Sibling Similarity Can Reveal Key Insights into Genetic Architecture

Tade Souaiaia[1,✉], **Hei Man Wu**[2], **Clive Hoggart**[2,✉,*], **and Paul O'Reilly**[2,✉,*]

[1]Department of Cellular Biology, Suny Downstate Health Sciences, Brooklyn, NY, USA

[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, 1 Gustave L Levy Pl, NY, NY, USA

[*]These authors contributed equally to this work

Correspondence: *tade.souaiaia@downstate.edu, clive.hoggart@mssm.edu, paul.oreilly@mssm.edu*

Abstract **The use of siblings to infer the factors influencing complex traits has been a cornerstone of quantitative genetics. Here we utilise siblings for a novel application: the inference of genetic architecture, specifically that relating to individuals with extreme trait values (e.g. in the top 1%). Inferring the genetic architecture most relevant to this group of individuals is important because they are at greatest risk of disease and may be more likely to harbour rare variants of large effect due to natural selection. We develop a theoretical framework that derives expected distributions of sibling trait values based on an index sibling's trait value, estimated trait heritability, and null assumptions that include infinitesimal genetic effects and environmental factors that are either controlled for or have combined Gaussian effects. This framework is then used to develop statistical tests powered to distinguish between trait tails characterised by common polygenic architecture from those that include substantial enrichments of *de novo* or rare variant (Mendelian) architecture. We apply our tests to UK Biobank data here, although we note that they can be used to infer genetic architecture in any cohort or health registry that includes siblings and their trait values, since these tests do not use genetic data. We describe how our approach has the potential to help disentangle the genetic and environmental causes of extreme trait values, and to improve the design and power of future sequencing studies to detect rare variants.**

Introduction

The fields of quantitative genetics and genetic epidemiology have exploited the shared genetics and environment of siblings in many applications, notably to estimate heritability, using theory developed a century ago (1, 2), and more recently to infer a so-called 'household effect' (3), which contributes to genetic risk indirectly via correlation between genetics and household environment. Here we leverage sibling trait data to infer genetic architecture departing from polygenicity, specifically affecting the tails of the trait distribution and consistent with enrichment in rare variants of large effect.

The genetic architecture of a complex trait is typically inferred from findings of multiple independent studies: genome-wide association studies (GWAS) identifying common variants (4, 5), whole exome or whole genome sequencing studies detecting rare variants (6), and family sequencing studies designed to identify *de novo* and rare Mendelian mutations (7). The relative contribution of each type of variant to trait heritability is a function of historical selection pressures on the trait in the population (8, 9). If selection has recently acted to increase the average value of a trait, then the lower tail of the trait distribution may become enriched for large effect rare variants over time because trait-reducing alleles will be subject to negative selection, and so those with large effects - most likely to 'push' individuals to the lower tail - will be reduced to low frequencies (10). Similarly, if the trait is subject to stabilising selection, then both tails of the trait distribution may be enriched for rare variant aetiology (11). This can result in less accurate polygenic scores in the tails of the trait distribution (12), and can also produce

dissimilarity between siblings beyond what is expected under polygenicity. For example, studies on the intellectual ability of sibling pairs have demonstrated similarity for average intellectual ability (13), regression-to-the-mean for siblings at the upper tail of the distribution (13), and complete discordance when one sibling is at the lower extreme tail of the distribution (14). However, no theoretical framework has been developed to formally infer genetic architecture from sibling trait data.

We introduce a novel theoretical framework that allows widely available sibling trait data in population cohorts and health registries to be leveraged to perform statistical tests that can infer complex genetic architecture in the tails of the trait distribution. These tests are powered to differentiate between polygenic, *de novo* and Mendelian (i.e. rare variants of large effect) architectures (see ***Figure 1***); while these are simplifications of true complex architectures, our tests allow enrichments of the different architectures to be inferred and compared across traits. This framework establishes expectations about the trait distributions of siblings of index individuals with extreme trait values (e.g. the top 1% of the trait) according to null assumptions of polygenicity, random mating and environmental factors that are either controlled for or have combined Gaussian effects. Critical to our framework is our derivation of the "conditional-sibling trait distribution", which describes the trait distribution for one individual given the quantile value of one or more "index" siblings. Our statistical framework, derivation of the conditional-sibling trait distribution, and simulation study, allow us to develop statistical tests to infer genetic architecture from data on siblings (15) without the need for genetic data (see ***Model & Methods***). We validate the statistical power of our tests using simulated data and apply our tests across a range of traits from the UK Biobank. We also release *sibArc*, an open-source software tool that can be used to apply our tests to sibling trait data (16). Our novel framework can be extended to applications such as estimating heritability, characterising mating patterns, and inferring historical selection pressures.

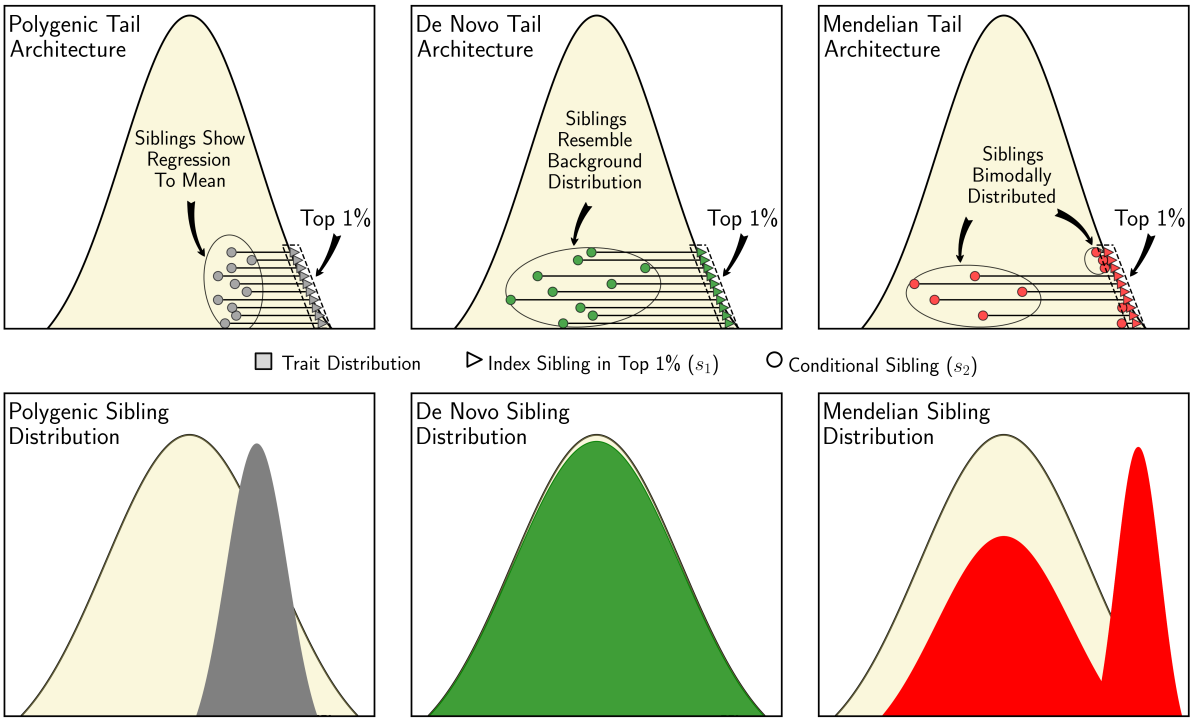


Fig. 1: Sibling Similarity Under Different Tail Architectures. Left to right: When an individual's extreme trait value (top 1%) is due to many alleles of small effect ("Polygenic"), then their siblings' trait values are expected to show regression-to-the-mean (grey). When an individual's extreme trait value is due to a *de novo* mutation of large effect, then their siblings are expected to have trait values that correspond to the background distribution (green). When an individual's extreme trait value is the result of an inherited rare allele of large effect ("Mendelian"), then their siblings are expected to have either similarly extreme trait values or trait values that are drawn from the background distribution (red), depending on whether or not they inherited the same large effect allele.

Model & Methods

Here we outline a framework that models the *conditional-sibling trait distribution*, which describes an individual's trait distribution conditioned on an *index* sibling's trait value. In this section, we describe our derivation of the distribution for a completely polygenic trait, outline development of statistical tests that infer enrichments of rare genetic architectures via departures from polygenicity, and describe the simulation study used to validate and benchmark our results. Full derivations that complement this high-level summary of our methodology, as well as extended results from application of our tests to UK Biobank data, are included in the ***Appendices***.

Conditional-Sibling Inference Framework

Assuming completely polygenic (i.e. infinitesimal) architecture, siblings of individuals with extreme trait values are expected to be less extreme. This regression-to-the-mean can be understood through the two factors (17, 18) that determine inherited genetic variation: **(i)** average genetic contribution of the parents (*mid-parent*) to the trait value, and, **(ii)** random genetic reassortment occurring during meiosis. Assuming that an individual presenting an upper-tail trait value does so due to *both* high mid-parent trait value *and* genetic reassortment favouring a higher value, then a sibling sharing common mid-parent trait value - but subject to independent reassortment - is likely to be less extreme. How much less extreme can be derived by first considering the conditional distribution that relates mid-parents and their offspring. In the simplest case, for a completely heritable continuous polygenic trait in a large population, the offspring trait values, s, are normally distributed around their mid-parent trait value, m, irrespective of selection and population structure (19, 20)

$$p(s \mid m) = \mathcal{N}\left(m, \sigma_f^2\right) \tag{1}$$

where σ_f^2 represents within family variance, which is constant across the population. This result does not require the trait distribution across the population to be Gaussian. If we assume that the population distribution is Gaussian and there is random mating and no selection, then $\sigma_f^2 =$ half the population variance (19, 21, 22). For a trait with heritability h^2, trait variance can be partitioned into genetic, σ_g^2, and environmental, σ_e^2, contributions, such that $\sigma^2 = \sigma_g^2 + \sigma_e^2$ and $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. Assuming mean-centred genetic and environmental trait contributions, and a trait variance of 1 (a standard normalised trait), the distribution for offspring conditioned on the mid-parent genetic trait value, m_g, is:

$$p(s \mid m_g) = \mathcal{N}\left(m_g, \sigma_g^2/2 + \sigma_e^2\right) = \mathcal{N}\left(m_g, 1 - \frac{h^2}{2}\right) \tag{2}$$

Since $p(m_g) = \mathcal{N}(0, h^2/2)$, from Bayes' Theorem it can be shown that the mid-parent genetic trait value conditional on offspring trait value, s, is distributed as follows:

$$p(m_g \mid s) = \mathcal{N}\left(\frac{1}{2}sh^2, \frac{h^2}{2}\left(1 - \frac{h^2}{2}\right)\right) \tag{3}$$

From ***Equations 2 & 3***, the sibling trait distribution conditional on an index sibling can be calculated as:

$$\begin{aligned} p(s_2 \mid s_1) &= \int p(s_2 \mid m_g) p(m_g \mid s_1)\, dm_g \\ &= \mathcal{N}\left(\frac{1}{2}s_1 h^2, 1 - \frac{h^4}{4}\right) \end{aligned} \tag{4}$$

A full derivation is provided on Appendix 1. This derivation via mid-parents can be generalised to cases where assumptions of random mating, Gaussian population trait distribution and no selection do

not hold. When these assumptions do hold, then the conditional sibling distribution can also be derived from the joint sibling distribution defined by the relationship matrix (23). In the Appendix, we use the joint sibling distribution to derive the distribution conditional on two sibling values. From (4), we can infer that under complete heritability, the siblings of an individual with a standard normal trait value of z, will have a mean trait value of $\frac{z}{2}$, with variance equal to three quarters of the population variance. In ***Appendix 1***, we further generalise the result to binary phenotypes, increasing the utility of this framework for further applications and theoretical development.

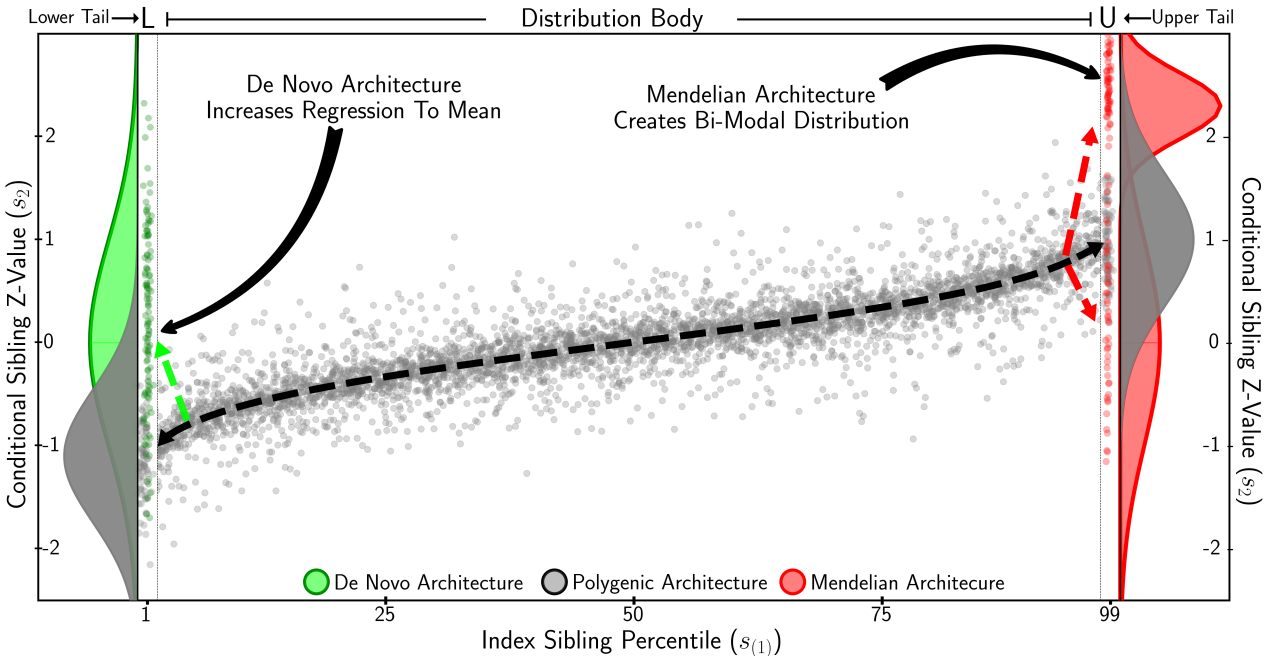


Fig. 2: Identifying Complex Tail Genetic Architecture. Conditional-sibling z-values plotted against index-sibling quantiles. Grey depicts complete polygenic architecture across index-sibling values. In the lower tail, an extreme scenario of *de novo* architecture is shown in green, resulting in sibling discordance. In the *Upper Tail*, extreme Mendelian architecture is shown in red, whereby siblings are half concordant and half discordant, resulting in a bimodal conditional sibling trait distribution. Statistical tests to infer each type of complex tail architecture are designed to exploit these expected trait distributions.

Statistical Tests for Complex Tail Architecture

In ***Figure 2***, the strategy employed to develop statistical tests for complex tail architecture is depicted. Our approach corresponds to testing for deviations from the expected conditional-sibling trait distribution under the null hypothesis of complete polygenicity in trait tails (see above for other null assumptions): excess discordance is indicative of an enrichment of *de novo* mutations, while excess concordance indicates an enrichment of Mendelian variants, i.e. large effect variants segregating in the population. The heritability, h^2, required to define the null distribution, is estimated by maximising the log-likelihood of the conditional-sibling trait distribution (***Equation 4***) with respect to h^2:

$$\log \mathcal{L} \propto -n \log\left(1 - \frac{h^4}{4}\right) - \frac{1}{1 - \frac{h^4}{4}} \sum_{i=1}^{n} \left(s_2^{(i)} - \frac{s_1^{(i)} h^2}{2}\right)^2 \tag{5}$$

where s_1 and s_2 represent index and conditional-sibling trait values, respectively, and n is number of sibling pairs. This allows h^2 to be estimated for given quantiles of the trait distribution by restricting sibling pair observations to those index siblings in the quantile of interest. To maximise power to detect non-polygenic architecture in the tails of the trait distribution, we estimate "polygenic heritability", h_p^2, from sibling pairs for which the index sibling trait value is between the 5th and 95th percentile (labelled

"Distribution Body" in ***Figure 2***).

Tests for complex architecture are then performed in relation to index siblings whose trait values are in the tails of the distribution (e.g. the lower and upper 1%). Below, A_q denotes the set of sibling pairs for which the index sibling is in quantile q, such that $s_1 > \Phi^{-1}(1-q)$ and $s_1 < \Phi^{-1}(q)$ for the upper and lower tails, respectively, where Φ^{-1} is the inverse normal cumulative distribution function.

It should be noted that our estimate of h^2 in ***Equation 5*** assumes no effects of shared environment. Polderman and colleagues (24) found limited contribution of shared environment for most complex traits and, critically, our statistical tests are robust to shared environmental effects with consistent effects throughout the trait distribution (see ***Discussion***).

Statistical Test for De Novo Architecture

For inference of *de novo* architecture in the tails of the trait distribution, we introduce a parameter, α, to the log-likelihood defined by the conditional-sibling trait distribution ***Equation 5***:

$$\log \mathcal{L} \propto \frac{1}{2(1-\frac{h^4}{4})} \sum_{i \in A_q} \left(s_2^{(i)} - \left(\frac{1}{2} s_1^{(i)} h^2 + \alpha \right) \right)^2 \tag{6}$$

Values of $\alpha > 0$ in the lower tail and $\alpha < 0$ in the upper tail indicate excess regression-to-the-mean and, thus, high sibling discordance, consistent with an enrichment of *de novo* mutations among the index siblings. The z-statistic of the one-sided score test for $\alpha > 0$ in the lower quantile, q, relative to the null of $\alpha = 0$ is (see ***Appendix 2*** for derivation):

$$z = \frac{\sum_{i \in A_q} s_2^{(i)} - \frac{1}{2} s_1^{(i)} h^2}{\sqrt{n(1-\frac{h^4}{4})}} \tag{7}$$

For the upper tail test of $\alpha < 0$, the above is multiplied by -1.

Statistical Test for Mendelian Architecture

For inference of Mendelian architecture in the tails of the trait distribution, we compare the observed and expected tail sibling concordance, defined by the number of sibling pairs for which both siblings have trait values in the tail. For each index sibling in A_q, we calculate the probability that the conditional sibling is also in A_q, which, for the upper tail, is given by:

$$P\left(s_2 > \Phi^{-1}(q) \mid s_1\right) = 1 - \Phi\left(\frac{\Phi^{-1}(q) - \frac{s_1 h^2}{2}}{\sqrt{1-\frac{h^4}{4}}} \right) \tag{8}$$

where Φ represents the normal cumulative distribution function. Denoting the mean of $P(s_2^{(i)} > \Phi^{-1}(q))$ across all index siblings in A_q by π_o, the expected sibling concordance is $n\pi_o$, where n is the number of index siblings in A_q. Given an observed number of concordant siblings, r, the z-statistic for a one-sided score test for excess concordance is (see ***Appendix 2*** for derivation) given by:

$$z = \frac{r - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} \tag{9}$$

Simulation of Conditional Sibling Data

We perform simulations using publicly available GWAS data on multiple traits to validate our analytical model and to benchmark our tests for complex architecture. ***Figure 3*** depicts the different stages of our simulation procedure. We start by simulating a "parent population" (**step A**), utilsing the allele frequencies and effect sizes of the first 100k SNPs from a trait GWAS to sample genotypes and subsequently trait values assuming an additive model. Next, parents are randomly paired and their genetic trait values averaged to produce mid-parent trait values (**step B**), and genotypes of two offspring (***Equation 1***) and corresponding genetic trait values, G, are calculated (**step C**) assuming independent reassortment of parental alleles and unlinked SNPs.

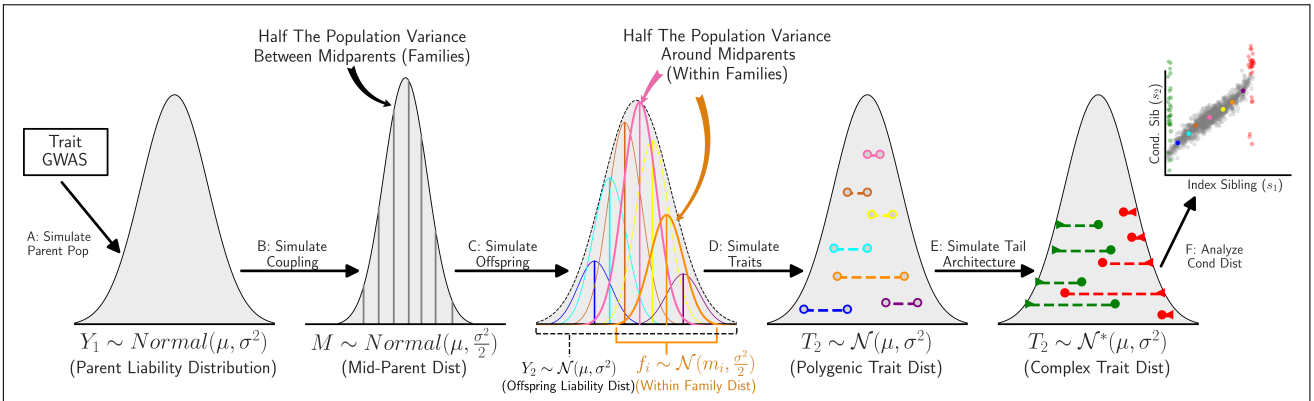


Fig. 3: Simulation Schematic. Publicly available GWAS allele frequency and effect size data is used to simulate parent genetic trait value (**A**). Midparent genetic trait value (**B**) is simulated assuming random mating. Offspring genotype and genetic trait value (**C**) is simulated assuming complete recombination. Environmental variation (**D**) is added to compare with theoretical polygenic conditional sibling distribution. *De novo* and Mendelian rare-variant effects are simulated (**E**) to benchmark tests for complex architecture (**F**).

In **step D**, we generate offspring trait values for different degrees of heritability by adding a Gaussian environmental effect. For heritability, h^2, offspring trait values are given by $T = hG + E$, where G is the genetic effect, standardised to have mean 0 and variance 1, and the environmental effect E is drawn from a normal distribution with mean 0 and variance $(1 - h^2)$. The simulated trait has a $\mathcal{N}(0, 1)$ distribution.

In **step E**, we simulate the effect of complex tail architecture on the conditional-sibling trait distribution. We assume that rare variants are sufficiently penetrant to move individuals into the tails of the distribution, independent of their polygenic contribution. We, thus, modify sibling trait values for individuals already in the tails (from Step D) to minimise perturbation of the trait distribution. We simulate *de novo* tail architecture by resampling the less extreme sibling from the background distribution, and simulate Mendelian tail architecture by resampling the less extreme sibling from the background distribution with probability 0.5, and from the same tail as the extreme sibling with probability 0.5.

Application to UK Biobank Data

The UK Biobank includes data from over 21,000 siblings (25). To apply our tests to this dataset we began by identifing continuous traits (at least 50 unique values) with at least 5,000 sibling pairs, as defined by kinship coefficient 0.18 - 0.35 and > 0.1% SNPs with IBS0 to distinguish from parent-offspring (26, 27). After removing outliers with absolute trait value > 8 standard deviations from the mean, we removed all traits with absolute skew or excess kurtosis greater than 0.5 to reduce the likelihood that skewed or heavy-tailed trait distributions impact our statistical tests. The remaining traits were standardised using rank-based inverse normal transformation and adjusted for age, sex,

recruitment centre, batch covariates and the first 40 principle components. After this, to ensure a primarily additive polygenic aetiology, we required that traits have heritability $h^2_{SNP} > 0.1$ (28) and that no single SNP contribute more than 0.01 to h^2.

We applied our method to estimate heritability from sibling pairs (***Equation 5***) across the distribution and within the trait body (5th to 95th percentiles) on the remaining traits, and selected the eighteen traits for which both measures exceeded 30% for further analysis. For each of these eighteen traits, siblings were randomly assigned index and conditional status and both trait tails were tested for departures from polygenicity using our *De Novo* and Mendelian tests, as well as a general Kolmogorov-Smirnov test (29) to identify departures from the conditional sibling distribution assuming polygenicity (***Appendix 2***).

Results

Here we illustrate the conditional-sibling trait distribution, validate the accuracy of our analytical model using simulation, perform power analyses for our statistical tests for complex genetic tail architecture (see ***Model & Methods***), and apply our tests to trait data on thousands of siblings from the UK Biobank.

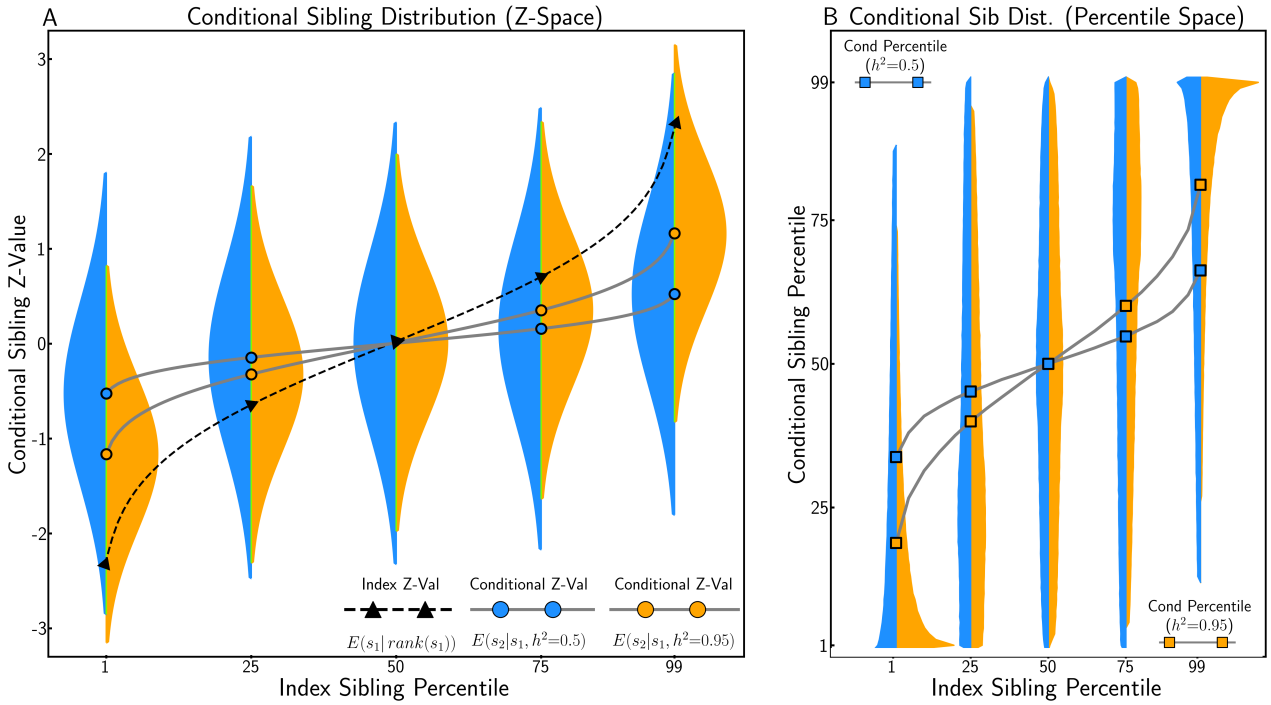


Fig. 4: Conditional-Sibling Trait Distribution under Polygenic Architecture. A: The conditional-sibling trait distribution according to ***Equation 4*** for index siblings at the 1st, 25th, 50th, 75th, and 99th percentile of the standardised trait distribution, when heritability is high ($h^2 = 0.95$, in orange) and moderate ($h^2 = 0.5$, in blue). When heritability is 0.95 conditional-sibling expectation is almost half of the index sibling z-score, when heritability is 0.5 the conditional-sibling expectation is equal to $1/4$ of the index sibling z-score. B: The conditional distribution transformed into rank space. An individual whose sibling is at the 99% percentile is expected to have a trait value in the 80% percentile when heritability is high and in the 67% percentile when heritability is moderate.

Conditional-Sibling Trait Distribution

In ***Figure 4***:A the conditional-sibling trait distribution (***Equation 4***) is illustrated at different index sibling trait values (ranked percentiles). For an almost entirely heritable polygenic trait (orange), siblings of individuals at the 99th percentile ($z = 2.32$) have mean z-scores approximately halfway between the population mean and index mean (i.e. $z = 1.1$). This regression-to-the-mean is greater when trait heritability is lower (blue), assuming (as here) independent environmental risk among siblings.

In ***Figure 4***:B the conditional-sibling z-distribution is transformed into percentiles for interpretation in rank space. This distribution is skewed, especially at the tails, due to truncation at extreme quantiles (i.e. siblings cannot be more extreme than the top 1%). For a trait with $h^2 = 0.95$, siblings of individuals at the 99th percentile (z=2.32) have a mean trait value at the 80th percentile. Note that this is less extreme than the result of transforming their expected z-value into percentile space ($\Phi^{-1}(1.1) = 86\%$), which is a consequence of Jensen's inequality (30) given that the inverse cumulative distribution function of the normal is convex above zero.

We compared the theoretical conditional-sibling trait distributions to those generated from simulated data (see ***Appendix 3***) and found that irrespective of trait used to simulate data (e.g. fluid intelligence, height) the two distributions did not differ significantly, suggesting that our analytically derived distributions are a valid model for the conditional-sibling trait distribution (***Equation 4***).

Power of Statistical Tests to Identify Complex Tail Architecture

Building on the theoretical framework introduced in the ***Model & Methods*** and illustrated in the previous section, we develop statistical tests to identify complex architecture in the tails of the trait distribution. These tests leverage the fact that the similarity (or dissimilarity) in trait values among siblings provides information about the underlying genetic architecture (see ***Figure 1***). For example, high-impact *de novo* mutations generate large dissimilarity between siblings when only one carries the unique mutant allele, while Mendelian variants can create excess similarity in the tails of the distribution when siblings both inherit the same mutant allele.

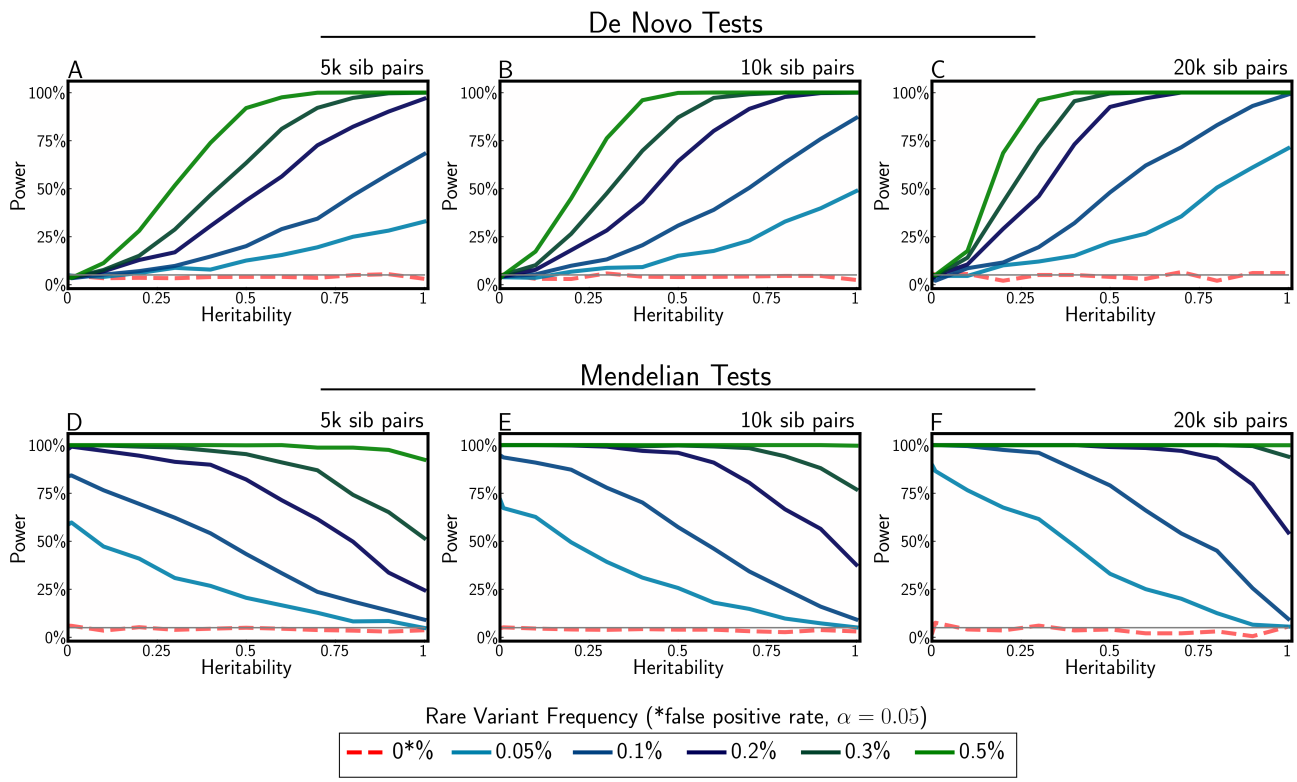
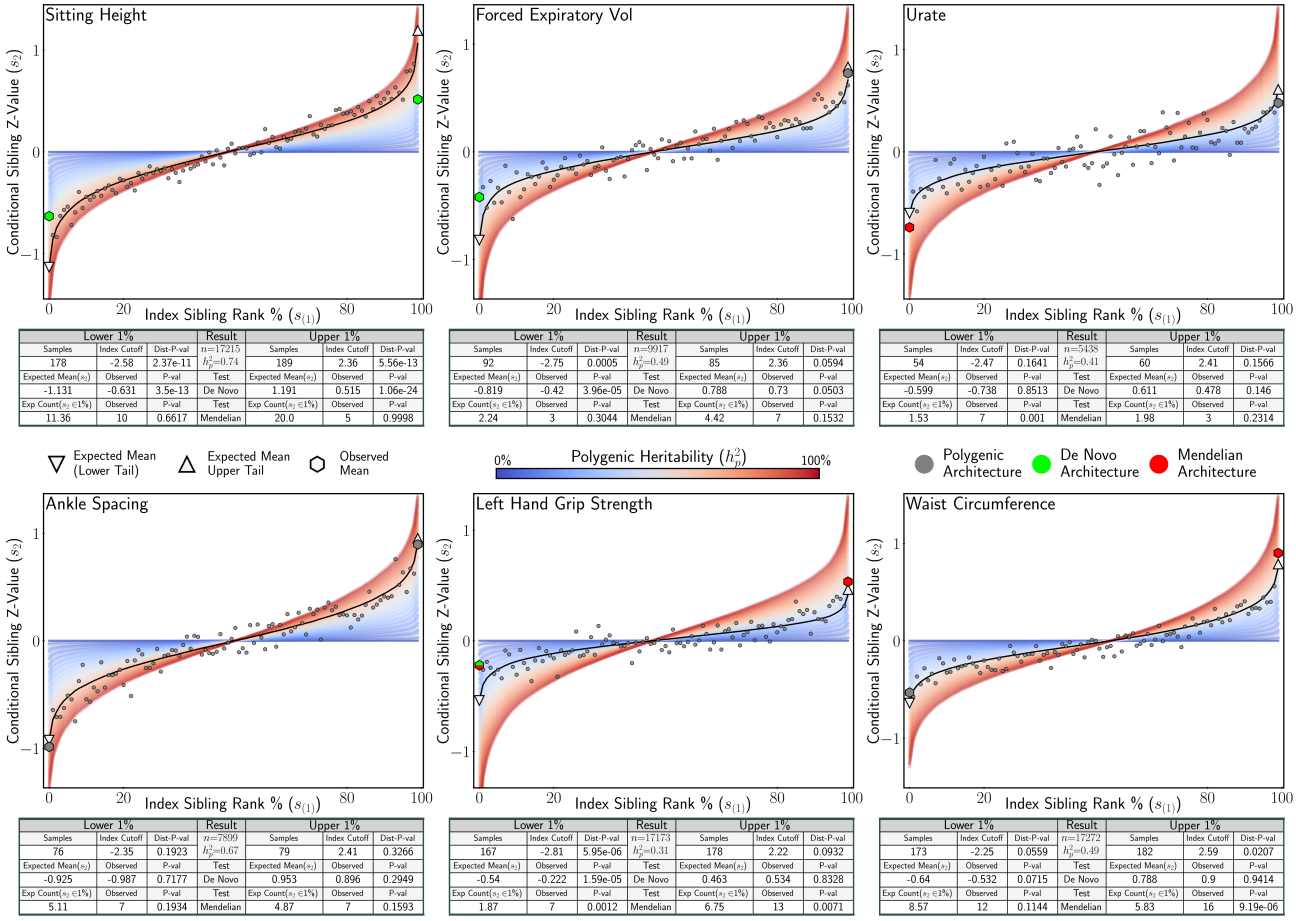


Fig. 5: Power to to detect complex tail architecture for different heritability levels, *de novo* and Mendelian frequencies and sample sizes. Simulation assumes highly penetrant *de novo* and Mendelian frequencies of 0.05%,0.1%,0.2%,0.3%, and 0.5%. The false-positive rate was set at 0.05. Null simulations (red dashed line) demonstrate tests are well calibrated.

In ***Appendix 2***, we provide detailed derivations for the statistical tests described at a high level in ***Model & Methods*** and explain how they identify tail signatures in contrast to a polygenic background where conditional siblings regress-to-the-mean at a rate proportional to $h^2/2$ (***Equation 4***). The performance

of the two tests evaluated using simulated sibling data is shown in ***Figure 5***. These tests demonstrate that power to identify *de novo* architecture is greatest when heritability is high, while power to identify Mendelian architecture is greatest when heritability is low. These patterns can be explained by the fact that high heritability should lead to relatively high similarity among siblings, and low heritability to low similarity, under polygenicity. When heritability is estimated near 50% and at least 0.1% of the population has high-impact rare aetiology, both tests are well-powered to identify each class of complex tail architecture.

Identifying Complex Tail Architecture in UK Biobank Data



Lower 1%			Result	Upper 1%		
Samples	Index Cutoff	Dist-P-val	n=17215	Samples	Index Cutoff	Dist-P-val
178	-2.58	2.37e-11	h_p^2=0.74	189	2.36	5.56e-13
Expected Mean(s_2)	Observed	P-val	Test	Expected Mean(s_2)	Observed	P-val
-1.131	-0.631	3.5e-13	De Novo	1.191	0.515	1.06e-24
Exp Count($s_2 \in 1\%$)	Observed	P-val	Test	Exp Count($s_2 \in 1\%$)	Observed	P-val
11.36	10	0.6617	Mendelian	20.0	5	0.9998

Lower 1%			Result	Upper 1%		
Samples	Index Cutoff	Dist-P-val	n=9917	Samples	Index Cutoff	Dist-P-val
92	-2.75	0.0005	h_p^2=0.49	85	2.36	0.0594
Expected Mean(s_2)	Observed	P-val	Test	Expected Mean(s_2)	Observed	P-val
-0.819	-0.42	3.96e-05	De Novo	0.788	0.73	0.0503
Exp Count($s_2 \in 1\%$)	Observed	P-val	Test	Exp Count($s_2 \in 1\%$)	Observed	P-val
2.24	3	0.3044	Mendelian	4.42	7	0.1532

Lower 1%			Result	Upper 1%		
Samples	Index Cutoff	Dist-P-val	n=5438	Samples	Index Cutoff	Dist-P-val
54	-2.47	0.1641	h_p^2=0.41	60	2.41	0.1566
Expected Mean(s_2)	Observed	P-val	Test	Expected Mean(s_2)	Observed	P-val
-0.599	-0.738	0.8513	De Novo	0.611	0.478	0.146
Exp Count($s_2 \in 1\%$)	Observed	P-val	Test	Exp Count($s_2 \in 1\%$)	Observed	P-val
1.53	7	0.001	Mendelian	1.98	3	0.2314

Lower 1%			Result	Upper 1%		
Samples	Index Cutoff	Dist-P-val	n=7899	Samples	Index Cutoff	Dist-P-val
76	-2.35	0.1923	h_p^2=0.67	79	2.41	0.3266
Expected Mean(s_2)	Observed	P-val	Test	Expected Mean(s_2)	Observed	P-val
-0.925	-0.987	0.7177	De Novo	0.953	0.896	0.2949
Exp Count($s_2 \in 1\%$)	Observed	P-val	Test	Exp Count($s_2 \in 1\%$)	Observed	P-val
5.11	7	0.1934	Mendelian	4.87	7	0.1593

Lower 1%			Result	Upper 1%		
Samples	Index Cutoff	Dist-P-val	n=171173	Samples	Index Cutoff	Dist-P-val
167	-2.81	5.95e-06	h_p^2=0.31	178	2.22	0.0932
Expected Mean(s_2)	Observed	P-val	Test	Expected Mean(s_2)	Observed	P-val
-0.54	-0.222	1.59e-05	De Novo	0.463	0.534	0.8328
Exp Count($s_2 \in 1\%$)	Observed	P-val	Test	Exp Count($s_2 \in 1\%$)	Observed	P-val
1.87	7	0.0012	Mendelian	6.75	13	0.0071

Lower 1%			Result	Upper 1%		
Samples	Index Cutoff	Dist-P-val	n=17272	Samples	Index Cutoff	Dist-P-val
173	-2.25	0.0559	h_p^2=0.49	182	2.59	0.0207
Expected Mean(s_2)	Observed	P-val	Test	Expected Mean(s_2)	Observed	P-val
-0.64	-0.532	0.0715	De Novo	0.788	0.9	0.9414
Exp Count($s_2 \in 1\%$)	Observed	P-val	Test	Exp Count($s_2 \in 1\%$)	Observed	P-val
8.57	12	0.1144	Mendelian	5.83	16	9.19e-06

Fig. 6: Analysis of Six UK Biobank Traits. Application of statistical tests for Mendelian and *de novo* tail architecture to sibling trait data of six UK Biobank traits. For each trait the conditional sibling mean is plotted under polygenicity (black line) for the heritability estimated estimated from the data. The red (high) and blue (low) bands represent the expected conditional sibling mean under polygenicity at different heritability values. Statistical tests for *de novo* architecture, Mendelian architecture, and general departure from polygenicity (Kolmogorov-Smirnov Test, Dist P-val), were applied to conditional siblings with index siblings in the upper and lower 1% of the distribution. Significant associations for the Mendelian and *de novo* tests are shown in red and green respectively. Tail architecture that is not distinct from polygenic expectation is denoted in grey.

We applied our statistical tests for complex tail architecture to sibling-pair data on eighteen traits from the UK Biobank (31). Here were present results from a set of six traits with varied tail architecture: *Sitting Height, Forced Expiratory Volume, Urate, Ankle Spacing, Left Hand Grip Strength, and Waist Circumference.* For each trait we estimated conditional sibling heritability via ***Equation 5***, and performed tests to identify *de novo* and Mendelian architecture in the lower and upper tails of the distribution of each trait. Additionally, we also performed a Kolmogorov-Smirnov test (29) to provide a general test for any departures from our null model. We observed expected polygenic architecture in both tails for

Ankle Spacing. We inferred Mendelian architecture in the lower tail for *Urate*, and the upper tail for *Wait Circumference* and *Left Hand Grip Strength*. *De novo* architecture was inferred in the lower tail for *Forced Expiratory Volume* and strongly inferred in both tails for *Sitting Height*, which is supported by evidence from deep sequencing studies indicating that rare variants play a substantial role in the genetic aetiology for this trait (32, 33). In the lower tail for *Left Hand Grip Strength* we infer the presence of both *De Novo* (greater than expected mean) and Mendelian (more concordant siblings than expected) architecture. We note that this could occur as a result of highly penetrant variants that are only shared among some siblings or perhaps because at the extremes, siblings with different handedness are unexpectedly divergent and siblings with matching handedness are unexpectedly concordant. Extended results for all eighteen traits analysed can be found in ***Appendix 4***.

Discussion

In this paper, we present a novel approach to infer the genetic architecture of continuous traits, specifically in the tails of the distributions, from sibling trait data alone. Our approach is based on a theoretical framework that we develop, which derives the expected trait distributions of siblings conditional on the trait value of an index-sibling and the trait heritability, assuming polygenicity, random mating and environmental factors that are either controlled for or have combined Gaussian effects. The key intuition underlying the approach is that departures from the expected *conditional-sibling trait distribution* in relation to index-siblings selected from the trait tails may be due to non-polygenic architecture in the tails.

We demonstrate the validity of our conditional-sibling analytical derivations through simulations and show that our tests for identifying *de novo* and Mendelian architecture in the tails of trait distributions are well-powered when large effect alleles are present in the population on the order of 1 out of 1000 individuals. Applying our test to a subset of well-powered traits in the UK Biobank, we find evidence of complex genetic architecture in at least one tail ($\alpha < 0.05$) in sixteen of eighteen traits and find *de novo* architecture occurring more frequently than Mendelian architecture (19 vs 6 of 36 total tails).

There are several areas in which our work could have short-term utility. Firstly, those individuals inferred as having rare variants of large effect could be followed up in multiple ways to gain individual-level insights. For example, they could undergo clinical genetic testing to identify potential pathogenic variants with effects beyond the examined trait, either in the form of diseases or disorders that the individual has already been diagnosed with or else that they have yet to present with but may be at high future risk for. Furthermore, investigation of their environmental risk profile may indicate an alternative - environmental - explanation for their extreme trait value (see below), rather than the rare genetic architecture inferred by our tests.

Our framework could also help to refine the design of sequencing studies for identifying rare variants of large effect. Such studies either sequence entire cohorts at relatively high cost (34) or else perform more targeted sequencing of individuals in the trait tails with the goal of optimising power per cost (35). However, even the latter approach is usually performed blind to evidence of enrichment of rare variant aetiology in the tails. Since our approach enables identification of individuals that may be most likely to harbour rare variants, then these individuals could be prioritised for (deep) sequencing. Moreover, our ability to distinguish between *de novo* and Mendelian architecture could influence the broad study design, with the former suggesting that a family trio design may be more effective than population sequencing, which may be favoured if Mendelian architecture is inferred. Furthermore, our approach could be applied as a screening step to prioritise those traits, and corresponding tails, most likely to harbour rare variant architecture. Finally, if sequence data have already been collected, either cohort-wide or using a more targeted design, then our approach could be utilised to increase the power of statistical methods for detecting rare variants by upweighting individuals most likely to harbour rare variants.

This study has several limitations. First and foremost, departures from the expected conditional-sibling trait distributions could be due to environmental risk factors, such as medication-use or work related exposures, rather than rare genetic architecture. Thus, while we believe that tail-specific deviations from polygenic expectation are interesting whether they arise primarily from genetic or environmental factors, we caution against over-interpretation of our results. Rejection of the null hypothesis from our tests should be considered only as indicating effects *consistent with* non-polygenic genetic architecture, alongside alternative explanations such as tail-specific unshared (*de novo*) or shared ("Mendelian") environmental risks. We suggest that further investigation of individuals' clinical, environmental and genetic profiles are required to achieve greater certainty about the causes of their extreme trait values. Nevertheless, given knowledge that rare variants of large effect contribute to complex trait architecture, we expect that traits for which we infer non-polygenic architecture, will, on average, be more enriched for rare architecture in the tail(s) than other traits. Secondly, our modelling assumes that environmental risk factors of siblings are independent of each other. If in fact shared environmental risk factors contribute significantly to trait similarity among siblings, then our heritability estimates will be upwardly biased. However, this would only impact our tests if the degree of shared environmental risk differed in the tails relative to the rest of the trait distribution. Moreover, a large meta-analysis of heritability estimates from twin studies (24) concluded that the contribution of shared environment among siblings (even twins) is insubstantial, and so we might expect this to have limited impact on results from our tests. Thirdly, our modelling assumes random mating and so the results from our tests in relation to traits that may be the subject of assortative mating should be considered with caution. Likewise, our modelling assumes additivity of genetic effects and so, while additivity is well-supported by much statistical genetics research (36, 37), results from our tests should be reconsidered for any traits with evidence for significant non-additive genetic effects.

Our approach not only provides a novel way of inferring genetic architecture (without genetic data) but can do so specifically in the tails of trait distributions, which are most likely to harbour complex genetic architecture, due to selection, and are a key focus in biomedical research given their enrichment for disease. This work could also have broader implications in quantitative genetics since we derive fundamental results about the relationship among family members' complex trait values. The conditional-sibling trait distribution provides a simple way of understanding the expected trait values of individuals according to their sibling's trait value, which could be used to answer questions of societal importance and inform future research. For example, it can be used to answer questions such as: as a consequence of genetics alone, how much overlap should there be in the traits of offspring of mid-parents at the 5th and 95th percentile and how does that contrast with what we observe in highly structured societies? Moreover, further development of the theory described here could lead to a range of other applications, for example, estimating levels of assortative mating, inferring historical selection pressures, and quantifying heritability in specific strata of the population.

Acknowledgments

We thank the participants in the UK Biobank and the scientists involved in the construction of this resource for making the sibling data used in this manuscript available. The work in this manuscript has been conducted using the UK Biobank Resource under application 18177 (Dr O'Reilly). We would also like to thank Dr. Avi Reichenberg for early discussions and Dr. Peter Visscher for highlighting key references related to the topic, and Dr. Shai Carmi for providing feedback on a draft version of the article.

References

1. Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.
2. Jay L Lush. Intra-sire correlations or regressions of offspring on dam as a method of estimating heritability of characteristics. *Journal of animal science*, 1940(1):293–301, 1940.
3. Saskia Selzam, Tom A McAdams, Jonathan RI Coleman, Susan Carnell, Paul F O'Reilly, Robert Plomin, and Clare H Llewellyn. Evidence for gene-environment correlation in child feeding: Links between common genetic variation for bmi in children and parental feeding practices. *PLoS genetics*, 14(11):e1007757, 2018.
4. Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U Eliasen, Yunxuan Jiang, Sridharan Raghavan, et al. A saturated map of common genetic variants associated with human height. *Nature*, 610(7933):704–712, 2022.
5. Margaret Sunitha Selvaraj, Xihao Li, Zilin Li, Akhil Pampana, David Y Zhang, Joseph Park, Stella Aslibekyan, Joshua C Bis, Jennifer A Brody, Brian E Cade, et al. Whole genome sequence analysis of blood lipid levels in> 66,000 individuals. *Nature communications*, 13(1):5995, 2022.
6. Tarjinder Singh, Timothy Poterba, David Curtis, Huda Akil, Mariam Al Eissa, Jack D Barchas, Nicholas Bass, Tim B Bigdeli, Gerome Breen, Evelyn J Bromet, et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*, 604(7906):509–516, 2022.
7. Conrad Iyegbe and Paul O'Reilly. Genetic origins of schizophrenia find common ground. *Nature*, 604, 04 2022. doi: 10.1038/d41586-022-00773-5.
8. E Koch, N Connally, N Baya, MP Reeve, M Daly, B Neale, ES Lander, A Bloemendal, and S Sunyaev. Genetic association data are broadly consistent with stabilizing selection shaping human common diseases and traits. *bioRxiv*, pages 2024–06, 2024.
9. Lawrence H Uricchio. Evolutionary perspectives on polygenic selection, missing heritability, and gwas. *Human genetics*, 139(1):5–21, 2020.
10. Lucia Corte, Lathan Liou, Paul F O'Reilly, and Judit García-González. Trumpet plots: visualizing the relationship between allele frequency and effect size in genetic association studies. *GigaByte*, 2023, 2023.
11. Yukihide Momozawa and Keijiro Mizukami. Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of human genetics*, 66(1):11–23, 2021.
12. Yingleong Chan, Oddgeir L Holmen, Andrew Dauber, Lars Vatten, Aki S Havulinna, Frank Skorpen, Kirsti Kvaløy, Kaisa Silander, Thutrang T Nguyen, Cristen Willer, et al. Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. *PLoS genetics*, 7(12):e1002439, 2011.
13. Nicholas G Shakeshaft, Maciej Trzaskowski, Andrew McMillan, Eva Krapohl, Michael A Simpson, Avi Reichenberg, Martin Cederlöf, Henrik Larsson, Paul Lichtenstein, and Robert Plomin. Thinking positively: the genetics of high intelligence. *Intelligence*, 48:123–132, 2015.
14. Abraham Reichenberg, Martin Cederlöf, Andrew McMillan, Maciej Trzaskowski, Ori Kapra, Eyal Fruchter, Karen Ginat, Michael Davidson, Mark Weiser, Henrik Larsson, et al. Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proceedings of the National Academy of sciences*, 113(4):1098–1103, 2016.
15. Yoon-Mi Hur, Jong Woo Kim, Kee Wha Chung, Joong Sik Shin, Hoe-Uk Jeong, and Emmanuel Auta. The nigerian twin and sibling registry. *Twin Research and Human Genetics*, 16(1):282–284, 2013.
16. sibarc: Software for inference of genetic architecture. http://www.sibArc.net, 2024. Accessed: 2024-03-30.
17. Douglas Scott Falconer and Trudy FC Mackay. *Quantitative genetics*. Longman London, UK, 1983.
18. Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Meiosis. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.
19. Toby Johnson and Nick Barton. Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1411–1425, 2005.
20. Nicholas H Barton, Alison M Etheridge, and Amandine Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical population biology*, 118:50–73, 2017.
21. Michael Bulmer. Galton's law of ancestral heredity. *Heredity*, 81(5):579–585, 1998.
22. Bart ML Baselmans, Loïc Yengo, Wouter van Rheenen, and Naomi R Wray. Risk in relatives, heritability, snp-based heritability, and genetic correlations in psychiatric disorders: a review. *Biological Psychiatry*, 89(1):11–19, 2021.
23. Todd Lencz, Daniel Backenroth, Einat Granot-Hershkovitz, Adam Green, Kyle Gettler, Judy H Cho, Omer Weissbrod, Or Zuk, and Shai Carmi. Utility of polygenic embryo screening for disease depends on the selection strategy. *Elife*, 10: e64716, 2021.
24. Tinca JC Polderman, Beben Benyamin, Christiaan A De Leeuw, Patrick F Sullivan, Arjen Van Bochoven, Peter M Visscher, and Danielle Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, 47(7):702–709, 2015.

25. Louis Lello, Maximus Hsu, Erik Widen, and Timothy G Raben. Sibling variation in polygenic traits and dna recombination mapping with uk biobank and ivf family data. *Scientific Reports*, 13(1):376, 2023.
26. Rosa Cheesman, Jonathan Coleman, Christopher Rayner, KL Purves, Genevieve Morneau-Vaillancourt, Kylie Glanville, Shing W Choi, Gerome Breen, and Thalia C Eley. Familial influences on neuroticism and education in the uk biobank. *Behavior genetics*, 50(2):84–93, 2020.
27. Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.
28. Guiyan Ni, Gerhard Moser, Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, et al. Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *The American Journal of Human Genetics*, 102(6):1185–1194, 2018.
29. Marco Marozzi. Nonparametric simultaneous tests for location and scale testing: a comparison of several methods. *Communications in Statistics-Simulation and Computation*, 42(6):1298–1317, 2013.
30. Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.
31. Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.
32. Catherine Tcheandjieu, Matthew Aguirre, Stefan Gustafsson, Priyanka Saha, Praneetha Potiny, Melissa Haendel, Erik Ingelsson, Manuel A Rivas, and James R Priest. A phenome-wide association study of 26 mendelian genes reveals phenotypic expressivity of common and rare variants within the general population. *PLoS genetics*, 16(11):e1008802, 2020.
33. Gyda Bjornsdottir, Lilja Stefansdottir, Gudmar Thorleifsson, Patrick Sulem, Kristjan Norland, Egil Ferkingstad, Asmundur Oddsson, Florian Zink, Sigrun H Lund, Muhammad S Nawaz, et al. Rare slc13a1 variants associate with intervertebral disc disorder highlighting role of sulfate in disc pathology. *Nature communications*, 13(1):1–13, 2022.
34. Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021.
35. Jinliang Yang, Haiying Jiang, Cheng-Ting Yeh, Jianming Yu, Jeffrey A Jeddeloh, Dan Nettleton, and Patrick S Schnable. Extreme-phenotype genome-wide association study (xp-gwas): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. *The Plant Journal*, 84(3):587–596, 2015.
36. Naomi R Wray, Cisca Wijmenga, Patrick F Sullivan, Jian Yang, and Peter M Visscher. Common disease is more complex than implied by the core gene omnigenic model. *Cell*, 173(7):1573–1580, 2018.
37. Valentin Hivert, Julia Sidorenko, Florian Rohart, Michael E Goddard, Jian Yang, Naomi R Wray, Loic Yengo, and Peter M Visscher. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *The American Journal of Human Genetics*, 108(5):786–798, 2021.
38. Nicholas H Barton and Peter D Keightley. Understanding quantitative genetic variation. *Nature Reviews Genetics*, 3(1): 11–21, 2002.
39. Nicholas H Barton. The "new synthesis". *Proceedings of the National Academy of Sciences*, 119(30):e2122147119, 2022.
40. Ole Barndorff-Nielsen, John Kent, and Michael Sørensen. Normal variance-mean mixtures and z distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 145–159, 1982.
41. Zachary R McCaw, Jacqueline M Lane, Richa Saxena, Susan Redline, and Xihong Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, 76(4):1262–1272, 2020.
42. J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 1994.
43. DS Falconer. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of human genetics*, 31(1):1–20, 1967.
44. Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons, 1995.
45. Ti Reich, JW James, and CA Morris. The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. *Annals of human genetics*, 36(2):163–184, 1972.
46. Douglas S Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1):51–76, 1965.
47. Benjamin Neale. Neale lab data: http://www.nealelab.is/uk-biobank/, 2018.

Appendix 1: Derivation of Conditional Sibling Distributions

Here we derive the distribution that describes the probability density of the "conditional" sibling (S_2) given the genetic liability, trait value and case-status of one or more index siblings. In each case we assume a population of unrelated parents and rely on the results from the *infinitesimal polygenic model* that show that within family variance is normally distributed around midparent genetic liability (average of parents) with half the ancestral trait variance even when selection, drift, population structure or dominance effects alter the between family trait distribution (20, 38, 39).

Case 1) Index liability known, continuous trait ($h^2 = 1$): $P(S_2 \mid S_1 = s_1)$

We begin with the simplest case, a polygenic normally distributed trait which is fully heritable ($h^2 = 1$) where the genetic liability of an index sibling in a population is known. Throughout we denote the midparent, index sibling and conditional sibling by M, S_1 and S_2. We begin by calculating the midparent distribution conditional on index liability using Bayes theorem:

$$\begin{aligned} p(M = m \mid S = s) &\propto p(S = s \mid M = m)p(M = m) \\ &= \mathcal{N}\left(S \mid m, \frac{\sigma^2}{2}\right)\mathcal{N}(M \mid 0, \sigma^2/2) \\ &\propto \exp\left\{-\frac{(s-m)^2}{\sigma^2}\right\}\exp\left\{-\frac{m^2}{\sigma^2}\right\} \\ &= \mathcal{N}\left(M \mid s/2, \sigma^2/4\right) \end{aligned} \tag{10}$$

Then using the following identity (40):

$$\int \mathcal{N}(x \mid \alpha z, \sigma_1^2)\mathcal{N}(z \mid \mu, \sigma_2^2)\, dz = \mathcal{N}(x \mid \alpha\mu, \sigma_1^2 + \alpha^2\sigma_2^2) \tag{11}$$

We calculate the conditional sibling distribution similarly:

$$\begin{aligned} p(S_2 \mid S_1 = s_1) &= \int p(S_2, M = m \mid S_1 = s_1)\, dm \\ &= \int p(S_2 \mid M = m, S_1 = s_1)p(M = m \mid S_1 = s_1)\, dm \\ &= \int p(S_2 \mid M = m)p(M = m \mid S_1 = s_1)\, dm \\ &= \int \mathcal{N}\left(S_2 \mid m, \frac{\sigma^2}{2}\right)\mathcal{N}\left(M = m \mid \frac{s_1}{2}, \frac{\sigma^2}{4}\right) dm \\ p(S_2 \mid S_1 = s_1) &= \mathcal{N}\left(S_2 \mid \frac{s_1}{2}, \frac{3\sigma^2}{4}\right) \end{aligned} \tag{12}$$

Thus, as predicted by the *infinitesimal polygenic model* the conditional sibling liability is normally distributed around the midparent liability distribution with additional variance equal to half the population liability variance.

Case 2) Index trait value known, continuous trait ($h^2 \neq 1$): $P(S_2 \mid S_1 = s_1)$

In this case, the primary result considered in this manuscript, a trait z-value, or equivalently, the percentile rank of an index sibling in genome wide association where the rank-based inverse transformation has been applied (41) is known. Transformation to a Z distribution ($\sigma^2 = 1$) means that for heritability h^2 the genetic liability and environmental contributions to trait variance are are $\sigma_g^2 = h^2$ and $\sigma_e^2 = 1 - h^2$, respectively. Similar to the previous case we begin by calculating the conditional mid-parent liability from Bayes' theorem:

$$
\begin{aligned}
p(M_g = m_g \mid S_1 = s_1) &\propto p(S_1 = s_1 \mid M_g = m_g) p(M_g = m_g) \\
&= \mathcal{N}\left(S_1 = s_1 \mid m_g, 1 - \frac{h^2}{2}\right) \mathcal{N}(M_g = m_g \mid 0, h^2/2) \\
&\propto \exp\left\{-\frac{(s_1 - m_g)^2}{2 - h^2}\right\} \exp\left\{-\frac{m_g^2}{h^2}\right\} \\
&\propto \exp\left\{-\frac{2m_g^2 - 2m_g s_1 h^2}{2h^2(1 - h^2/2)}\right\} \\
&\propto \exp\left\{-\frac{(m_g - s_1 h^2/2)^2}{h^2(1 - h^2/2)}\right\} \\
&= \mathcal{N}\left(M_g = m_g \mid \frac{1}{2} s_1 h^2, \frac{1}{2} h^2 (1 - h^2/2)\right) \qquad \textbf{(13)}
\end{aligned}
$$

Then, we again use ***Equation 11*** to derive the distribution conditional sibling distribution:

$$
\begin{aligned}
p(S_2 = s_s \mid S_1 = s_1) &= \int p(S_2 = s_2 \mid M_g = m_g) p(M_g = m_g \mid S_1 = s_1)\, dm_g \\
&= \int \mathcal{N}\left(S_2 \mid m_g, 1 - \frac{h^2}{2}\right) \mathcal{N}\left(M_g \mid \frac{1}{2} s_1 h^2, \frac{1}{2} h^2 \left(1 - \frac{h^2}{2}\right)\right) dm_g \\
&= \mathcal{N}\left(S_2 = s_2 \mid \frac{1}{2} s_1 h^2, 1 - \frac{h^4}{4}\right) \qquad \textbf{(14)}
\end{aligned}
$$

Case 3) Multiple index trait values known, continuous trait: $P(S_3 \mid S_1 = s_1, S_2 = s_2)$
The conditional sibling distribution can also be derived using the joint trait distribution for related individuals:

$$y \sim \mathcal{N}(0, \mathbf{G})$$

where the covariance $\mathbf{G}$ is the genetic relationship matrix. Thus for a sibling pair

$$p(s_1, s_2) = \mathcal{N}_2\left(0, \begin{pmatrix} 1 & h^2/2 \\ h^2/2 & 1 \end{pmatrix}\right)$$

As shown by Bernardo and Smith (42), if X is multivariate normal $\mathcal{N}(\mu, \lambda^{-1})$, where $\lambda = \Sigma^{-1}$ is the precision matrix, and X is partitioned into x_1 and x_2, with corresponding partitions of μ and λ of:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$$

then the conditional distribution of x_1 given x_2 is also normal with mean and precision matrix:

$$\mu_1 - \lambda_{11}^{-1}\lambda_{12}(x_2 - \mu_2), \quad \lambda_{11} \tag{15}$$

Thus, given the joint distribution for three siblings:

$$p(s_1, s_2, s_3) = \mathcal{N}_3\left(0, \begin{pmatrix} 1 & h^2/2 & h^2/2 \\ h^2/2 & 1 & h^2/2 \\ h^2/2 & h^2/2 & 1 \end{pmatrix}\right)$$

The precision matrix $\lambda = \Sigma^{-1}$ is

$$\lambda = \frac{1}{2 + h^2 - h^4}\begin{pmatrix} 2+h^2 & -h^2 & -h^2 \\ -h^2 & 2+h^2 & -h^2 \\ -h^2 & -h^2 & 2+h^2 \end{pmatrix}$$

And we can calculate the conditional distribution for for two sibling using ***Equation 15***

$$p(S_3 \mid S_1 = s_1, S_2 = s_2) = \mathcal{N}\left(S_3 \mid \frac{h^2}{2+h^2}(s_1 + s_2), 1 - \frac{h^4}{2+h^2}\right)$$

Case 4) Binary trait: $P(S_2 \mid S_1 = \textit{Affected})$
Here we again assume an underlying distribution that is $\mathcal{N}(0,1)$ and made up of genetic and environmental components. However, we only know the index sibling's *status*, which, as described under the liability threshold model (43), is equivalent to conditioning on the event where than index sibling's trait value is above or below a z-value threshold T:

$$p(S_2 \mid S_1{=}\text{Affected}) = p(S_2 \mid S_1 > T) \quad \text{and} \quad p(S_2 \mid S_1{=}\cancel{\text{Affected}}) = p(S_2 \mid S_1 < T) \tag{16}$$

where $T = \Phi^{-1}(1-K)$, Φ^{-1} is the inverse normal cumulative distribution function, and K is the incidence of the binary trait in the population. Thus, the conditional distribution one sibling given an *Affected* index sibling can be can be calculated integrating over the normal distribution truncated at T:

$$p(S_2 \mid S_1 > T) = \int_T^{\infty} p(S_2 \mid S_1 = s_1) p(S_1 = s_1) ds_1 \tag{17}$$

The first two moments of the this truncated normal (44) are:

$$\begin{aligned}
\mathrm{E}(S_1 \mid S_1 > T) &= \frac{\phi(T)}{1-\Phi(T)} = \frac{\phi(T)}{K} = i \\
\mathrm{V}(S_1 \mid S_1 > T) &= 1 + T\frac{\phi(T)}{K} - \left(\frac{\phi(T)}{K}\right)^2 = 1 + iT - i^2
\end{aligned}$$

Approximating the index sibling distribution using a normal whose moments are taken from this truncated distribution, ***Equation 17*** becomes:

$$p(S_2 = s_s \mid S_1 > T) = \int \mathcal{N}\left(S_2 = s_2 \mid \frac{1}{2}s_1^* h^2, 1 - \frac{h^4}{4}\right) \mathcal{N}(S_1 = s_1^* \mid i, 1 + iT - i^2) ds_1^* \tag{18}$$

which can be solved using the identity given in ***Equation 11***:

$$\begin{aligned}
&= \mathcal{N}\left(S_2 = s_2 \mid \frac{1}{2}ih^2, 1 - \frac{h^4}{4} + \frac{1}{4}h^4(1 + iT - i^2)\right) \\
&= \mathcal{N}\left(S_2 = s_2 \mid \frac{1}{2}ih^2, 1 - \frac{h^4}{4}i(i-T)\right)
\end{aligned} \tag{19}$$

Thus, conditional on an *Affected* sibling, the probability of concordance is:

$$\begin{aligned}
p(S_2 = \text{Affected} \mid S_1{=}\text{Affected}) &= p(S_2 > T \mid S_1 > T) \\
&= 1 - \Phi\left(\frac{T - ih^2/2}{\sqrt{1 - h^4 i(i-T)/4}}\right)
\end{aligned}$$

which is equivalent to Reich's (45) correction to Falconer's (46) approximation where the relationship between the relatives is 0.5 for siblings. The probability of discordance given an *Unaffected* sibling can be calculated from Bayes' Theorem:

$$\begin{aligned}
p(S_2 = \text{Affected} \mid S_1{=}\cancel{\text{Affected}}) &= \frac{p(S_1{=}\cancel{\text{Affected}} \mid S_2 = \text{Affected}) p(S_2 = \text{Affected})}{p(S_1{=}\cancel{\text{Affected}})} \\
&= \frac{K - Kp(S_2{=}\text{Affected} \mid S_1{=}\text{Affected})}{1-K}
\end{aligned}$$

which allow the conditional probability of case status to be determined given a index sibling's status.

Appendix 2: Statistical Tests for Complex Architecture

Here we describe our statistical tests for complex tail architecture. Our tests identify changes in the conditional sibling distribution when ascertaining on an index sibling in the tail relative to polygenic expectation. To carry out these tests we establish a null distribution built on the assumption that indexing on siblings not in the tails reduces that likelihood that either sibling phenotype in the pair is driven by rare variants of large effect. We use the region from the 5th to the 95th percentile to estimate heritability. From the n sibling pairs where the index sibling $s_1^{(i)}$ is in the 5th to the 95th, we calculate the conditional likelihood ***Equation 14***:

$$\mathcal{L}(D|h^2) = \prod_i^n p(s_2^{(i)}|s_1^{(i)}, h^2) \propto \left(1 - \frac{h^4}{4}\right)^{-n/2} \exp\left(-\frac{\sum_i^n \left(s_2^{(i)} - \frac{s_1^{(i)} h^2}{2}\right)^2}{2\left(1 - \frac{h^4}{4}\right)}\right) \tag{20}$$

and maximize the log-likelihood with respect to h^2:

$$\log \mathcal{L} \propto -n \log\left(1 - \frac{h^4}{4}\right) - \frac{1}{1 - \frac{h^4}{4}} \sum_{i=1}^n \left(s_2^{(i)} - \frac{s_1^{(i)} h^2}{2}\right)^2 \tag{21}$$

to obtain an maximum likelihood estimate for h^2 that is used to define the null distribution for our statistical tests.

Statistical Test for De Novo Architecture

We identify *de novo* mutations of large effect by testing for discordance between siblings relative to the polygenic null using the conditional distribution of a sibling given index sibling from (***Equation 14***). Since *de novo* mutations typically result in trait values in the tail of the distribution the test conditions on those index siblings in a specified upper quantile q of the distribution, i.e. those sibling pairs $(s_1^{(i)}, s_2^{(i)})$ such that $s_1^{(i)} > \Phi(q)$, defined as the set A_q. We introduce an additional parameter α where values of $\alpha < 0$ in the right tail and $\alpha > 0$ in the left tail are indicative of discordant siblings with trait values closer to the mean, giving a log-likelihood:

$$\log \mathcal{L} = -\frac{1}{2(1 - \frac{h^4}{4})} \sum_{i \in A_q} \left(s_2^{(i)} - \left(\frac{1}{2} s_1^{(i)} h^2 + \alpha\right)\right)^2 \tag{22}$$

The null hypothesis $H_0 : \alpha = 0$ is tested via a score test:

$$U = \frac{d}{d\alpha} \log \mathcal{L} = \frac{1}{1 - \frac{h^4}{4}} \sum_{i \in A} s_2^{(i)} - \left(\frac{1}{2} s_1^{(i)} h^2 + \alpha\right) \tag{23}$$

$$I = \frac{d^2}{d\alpha^2} \log \mathcal{L} = -\frac{n}{1 - \frac{h^4}{4}} \tag{24}$$

And the score test for $H_0 : \alpha = 0$:

$$z = \frac{U}{\sqrt{-I}} = \frac{\sum_{i \in A_q} s_2^{(i)} - \frac{1}{2} s_1^{(i)} h^2}{\sqrt{n(1 - \frac{h^4}{4})}} \tag{25}$$

Statistical Test for Mendelian Architecture

Here we test for excess concordance between siblings in the tails of the distribution by testing for an excess number of observed siblings in the tail $S_2 > \Phi^{-1}(q)$ given the index sibling is in the tail $S_1 > \Phi^{-1}(q)$, where q is the quantile of interest. Denoting the set of index siblings in the tail by A_q and the size of the set by n, under the null of pologenicity we calculate the probability that the conditional sibling exceeds $\Phi^{-1}(q)$ from the normal cdf and compute the mean to define mean concordance under polygenicity π_0:

$$\begin{aligned}\pi_0 &= \frac{1}{n}\sum_{i\in A_q} P\left(S_2^{(i)} > \Phi^{-1}(q) \mid S_1^{(i)} = s_1^{(i)}\right)\\ &= \frac{1}{n}\sum_{i\in A_q}\left(1-\Phi\left(\frac{\Phi^{-1}(q)-\frac{s_1^{(i)}h^2}{2}}{\sqrt{1-\frac{h^4}{4}}}\right)\right)\end{aligned}$$

Denoting the observed concordance (number of sibling pairs both $> \Phi^{-1}(q)$) by r, the binomial log-likelihood (ignoring the constant) is:

$$\begin{aligned}\log\mathcal{L} &= r\log\pi + (n-r)\log(1-\pi)\\ U = \frac{d}{d\pi}\log L &= \frac{r}{\pi} - \frac{n-r}{1-\pi}\\ &= \frac{r(1-\pi)-(n-r)\pi}{\pi(1-\pi)}\\ &= \frac{r-n\pi}{\pi(1-\pi)}\\ I = \frac{d^2}{d\pi^2}\log L &= -\frac{r}{\pi^2} - \frac{n-r}{(1-\pi)^2}\\ &= -\frac{r(1-\pi)^2+(n-r)\pi^2}{\pi^2(1-\pi)^2}\end{aligned}$$

Assuming $r = n\pi$ such that I is not a function of any particular observation:

$$= -\frac{n}{\pi(1-\pi)}$$

And the score test for $H_0 : \pi = \pi_0$:

$$z = \frac{U}{\sqrt{-I}} = \frac{r-n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

Statistical Test for Non-Polygenic Architecture

General departures from polygenicity can be identified based on the degree of discordance in conditional sibling distribution from the expectation. Assuming polygenicity, given index and conditional sibling pairs $(s_1^{(i)}, s_2^{(i)})$, from ***Equation 4*** we can write

$$z_2^{(i)} = \frac{(s_2^{(i)} - \frac{1}{2}s_1^{(i)}h^2)}{\sqrt{1-\frac{h^4}{4}}} \sim \mathcal{N}(0,1)$$

Therefore, departures in polygenicity can be tested in trait quantiles by testing the observed distribution of $z_2^{(i)}, i \in A_q$, where A_q is the set of index siblings $s_1^{(i)}$ in the quantile of interest, relative to a standard normal distribution via the Kolmogorov-Smirnov test.

Appendix 3: Model Evaluation

Here we compare our theoretical derivations (***Equation 4***) that rely on the *infinitesimal polygenic model* (20) with simulated offspring data (**Figure 7**). We also compare our model to our empirical simulation (see ***Model & Methods***) that draws allele frequencies and effect sizes from publicly available GWAS data (47) for two traits to produce parent and offspring genotype and genetic liability (equivalent to trait value when $h^2 = 1$) (**Figure 8**).

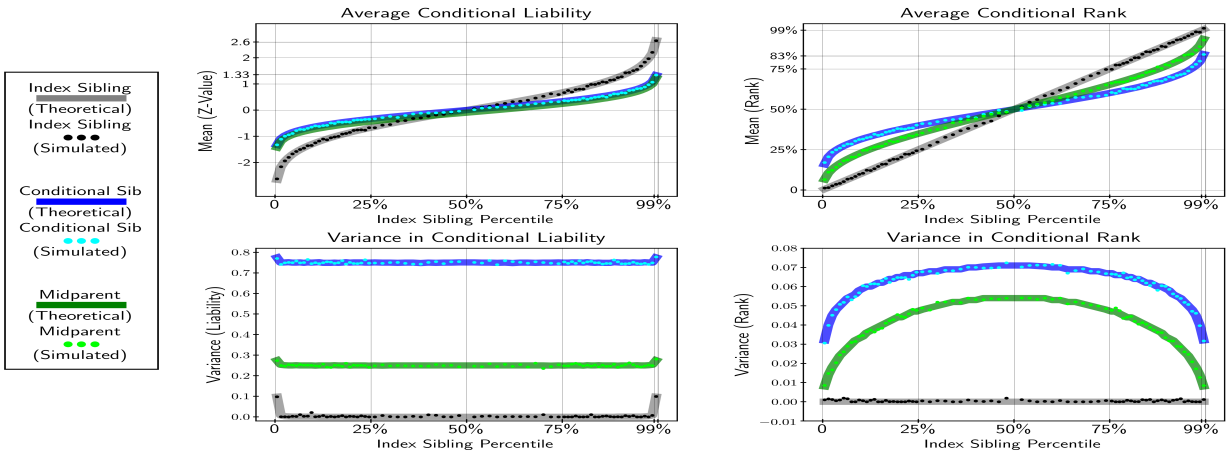


Fig. 7: Theoretical and simulated conditional expectation and variance in liability (z-score) and rank across index sibling percentiles for conditional sibling, midparents and index siblings. Simulation drew one-thousand parent liability values from $\mathcal{N}(0,1)$, these were randomly paired to produce to midparents with liability m_i, two offspring were subsequently drawn from ($\mathcal{N}(m_i, \frac{1}{2})$) and randomly assigned as index and conditional siblings.

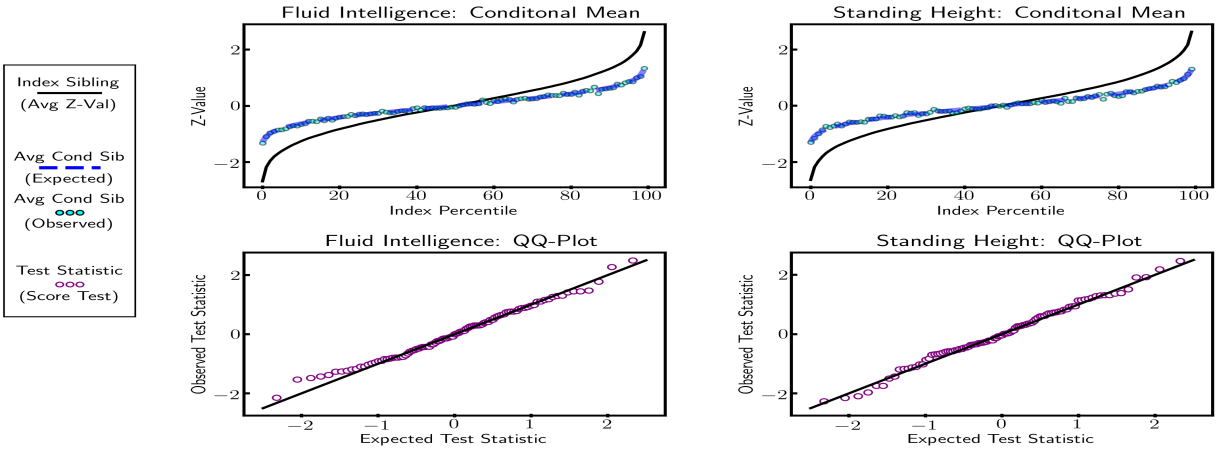


Fig. 8: For both *Fluid Intelligence* and *Standing Height* GWAS variants (on chromosome one) were used to simulate parent and offspring genotypes and liability values. Plots show that for both traits the offspring distribution is normal and that the sibling distribution is multivariate normal, in line with our theoretical prediction

These tests demonstrate that our theoretical framework accurately reflects an additive polygenic trait in an outcrossing population. Additionally these results demonstrate that deviations in the conditional sibling distribution can be interpreted as non-polygenic architecture or quantile specific environmental effects.

Appendix 4: Software Availability and Extended Results

We have made our code, sample data, and a brief tutorial available online at www.sibArc.net (16). Below we display the summary statistics and the tail results for all eighteen UK Biobank traits analyzed and referenced in the manuscript.

Trait Summary				Tail Architecture									
				Tail Summary				De-novo Test			Mendelian Test		
Trait Name (UKB-id)	Totals: Sib-Pairs, Unique Vals	Dist: Skew, Kurtosis	snp-h^2 (gwas), sib-h^2 (full), sib-h^2 (5-95)	Tail	samples	index-sib cutoff	KS-pv	Expected Mean(s_2)	Observed Mean	P-val	Exp Count ($s_2 \in 1\%$)	Observed Count	P-val
Left Hand Grip Strength (46)	17,174, 85	0.369, 2.731	0.1, 0.32, 0.36	Upper	178	2.217	0.0932	0.463	0.534	0.8328	6.75	13	0.0071
				Lower	167	-2.806	5.95e-06	-0.54	-0.222	1.59e-05	1.87	7	0.0012
Waist Circumference (48)	17,273, 716	0.423, 3.294	0.16, 0.49, 0.52	Upper	182	2.589	0.0207	0.788	0.9	0.9414	5.83	16	9.19e-06
				Lower	173	-2.245	0.0559	-0.64	-0.532	0.0715	8.57	12	0.1144
Ankle Spacing (3143)	7,900, 1,928	0.251, 3.112	0.24, 0.67, 0.69	Upper	79	2.406	0.3266	0.953	0.896	0.2949	4.87	7	0.1593
				Lower	76	-2.355	0.1923	-0.925	-0.987	0.7177	5.11	7	0.1934
Sitting Height (20015)	17,216, 352	0.041, 3.4	0.29, 0.73, 0.82	Upper	189	2.362	5.56e-13	1.191	0.515	1.06e-24	20.0	5	0.9998
				Lower	178	-2.581	2.37e-11	-1.131	-0.631	3.5e-13	11.36	10	0.6617
Forced Expiratory Vol (20150)	9,918, 662	0.402, 3.22	0.17, 0.49, 0.58	Upper	85	2.359	0.0594	0.788	0.73	0.0503	4.42	7	0.1532
				Lower	92	-2.747	0.0005	-0.819	-0.42	3.96e-05	2.24	3	0.3044
Body Fat Percent (23099)	16,785, 577	0.094, 2.562	0.18, 0.53, 0.57	Upper	165	2.329	0.5765	0.794	0.88	0.8754	9.25	17	0.0044
				Lower	172	-2.554	0.004	-0.846	-0.646	0.0032	6.64	9	0.1748
Whole Body Impedance (23106)	16,801, 771	0.239, 2.796	0.21, 0.57, 0.59	Upper	172	2.4	0.0565	0.833	0.683	0.0198	8.93	7	0.7465
				Lower	164	-2.462	0.012	-0.866	-0.714	0.0209	8.18	10	0.2566
Right Leg Impedance (23107)	16,793, 350	0.152, 3.342	0.2, 0.55, 0.58	Upper	175	2.315	0.7491	0.783	0.763	0.3919	9.89	9	0.6143
				Lower	163	-2.5	0.0003	-0.853	-0.595	0.0003	7.17	6	0.672
Left Leg Impedance (23108)	16,793, 352	0.144, 3.373	0.2, 0.54, 0.56	Upper	172	2.298	0.1393	0.737	0.829	0.8965	9.07	10	0.3752
				Lower	159	-2.502	0.0093	-0.81	-0.65	0.0181	6.35	5	0.7078
Right Arm Impedance (23109)	16,983, 475	0.329, 2.627	0.19, 0.51, 0.53	Upper	169	2.506	0.001	0.806	0.581	0.0012	6.91	3	0.9356
				Lower	168	-2.369	0.4972	-0.771	-0.721	0.252	8.7	10	0.3251
Left Arm Impedance (23110)	16,797, 455	0.337, 2.616	0.19, 0.51, 0.53	Upper	162	2.537	0.0023	0.805	0.663	0.0306	6.1	7	0.3548
				Lower	171	-2.399	0.599	-0.757	-0.688	0.1767	7.87	10	0.2183
Trunk Fat Percentage (23127)	16,780, 644	-0.083, 3.188	0.17, 0.51, 0.52	Upper	162	2.213	0.5881	0.658	0.683	0.6269	8.78	11	0.2204
				Lower	162	-2.512	0.1247	-0.739	-0.645	0.1056	5.47	9	0.0624
Red Blood Cell Count (30010)	16,307, 3,421	0.107, 3.308	0.17, 0.48, 0.53	Upper	177	2.295	0.0892	0.716	0.575	0.0256	9.16	8	0.653
				Lower	163	-2.49	1.58e-09	-0.817	-0.338	1.07e-10	7.08	4	0.8815
Haemoglobin Concentration (30020)	16,307, 1,119	-0.066, 3.361	0.17, 0.39, 0.44	Upper	168	2.282	0.0217	0.603	0.49	0.0343	7.24	4	0.8908
				Lower	159	-2.59	4.34e-06	-0.71	-0.315	1.67e-07	4.41	5	0.3878
Haematocrit Percentage (30030)	16,303, 2,872	-0.018, 3.35	0.13, 0.37, 0.41	Upper	167	2.289	0.0104	0.548	0.343	0.0034	6.37	3	0.9135
				Lower	158	-2.524	0.0004	-0.631	-0.352	0.0002	4.32	4	0.5622
Cholesterol (30690)	5,452, 7,028	0.369, 3.44	0.19, 0.47, 0.53	Upper	60	2.281	0.4059	0.707	0.573	0.1425	3.13	6	0.0479
				Lower	65	-2.355	0.0045	-0.736	-0.371	0.0012	3.08	5	0.1316
LDL Direct (30780)	5,432, 5,475	0.367, 3.356	0.17, 0.45, 0.5	Upper	67	2.298	0.2262	0.655	0.372	0.0083	3.04	3	0.5098
				Lower	60	-2.351	0.0112	-0.68	-0.514	0.0918	2.58	5	0.0614
Urate (30880)	5,439, 5,035	0.46, 3.158	0.2, 0.42, 0.42	Upper	60	2.413	0.1566	0.611	0.478	0.146	1.98	3	0.2314
				Lower	54	-2.472	0.1641	-0.599	-0.738	0.8513	1.53	7	0.001

Fig. 9: Application To UKB (Extended).