., O'Reilly, P.F., and Plomin, R. (2018). Multi-polygenic score approach to trait prediction. Mol. Psychiatry *23*, 1368–1374.

Kumar, A., Ryan, A., Kitzman, J.O., Wemmer, N., Snyder, M.W., Sigurjonsson, S., Lee, C., Banjevic, M., Zarutskie, P.W., Lewis, A.P., et al. (2015). Whole genome prediction for preimplantation genetic diagnosis. Genome Med. *7*, 35.

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al.; 23andMe Research Team; COGENT (Cognitive Genomics Consortium); Social Science Genetic Association Consortium (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. *50*, 1112–1121.

Lello, L., Avery, S.G., Tellier, L., Vazquez, A.I., de Los Campos, G., and Hsu, S.D.H. (2018). Accurate genomic prediction of human height. Genetics *210*, 477–497.

Lencz, T., Yu, J., Palmer, C., Carmi, S., Ben-Avraham, D., Barzilai, N., Bressman, S., Darvasi, A., Cho, J.H., Clark, L.N., et al. (2018). High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. Hum. Genet. *137*, 343–355.

Lin, M.-H., Wu, F.S.-Y., Lee, R.K.-K., Li, S.-H., Lin, S.-Y., and Hwu, Y.-M. (2013). Dual trigger with combination of gonadotropin-releasing hormone agonist and human chorionic gonadotropin significantly improves the live-birth rate for normal responders in GnRH-antagonist cycles. Fertil. Steril. *100*, 1296–1302.

Loos, R.J.F., and Janssens, A.C.J.W. (2017). Predicting polygenic obesity using genetic information. Cell Metab. *25*, 535–543.

Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480.

Marquez-Luna, C., Gazal, S., Loh, P.-R., Furlotte, N., Auton, A., and Price, A.L.; 23andMe Research Team (2018). Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv. https://doi.org/10.1101/375337.

Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649.

Meuwissen, T., Hayes, B., and Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. Anim. Front. *6*, 6.

Morris, T.T., Davies, N.M., Hemani, G., and Smith, G.D. (2019). Why are education, socioeconomic position and intelligence genetically correlated? bioRxiv. https://doi.org/10.1101/630426.

Mostafavi, H., Harpak, A., Conley, D., Pritchard, J.K., and Przeworski, M. (2019). Variable prediction accuracy of polygenic scores within an ancestry group. bioRxiv. https://doi.org/10.1101/629949.

Mullaart, E., and Wells, D. (2018). Embryo Biopsies for Genomic Selection. In Animal Biotechnology 2, H. Niemann and C. Wrenzycki, eds. (Springer), pp. 81–94.

National Academies of Sciences Engineering and Medicine (2017). Human genome editing: science, ethics, and governance (National Academies Press).

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.

Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S.F.W., et al.; LifeLines Cohort Study (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature *533*, 539–542.

Oltenacu, P.A., and Algers, B. (2005). Selection for increased production and the welfare of dairy cows: are new breeding goals needed? Ambio *34*, 311–315.

Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.

Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.

President's Council on Bioethics (US) (2003). Beyond therapy: Biotechnology and the pursuit of happiness (ReganBooks).

Privé, F., Vilhjalmsson, B.J., Aschard, H., and Blum, M.G.B. (2019a). Making the most of Clumping and Thresholding for polygenic scores. bioRxiv. https://doi.org/10.1101/653204.

Privé, F., Aschard, H., and Blum, M.G.B. (2019b). Efficient implementation of penalized regression for genetic risk prediction. Genetics *212*, 65–74.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

Rauw, W.M., Kanis, E., Noordhuizen-Stassen, E., and Grommers, F.J. (1998). Undesirable side effects of selection for high production efficiency in farm animals: a review. Livestock Prod. Sci. *56*, 15.

Robinson, M.R., Kleinman, A., Graff, M., Vinkhuyzen, A.A.E., Couper, D., Miller, M.B., Peyrot, W.J., Abdellaoui, A., Zietsch, B.P., Nolte, I.M., et al. (2017). Genetic evidence of assortative mating in humans. Nat. Hum. Behav. *1*, 16.

Rodenburg, T.B., and Turner, S.P. (2012). The role of breeding and genetics in the welfare of farm animals. Anim. Front. *2*, 16.

Roshandel, D., Klein, R., Klein, B.E.K., Wolffenbuttel, B.H.R., van der Klauw, M.M., van Vliet-Ostaptchouk, J.V., Atzmon, G., Ben-Avraham, D., Crandall, J.P., Barzilai, N., et al.; DCCT/EDIC Research Group (2016). New locus for skin intrinsic fluorescence in type 1 diabetes also associated with blood and skin glycated proteins. Diabetes *65*, 2060–2071.

Sathyan, S., Barzilai, N., Atzmon, G., Milman, S., Ayers, E., and Verghese, J. (2018). Genetic insights into frailty: Association of 9p21-23 locus with frailty. Front. Med. (Lausanne) *5*, 105.

Savage, J.E., Jansen, P.R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C.A., Nagel, M., Awasthi, S., Barr, P.B., Coleman, J.R.I., et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat. Genet. *50*, 912–919.

Savulescu, J. (2001). Procreative beneficence: why we should select the best children. Bioethics *15*, 413–426.

Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In 9th Python in Science Conference, pp. 57–61.

Selzam, S., Ritchie, S.J., Pingault, J.-B., Reynolds, C.A., O'Reilly, P.F., and Plomin, R. (2019). Comparing Within- and Between-Family Polygenic Score Prediction. Am. J. Hum. Genet. *105*, 351–363.

Sexton, C.E., Ebbert, M.T.W., Miller, R.H., Ferrel, M., Tschanz, J.A.T., Corcoran, C.D., Ridge, P.G., and Kauwe, J.S.K.; Alzheimer's Disease Neuroimaging Initiative (2018). Common DNA Variants Accurately Rank an Individual of Extreme Height. Int. J. Genomics *2018*, 5121540.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. Am. J. Hum. Genet. *99*, 139–153.

Shulman, C., and Bostrom, N. (2014). Embryo Selection for Cognitive Enhancement: Curiosity or Game-changer? Glob. Policy *5*, 85.

Smyrnis, N., Avramopoulos, D., Evdokimidis, I., Stefanis, C.N., Tsekou, H., and Stefanis, N.C. (2007). Effect of schizotypy on cognitive performance and its tuning by COMT val158 met genotype variations in a large population of young men. Biol. Psychiatry *61*, 845–853.

Stefanis, N.C., Smyrnis, N., Avramopoulos, D., Evdokimidis, I., Ntzoufras, I., and Stefanis, C.N. (2004). Factorial composition of self-rated schizotypal traits among young males undergoing military training. Schizophr. Bull. *30*, 335–350.

Stefanis, N.C., Trikalinos, T.A., Avramopoulos, D., Smyrnis, N., Evdokimidis, I., Ntzani, E.E., Ioannidis, J.P., and Stefanis, C.N. (2007). Impact of schizophrenia candidate genes on schizotypy and cognitive endophenotypes at the population level. Biol. Psychiatry *62*, 784–792.

Sullivan-Pyke, C., and Dokras, A. (2018). Preimplantation Genetic Screening and Preimplantation Genetic Diagnosis. Obstet. Gynecol. Clin. North Am. *45*, 113–125.

Sunkara, S.K., Rittenberg, V., Raine-Fenning, N., Bhattacharya, S., Zamora, J., and Coomarasamy, A. (2011). Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. Hum. Reprod. *26*, 1768–1774.

Tabery, J. (2015). Why Is Studying the Genetics of Intelligence So Controversial? Hastings Cent. Rep. *45* (5, Suppl), S9–S14.

Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv. https://doi.org/10.1101/563866.

Tenesa, A., Rawlik, K., Navarro, P., and Canela-Xandri, O. (2016). Genetic determination of height-mediated mate choice. Genome Biol. *16*, 269.

Teoh, P.J., and Maheshwari, A. (2014). Low-cost in vitro fertilization: current insights. Int. J. Womens Health *6*, 817–827.

The Economist. (2018). A slippery slope towards designer babies?. https://www.economist.com/science-and-technology/2018/11/14/a-slippery-slope-towards-designer-babies.

Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590.

Treff, N.R., Zimmerman, R., Bechor, E., Hsu, J., Rana, B., Jensen, J., Li, J., Samoilenko, A., Mowrey, W., Van Alstine, J., et al. (2019). Validation of concurrent preimplantation genetic testing for polygenic and monogenic disorders, structural rearrangements, and whole and segmental chromosome aneuploidy with a single universal platform. Eur. J. Med. Genet. *62*, 103647.

van der Werf, J. (2013). Genomic Selection in Animal Breeding Programs. In Genome-Wide Association Studies and Genomic Prediction, B. Hayes, ed. (Humana Press), p. 543.

Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.

Visscher, P.M., McEvoy, B., and Yang, J. (2010). From Galton to GWAS: quantitative genetics of human height. Genet. Res. *92*, 371–379.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.

Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. *17*, 9–18.

Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M., et al. (2019). Recovery of trait heritability from whole genome sequence data. bioRxiv. https://doi.org/10.1101/588020.

Ware, E.B., Schmitz, L.L., Faul, J.D., Gard, A., Mitchell, C., Smith, J.A., Zhao, W., Weir, D., and Kardia, S.L.R. (2017). Heterogeneity in polygenic scores for common human traits. bioRxiv. https://doi.org/10.1101/106062.

Wikler, D. (1999). Can we learn from eugenics? J. Med. Ethics *25*, 183–194.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507–515.

Wray, N.R., Kemper, K.E., Hayes, B.J., Goddard, M.E., and Visscher, P.M. (2019). Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. Genetics *211*, 1131–1141.

Yamashiro, C., Sasaki, K., Yabuta, Y., Kojima, Y., Nakamura, T., Okamoto, I., Yokobayashi, S., Murase, Y., Ishikura, Y., Shirane, K., et al. (2018). Generation of human oogonia from induced pluripotent stem cells in vitro. Science *362*, 356.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al.; LifeLines Cohort Study (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.

Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., and Visscher, P.M.; GIANT Consortium (2018a). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum. Mol. Genet. *27*, 3641–3649.

Yengo, L., Robinson, M.R., Keller, M.C., Kemper, K.E., Yang, Y., Trzaskowski, M., Gratten, J., Turley, P., Cesarini, D., Benjamin, D.J., et al. (2018b). Imprint of assortative mating on the human genome. Nat. Hum. Behav. *2*, 948–954.

Zeevi, D., Bloom, J.S., Sadhu, M.J., Ben Yehuda, A., Zangen, D., Levy-Lahad, E., and Kruglyak, L. (2019). Analysis of the genetic basis of height in large Jewish nuclear families. PLoS Genet. *15*, e1008082.

Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., Pourcain, B.S., et al.; Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics *33*, 272–279.

Please cite this article in press as: Karavani et al., Screening Human Embryos for Polygenic Traits Has Limited Utility, Cell (2019), https://doi.org/10.1016/j.cell.2019.10.033

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
The Longevity study	Sathyan et al., 2018; Roshandel et al., 2016; Eny et al., 2014; Chang et al., 2014	https://www.einstein.yu.edu/centers/aging/longevity-genes-project/
The Schizophrenia study (ASPIS)	Smyrnis et al., 2007; Hatzimanolis et al., 2015; Stefanis et al., 2007	N/A
Height study genotypes and phenotypes	Zeevi et al., 2019	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001852.v1.p1
The Ashkenazi Genome Consortium data	Lencz et al., 2018	https://ega-archive.org/studies/EGAS00001000664
Summary statistics for height and BMI	Yengo et al., 2018a	http://cnsgenomics.com/data.html
Summary statistics for cognitive ability (our stats were modified to exclude the ASPIS cohort)	Savage et al., 2018	https://ctg.cncr.nl/software/summary_statistics
Genetic maps	International HapMap 3 Consortium et al., 2010	ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/
Software and Algorithms		
PLINK	Chang et al., 2015	https://www.cog-genomics.org/plink2/
SHAPEIT2	O'Connell et al., 2014	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
Impute2	Howie et al., 2009	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
Python code implementing the data analyses	This study	https://bitbucket.org/ehudk/embryo-pgs-selection
An R implementation of the quantitative genetic model	This study	https://github.com/orzuk/EmbryoSelectionCalculator.

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Shai Carmi (shai.carmi@huji.ac.il). This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cohorts for simulating offspring

Longevity

Our data included 208 individuals from 104 couples who were part of the LonGenity study of longevity and aging in Ashkenazi Jews (the "Longevity" cohort). Genotyping was performed using Illumina HumanOmniExpress array. Genotyping and QC were previously described (Chang et al., 2014; Eny et al., 2014; Roshandel et al., 2016; Sathyan et al., 2018). The number of SNPs was 704,759, with an average missing rate 0.2%. We removed duplicate variants and variants with missingness rate > 1%. Height was available for all individuals except two who were discarded along with their spouses. Height was 177 ± 6cm (mean ± SD) in males (range 163-191) and 163 ± 6cm in females (range 147-175). BMI was available for 203 individuals. BMI was 26.5 ± 3.9 (kg/m^2) (mean ± SD) in males (range 15.9-42.9) and 25.4 ± 5.0 (kg/m^2) in females (range 18.0-51.2).

ASPIS

The Athens Study of Psychosis Proneness and Incidence of Schizophrenia (Stefanis et al., 2004) (henceforth "ASPIS") included 1066 randomly selected young male conscripts aged 18 to 24 years from the Greek Air Force in their first two weeks of admission. All participants were free of serious medical conditions. Cognitive measures included: Raven Progressive Matrices Test (Raven Matrices; raw score); Continuous Performance Task, Identical Pairs version (CPT-IP; d-prime score); Verbal N-Back working memory task (Verbal NBack; total accuracy); and Spatial N-Back working memory task (Spatial NBack; total accuracy). General cognitive ability scores

(*g*) were generated using the first principal component. We transformed the scores to IQ points by scaling the mean to 100 and the standard deviation to 15 (range 47-140). We note that this measure of cognitive ability is only an estimate of the IQ as would have been obtained from standard tests (such as the Wechsler Adult Intelligence Scale), and hence the somewhat wide range. Genotyping was performed on Affymetrix 6.0 arrays (Hatzimanolis et al., 2015; Smyrnis et al., 2007; Stefanis et al., 2007). The number of SNPs was 487,126, with an average missingness rate of 0.3%. Out of the 1066 genotyped samples, 147 had their cognitive function scores missing and were discarded from the analysis, leaving 919 individuals.

Nuclear families

We used 28 large nuclear Jewish families with an average of 9.6 adult offspring (full-siblings) per family who have completed their growth. The families were recruited in Israel and in the US after obtaining IRB approvals in both locations. Details on the cohort, measurements, and genotyping appear elsewhere (Zeevi et al., 2019). In short, participants signed a consent form and filled a medical questionnaire (to ensure there were no medical conditions that could have affected their growth), and their heights were measured with four technical repeats at an accuracy of ± 0.1cm. All 308 consented participants were genotyped on the Affymetrix Axiom Biobank array (≈630,000 SNPs). One from each of six pairs of monozygotic twins was excluded. Heights were corrected for age and age^2, then standardized to *Z*-scores in each sex separately, then reported as 173.0 + 5.6*Z*cm.

METHOD DETAILS

Phasing

We phased the Longevity and ASPIS cohorts (separately) using SHAPEIT2 (O'Connell et al., 2014). Default parameters were used, except for using 200 states (to improve precision), and an effective population size of 12k, similar to the value suggested for Europeans. The genetic map used was from HapMap (International HapMap 3 Consortium, 2010).

Polygenic score calculation

Longevity cohort: height and BMI

We used summary statistics from Yengo et al. (2018a), a meta-analysis based on Wood et al. (2014) and the UK Biobank (Bycroft et al., 2018). Effect sizes were available for 2,334,001 SNPs, of which 1,789,210 were missing from the Longevity panel. Another 241 variants had mismatching alleles, leaving a total of 544,550 for downstream analyses. Scoring of individuals based on the summary statistics was performed in PLINK (Chang et al., 2015) with the no-mean-imputation flag.

Given a polygenic score (PS), we predicted height in a two-step approach. First, the heights of the Longevity individuals were regressed against age and sex. Second, the residuals from the first step were regressed against their PS. The regression line from the second step was used to predict the height of the simulated offspring.

To optimize the PS, we first determined whether imputation had an effect on prediction accuracy. We used IMPUTE2 (Howie et al., 2009) and The Ashkenazi Genome Consortium reference panel (Lencz et al., 2018). Imputed data was post-processed to include only single nucleotide variants present in the summary statistics and with IMPUTE2 INFO-score > 0.9. The r^2_{ps} for height prediction (using all SNPs) was 0.201, which was slightly lower than for the PS generated without imputation, consistent with previous reports (Ware et al., 2017). Since imputation incurs a significant computational and storage burden, we proceeded with the genotyped SNPs only.

Next, we considered the effect of linkage-disequilibrium (LD) pruning and p value thresholds. LD-clumping was performed in PLINK (Chang et al., 2015) with window size of 250kb and r^2 threshold of 0.1. LD was estimated based on 574 genomes from The Ashkenazi Genome Consortium (Lencz et al., 2018), reduced to the 657,179 SNPs intersecting with the Longevity study. The number of remaining SNPs after LD-clumping was 93,345. We considered p value thresholds between 10^{-7} to 1 in multiples of 10. We then searched for the parameter combination giving the maximum r^2_{ps} between predicted and actual phenotypes. Without LD-pruning, the maximal r^2_{ps} was 0.207 (using a p value cutoff of 0.1). With LD-pruning, the maximal r^2_{ps} was 0.243 (comparable to Yengo et al., 2018a; Figure S4), using a p value cutoff of 0.001. Thus, our final score used LD-pruning and $p < 0.001$, and included 15,752 SNPs.

We used the same GWAS (Yengo et al., 2018a) to obtain summary statistics for BMI. We regressed BMI on age and sex, and then we regressed the residuals on the PS. The optimal parameters were $r^2 = 0.1$ and $p = 0.1$, and the optimal score included 15,695 SNPs and explained 3.1% of the variance (Figure S4). This is less than previously reported (≈10% of the variance) but was significantly non-zero. Scores for BMI were only used for the analysis of selection for multiple traits.

Nuclear families: height

We used the set of 15,752 SNPs obtained for the Longevity cohort with the thresholds $p < 0.001$ and LD $r^2 < 0.1$. Of these, we used 15,124 SNPs that were present on the array or could be imputed from the AJ reference panel (Carmi et al., 2014). We excluded SNPs homozygous in all participants. The weight of each SNP was its effect size (Yengo et al., 2018a), zero centered for the cohort. Scores were standardized into *Z*-scores and reported as for the actual heights.

ASPIS: general cognitive ability

We used summary statistics from Savage et al. (2018), based on a meta-analysis of intelligence (excluding the ASPIS cohort). Out of total of 9,145,263 SNPs, 468,809 intersected with the ASPIS panel. Following the results from height, we did not consider imputation. The optimal LD-clumping threshold and p value threshold were $r^2 = 0.3$ and 1, respectively, leaving 130,199 SNPs and reaching $r^2_{ps} = 0.043$ (Figure S4). For improving the accuracy of LD estimation, we considered the entire 1066 genotyped individuals, including those without phenotypes.

We note that other approaches for genetic prediction may have slightly higher predictive power. However, an extensive benchmarking of methods and thresholds for trait prediction is beyond the scope of this paper. Our quantitative model can approximate the utility of any score based on its proportion of variance explained.

Simulating embryos

The Longevity cohort included actual couples, and these were used to simulate offspring ("actual matching"). For both the Longevity and the ASPIS cohorts, we also matched parents randomly ("random matching"). Given a pair of parents, we simulated offspring (embryos) by specifying the locations of crossovers in each parent. Recombination was modeled as a Poisson process, with distances measured in cM using the HapMap genetic map (International HapMap 3 Consortium, 2010). For each parent, we drew the number of crossovers in each chromosome from a Poisson distribution with mean equal to the chromosome length in Morgans. Random positions along the chromosome (in Morgans) represented the locations of the crossovers. We mixed the phased paternal and maternal chromosomes of the parent according to the crossovers' locations, and randomly chose one of the resulting sequences as the chromosome transmitted from that parent. Note that due to phase switch errors, the paternal and maternal chromosomes are each a mixture of both. Nevertheless, phasing is expected to be accurate over short distances (switch error rate around 1%) (Choi et al., 2018), thus correctly representing LD blocks.

We repeated the process to generate either 10 or 50 embryos per couple (whether a true couple or randomly matched). The number of couples for random matches was such that the total number of embryos was 5000. For a number of embryos other than 10 or 50, we downsampled embryos from the $n = 50$ simulations.

The Sets of Simulated Embryos

Cohort	Phenotype	Matching	Matches (n)	Offspring per Couple (n)
Longevity	height/BMI	random	500	10
Longevity	height/BMI	random	100	50
Longevity	height	actual	102	10
Longevity	height	actual	102	50
ASPIS	cognitive ability	random	500	10
ASPIS	cognitive ability	random	100	50

To calculate the polygenic scores for the synthetic embryos, we used the same summary statistics as for the parents. To predict the phenotypes of the embryos, we used the regression model that we have generated for the parents. The predicted phenotype is thus in its natural units (cm, kg/m^2, or IQ points). Adding sex- or age-specific means was unnecessary, as we considered only the differences between embryos attributed to their polygenic scores.

Multiple traits

We used the Longevity cohort, which had data on both height and BMI. We used the same sets of simulated embryos as for height. For each embryo, we computed the scores for height and BMI, and normalized the scores by the standard deviations of the predicted phenotypes (2.89cm for height and 0.78kg/m^2 for BMI). The combined score per embryo was the normalized height score minus the normalized BMI score (to simulate selection for *lower* BMI). The gain for height was the predicted height for the embryo with the highest combined score, and similarly for BMI. The correlation between the scores of height and BMI was -0.16, which we used in the equations for the gain (Equations 3 and 4).

QUANTIFICATION AND STATISTICAL ANALYSIS

Polygenic scores were calculated with PLINK (Chang et al., 2015). Other data analyses were performed using custom Python and R scripts.

In Figures 1, 2, S1, and S2, 95% confidence intervals are based on $\pm$ 1.96 the standard error of the mean (SEM) over the simulated families. Regressions (as in Figure S4) were performed using statsmodels (Seabold and Perktold, 2010). For the regression of the trait on the PS, the proportion of variance explained was the squared correlation coefficient, and the p value for a non-zero correlation coefficient was computed with scipy.stats.pearsonr. The mean and 95% confidence bands in Figure S4 were generated by bootstrapping with seaborn.lmplot.

The quantitative genetic model

We modeled the vector of polygenic scores for a set of embryos as a multivariate normal variable with zero means, and derived its covariance matrix. The model implies that the score of each embryo can be represented as a sum of two normal variables, one shared across embryos and one independent, both with variance equal to half the variance in the trait explained by the PS. The maximal score, and thereby the gain, could be written using the maximum of n independent normal variables. We derived formulas for the mean and variance of the gain, and then: the mean gain conditional on the parental scores and phenotypes, a prediction interval

for the phenotype, the difference between the maximal-predicted and the actual maximal trait value, and the gain when selecting for multiple traits. Full details are available in Methods S1.

DATA AND CODE AVAILABILITY

Python code implementing the analyses described in this paper is available at https://bitbucket.org/ehudk/embryo-pgs-selection.

R code that implements some of the calculations of the gain under the quantitative genetic model can be found at https://github.com/orzuk/EmbryoSelectionCalculator.

Cell

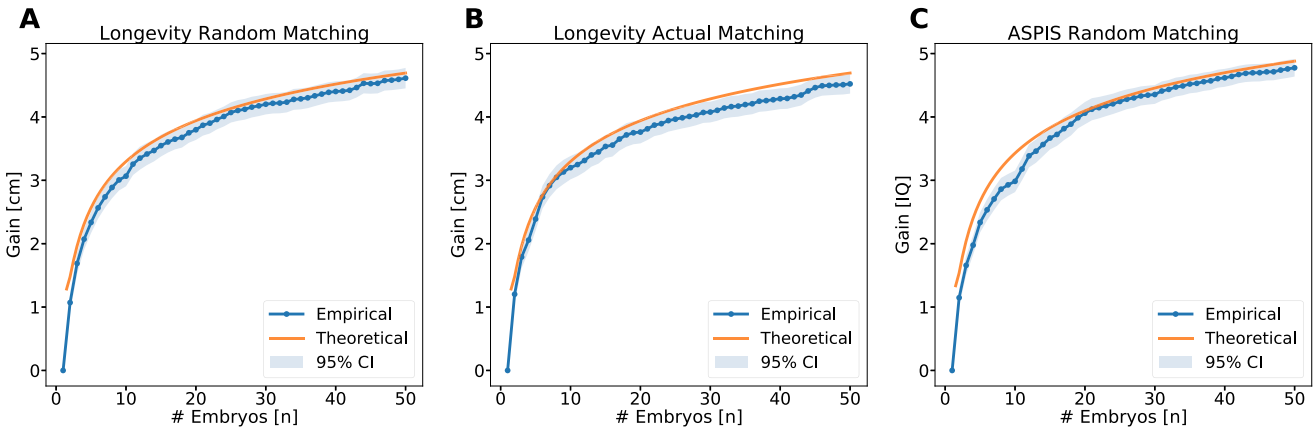


Figure S1. The Mean Gain in Embryo Selection versus the Number of Embryos *n*, Related to Figure 2

All details are the same as in Figure 2. The theoretical prediction here is based on extreme value theory, as given in Methods S1 Equation 33, providing a slightly better fit compared to main text Equation 1.

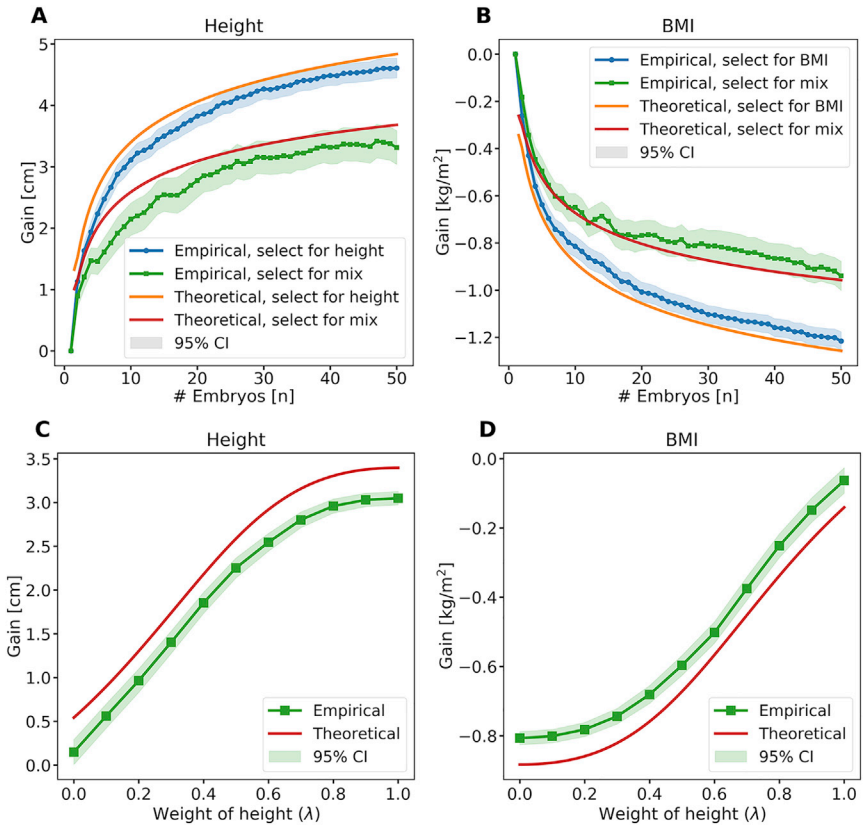


Figure S2. Selection for Multiple Traits, Related to Figure 2

We simulated up to 50 embryos per (random) couple from the Longevity cohort, and calculated PSs for height and BMI. In (A) (height) and (B) (BMI) we show the gain per trait when selecting either for the focal trait (height in (A) and (negative) BMI in (B), blue dots along with their 95% confidence intervals; as in Figures 2 or S1) or when selecting for the sum of the scores for height and negative BMI, after normalizing each score by its standard deviation (green dots along with their 95% confidence intervals). Note that the gain for BMI is negative, since we select for lower values of BMI. As expected, selection for multiple traits leads to lower gains (in absolute value) per trait. The orange lines correspond to Equation 33 of Methods S1 for the mean gain when selecting for a single trait. The red lines correspond to Equation 93 of Methods S1, where we used the expression for the mean of the maximum of n normal variables based on extreme value theory (Equation 33). In (C) and (D), we plot the gains for height and BMI, respectively, when the (normalized) score of height is weighted by λ and the (normalized) score for BMI is weighted by $-(1-\lambda)$. Green dots and 95% confidence intervals are based on simulations, whereas the red lines are based on Equation 99 of Methods S1. The gain in height increases and the gain in BMI decreases (in absolute value) with λ, as expected. Note that the gain is non-zero even for $\lambda = 0$ or $\lambda = 1$, due to the correlation between height and BMI.

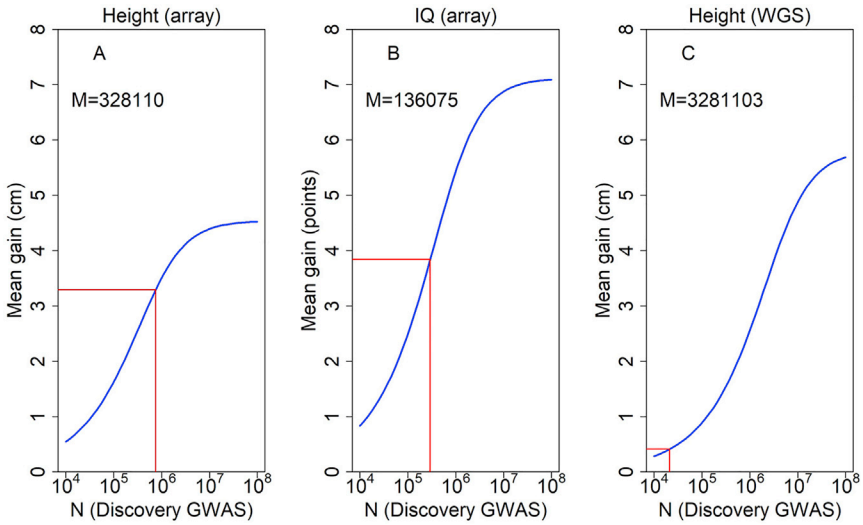


Figure S3. The Expected Increase in the Mean Gain with Discovery GWAS Sample Sizes, Related to Figure 1

To evaluate the expected gain given a GWAS sample size N, we used the relation $r^2_{ps} = h^2_{snp}(1 + M/(Nh^2_{snp}))^{-1}$ (Pasaniuc and Price, 2017; Wray et al., 2019). In this equation, h^2_{snp} is the SNP-based or chip heritability (the variance in the trait explained by all SNPs on the array) and M is the effective number of SNPs. To estimate M for height, we substituted $h^2_{snp} = 0.48$, $N \approx 700 \cdot 10^3$, and $r^2_{ps} \approx 0.24$ (Yengo et al., 2018a), which gave $M = 328 \cdot 10^3$. For IQ, $h^2_{snp} = 0.19$, $N \approx 270 \cdot 10^3$, and $r^2_{ps} = 0.05$ (Savage et al., 2018), which gave $M = 136 \cdot 10^3$. Given these values of M, we calculated the expected r^2_{ps} for a range of GWAS sample sizes. To compute the expected gain when selecting one embryo out of $n = 10$, we used an exact numerical solution for the mean of the maximum of independent normal variables (Methods S1 Equation 28), and assumed standard deviations of 6cm for height in (A) and 15 points for IQ in (B). The red lines denote the gain with current GWAS sizes. (C) The expected gain in height for scores based on whole-genome sequencing (WGS) data. Based on Wray et al. (2019), we used a value of M 10x larger compared to that of arrays, giving $M = 3.28 \cdot 10^6$. Instead of h^2_{snp}, we used $h^2_{wgs} = 0.79$ (Wainschtein et al., 2019).

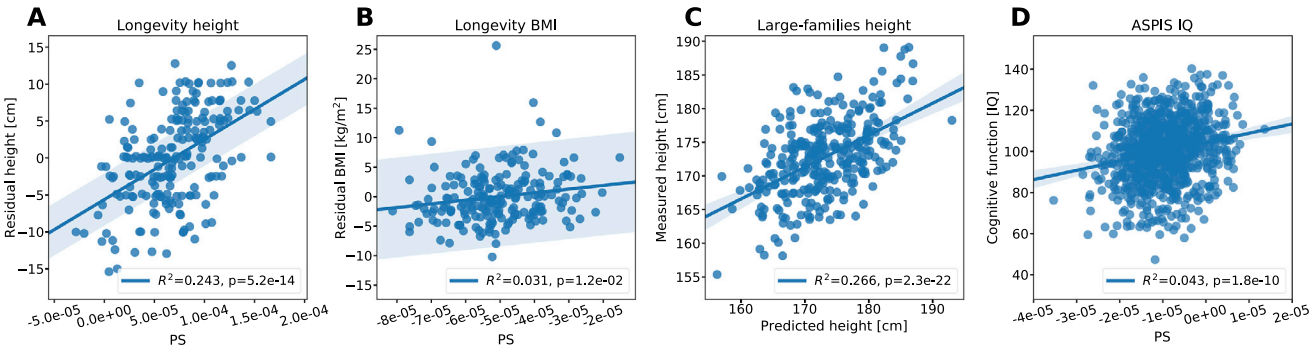


Figure S4. Height, BMI, and Cognitive Ability (IQ) versus Their Polygenic Scores, Related to STAR Methods

Results are shown for the heights and BMI of 204 individuals in the Longevity cohort (A) and (B), respectively), the heights of 308 individuals from the large nuclear families (C), and the IQ of 919 individuals from the ASPIS cohort (D). Also shown are the regression lines with 95% bootstrap confidence intervals (seaborn.lmplot), the proportions of variance explained, and the p values (scipy.stats.pearsonr). The proportions of variance explained by the polygenic scores are ≈25%–27% for height, ≈3% for BMI, and ≈4.3% for IQ.