

Genetics and Genomics

Sibling Similarity Can Reveal Key Insights Into Genetic Architecture

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted

June 20, 2023 (this version)

Sent for peer review

March 10, 2023

Posted to bioRxiv

February 21, 2023

Tade Souzaiaia ✉, Hei Man Wu, Clive Hoggart, Paul O'Reilly ✉

Department of Cellular Biology, Suny Downstate Health Sciences, Brooklyn, NY, USA • Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, 1 Gustave L Levy Pl, NY, NY, USA

 (https://en.wikipedia.org/wiki/Open_access)

 (<https://creativecommons.org/licenses/by/4.0/>)

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 identification of genetic architecture, specifically that in individuals with extreme trait values (e.g. in the top 1%). Establishing genetic architecture in these individuals is important because they are at greatest risk of disease and are most likely to harbour rare variants of large effect due to natural selection. We develop a theoretical framework that derives expected trait distributions of siblings based on an index sibling's trait value and trait heritability. This framework is used to develop statistical tests that can infer complex genetic architecture in trait tails, distinguishing between polygenic, *de novo* and Mendelian tail architecture. We apply our tests to UK Biobank data here, while they can be used to infer genetic architecture in any cohort or health registry that includes siblings, without requiring genetic data. We describe how our approach has the potential to help disentangle the genetic and environmental causes of extreme trait values, to identify individuals likely to carry pathogenic variants for follow-up clinical genetic testing, and to improve the design and power of future sequencing studies to detect rare variants.

eLife assessment

The authors present **valuable** findings on how to determine the genetic architecture of extreme phenotype values by using data on sibling pairs. While the authors' derivations of the method are correct, the scenarios considered are **incomplete**, making it difficult to have confidence in the interpretation of the results as demonstrating the influence of *de-novo* or Mendelian (rare, penetrant-variant) architectures. The method shows nevertheless promise and will be of interest to researchers studying complex trait genetics.

Introduction

The fields of quantitative genetics and genetic epidemiology have exploited the shared genetics and environment of siblings in a range of applications, notably to estimate

heritability, using theory first developed a century ago (1, 2), and more recently to infer a so-called ‘household effect’ (3), which contributes to genetic risk indirectly via a correlation between genetics and the household environment. Here we leverage trait information on siblings to infer the genetic architecture of those traits, not only at the population-level, but also in relation to whether the genetic liability of each individual - specifically those with extreme trait values - is a result of a few large effect alleles or many small effect alleles.

The genetic architecture of a complex trait is typically inferred from the findings of multiple studies: with 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 will be subject to negative selection and may be enriched for large effect rare variants (10), while 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), but can also produce dissimilarity between siblings beyond what is expected under polygenecity. 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). Findings from such studies such as *Reichenberg et al. 2016* (14) are consistent with the presence of *de novo* alleles of large effect in the trait tails, alongside alternative explanations such as specific environmental exposures. However, no theoretical framework has been developed to formally infer genetic architecture from sibling trait data.

We introduce a theoretical framework that allows widely available sibling trait data in population cohorts and health registries to be leveraged to perform statistical tests that estimate complex genetic architecture in the tails of the trait distribution. These tests can differentiate between polygenic, *de novo* alleles and rare variants of large effect (hereafter ‘Mendelian variants’ for shorthand). 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 polygenic and non-polygenic tail trait architecture (see [Figure 1](#)).

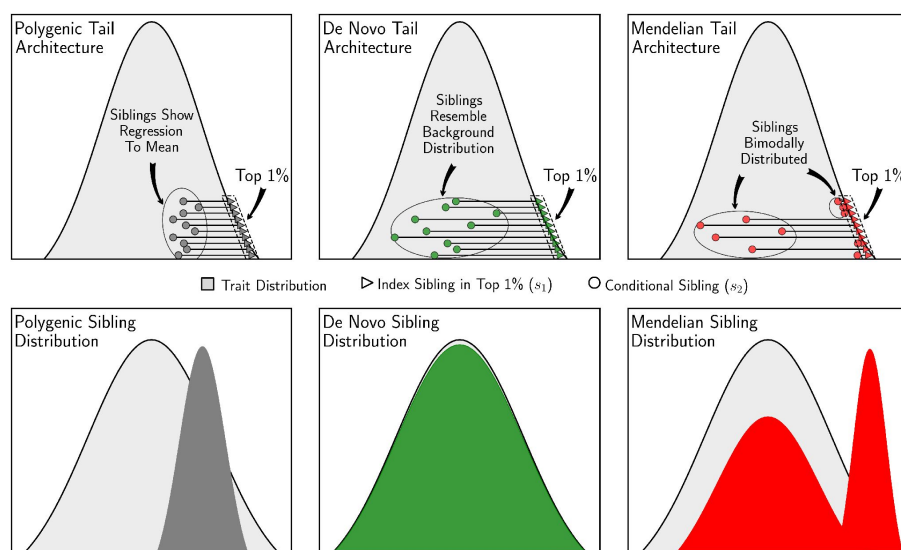


Fig. 1:

Sibling Similarity Under Different Tail Architectures.

From left to right: When individuals have extreme (top 1%) tail values as a result of many common alleles of small effect (polygenecity), then their siblings’ trait values are expected to show regression-to-the-mean (grey). When individuals have extreme trait values as a result of *de novo* mutations of large effect, then their siblings are expected to have trait values that correspond to the back-

ground distribution. When individuals have extreme trait values as a result of a rare variant of large effect inherited from

one of their parents (Mendelian), then their siblings are expected to have either similarly extreme trait values or trait values that are drawn from the background distribution, depending on whether or not they inherited the same large effect allele.

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 sibling registries (15) without the need for genetic data (see **Model & Methods**). We validate the statistical power of our tests using simulated data and real data from the UK Biobank. Our novel framework can be extended to applications such as estimating heritability, inferring assortative mating, and characterising historical selection pressures.

Model & Methods

Here we outline a framework that models the *conditional-sibling trait distribution* which describes an individual’s trait distribution conditioned on one or more *index* sibling(s). In this section we describe our derivation of the distribution for a completely polygenic trait, outline development of statistical tests that detect complex genetic architecture via departure polygenicity, and describe the simulation scheme used to validate and benchmark our results. Full derivations that complement this high-level summary of our methodology are included in the **Appendices**.

Conditional-Sibling Inference Framework

Assuming completely polygenic architecture, siblings of individuals with extreme trait values are expected to be much less extreme. This regression-to-the-mean can be understood through the two factors (16, 17) that determine inherited genetic liability: **(i)** average genetic liability of the parents, known as “*midparent*” liability, and, **(ii)** random genetic reassortment occurring during meiosis. Assuming that an individual presenting an upper-tail trait value does so due to *both* high midparent liability *and* genetic reassortment favouring a higher value, then a sibling sharing common midparent liability - but subject to independent reassortment - is likely less extreme. How much less extreme can be derived by first considering the conditional distribution that relates midparents and their offspring. In the simplest case, for a completely heritable continuous polygenic trait in a large randomly mating population (18), the offspring trait value, s , is normally distributed around the midparent trait value, m , as follows:

$$p(s | m) \sim \mathcal{N}(m, \sigma^2/2) \quad (1)$$

where σ^2 is the population trait variance. Note that the trait variance within families is half that of the population trait variance under neutrality and random mating, which is a key property in quantitative genetics (19–21). 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 midparent genetic liability, m_g , is:

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

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

$$p(m_g | s) \sim \mathcal{N}\left(\frac{1}{2}sh^2, \frac{h^2}{2}\left(1 - \frac{h^2}{2}\right)\right) \quad (3)$$

From [Equations 2 & 3](#), the sibling trait distribution conditional on an index sibling can be calculated:

$$\begin{aligned} p(s_2 | s_1) &\sim \int p(s_2 | m_g)p(m_g | s_1) dm_g \\ &\sim \mathcal{N}\left(\frac{1}{2}s_1h^2, 1 - \frac{h^4}{4}\right) \end{aligned} \quad (4)$$

Here we note the relative simplicity of this distribution, which means, for example, 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 [Results \(Figure 4\)](#) we illustrate how trait heritability and index quantile determine the conditional-sibling distribution in standard trait space and in population percentile space. In [Appendix 1](#) we provide a full derivation of this distribution ([Equation 4](#)) and generalize our analytical results across a range of scenarios, including binary phenotypes and multiple index siblings, increasing the utility of this framework for further applications and theoretical development.

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 polygenicity in the trait tails: 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 :

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

where s_1 and s_2 represent index and conditional-sibling trait values, respectively, and n is the number of sibling pairs. This estimation method 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 (labeled “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 $s_1 > \Phi(q)$ and $s_1 < \Phi(q)$ for the upper and lower tails, respectively, where Φ^{-1} is the inverse normal cumulative distribution function.

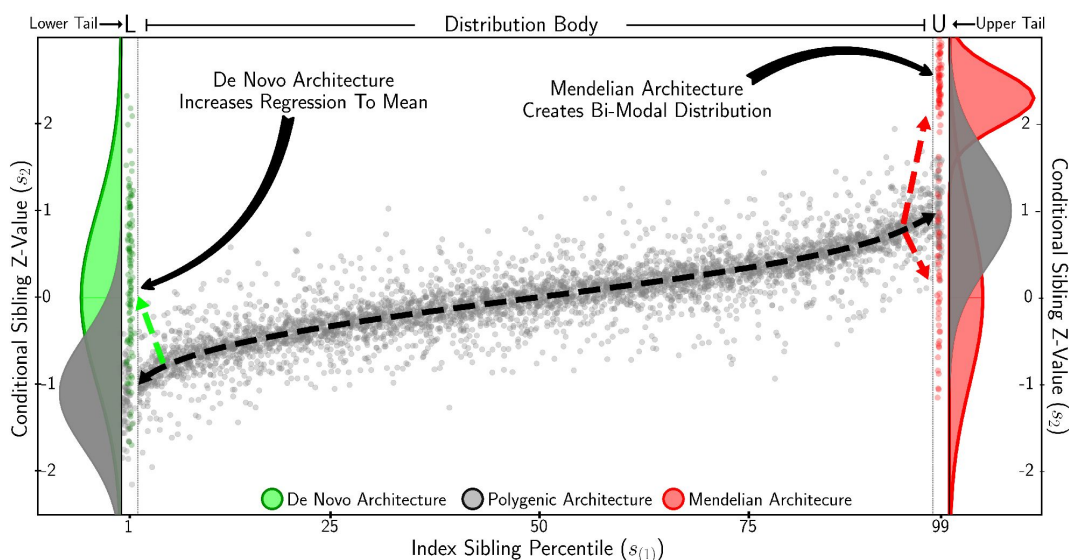


Fig. 2:

Identifying Complex Tail Genetic Architecture.

Conditional-sibling z-values plotted against index sibling quantiles. Grey depicts purely polygenic architecture across all 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*, an extreme scenario of Mendelian architecture is shown (in red), whereby half the siblings are concordant and half discordant, resulting in a bimodal conditional sibling distribution. Statistical tests to identify the presence of each type of complex tail architecture are designed to exploit these characteristics.

Statistical Test for De Novo Architecture

To identify *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:

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

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):

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

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

Statistical Test for Mendelian Architecture

To identify 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:

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

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_0 , the expected sibling concordance is $n\pi_0$ 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:

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

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**), assigning genotypes based on the allele frequencies of the first 100k SNPs from a trait GWAS. Additive parent liability can then be calculated based on the genotype effect size distribution of the GWAS. Next, parents are randomly paired, their liabilities averaged to produce midparent trait values (**step B**), and genotypes of two offspring ([Equation 1](#)) and corresponding genetic liabilities, G , are calculated (**step C**) assuming independent reassortment of parental alleles and unlinked SNPs.

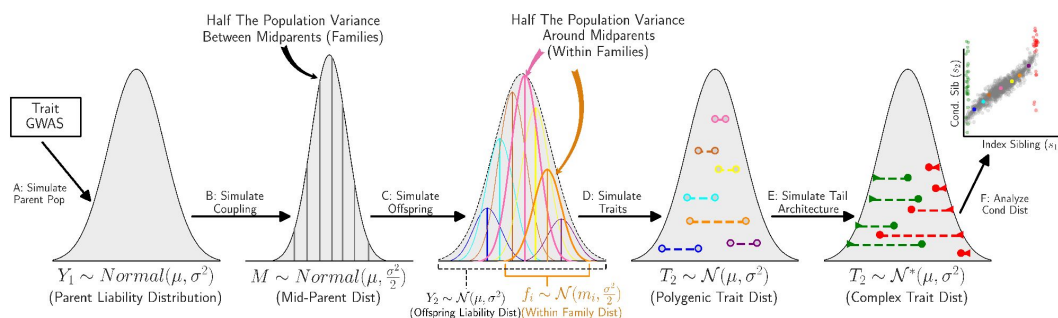


Fig. 3:

Simulation Schematic.

Publicly available GWAS allele frequency and effect size data is used to simulate parent liability (A). Midparent liability (B) is simulated assuming random mating. Offspring genotype and liability (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 an environmental random effect. For heritability, h^2 , offspring trait values are given by $T = hG + E$, where 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 and the correlation between the genetic liability and trait is equal to the heritability.

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 polygenic liability. We 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

For the UK Biobank analyses, we used six continuous traits (*Body Fat*, *Mean Corpuscular Haemoglobin*, *Neuroticism*, *Hell Bone Mineral Density*, *Monocyte Count*, and *Sitting Height*) with $h_{SNP}^2 > 5\%$ (22) and data on > 4,500 sibling pairs (sibling-pairs defined as having kinship coefficient 0.18 - 0.35 and > 0.1% SNPs with 0 IBD to distinguish from parent-offspring (23, 24). Outliers with absolute trait value > 6 standard deviations from the mean were removed and then cohort-wide trait values were standardised using a rank-based inverse normal transformation and adjusted for age, sex, recruitment centre, batch covariates and the first 40 principle components. The sub-sample corresponding to the sibling pairs was then re-normalised, and for each sibling pair one was randomly assigned as the index sibling and the other the conditional-sibling. Sibling pairs were then sorted by their index trait value and each sibling binned according to trait percentile rank among all siblings.

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 trait architecture (see **Model & Methods**), and apply our tests to trait data on thousands of siblings from the UK Biobank.

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.

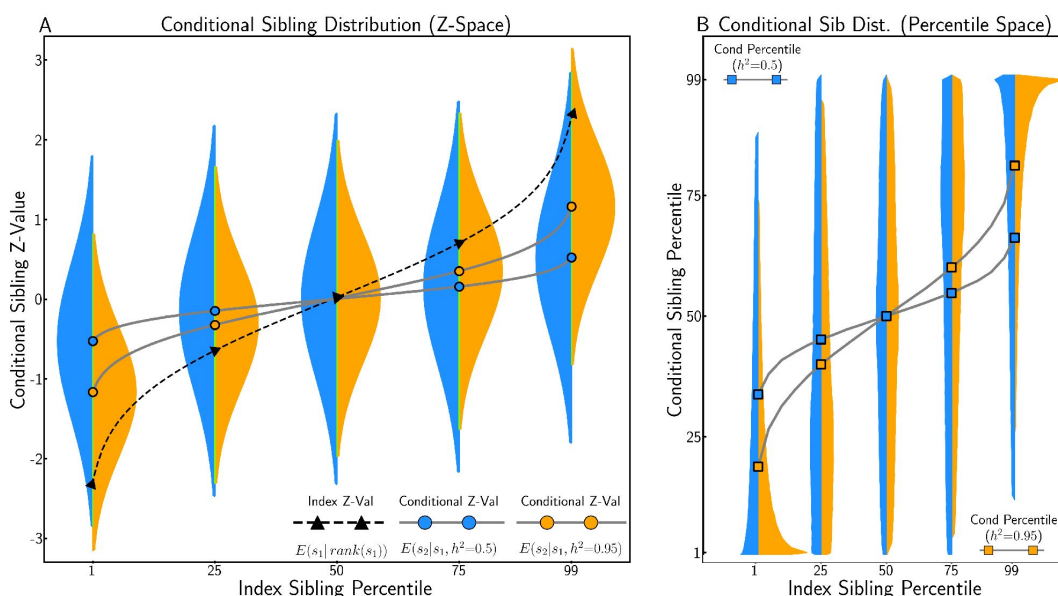


Fig. 4:

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.

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 (25) given that the inverse cumulative distribution functional of the normal distribution 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 genetic architecture underlying the trait (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 share both inherit the same mutant allele.

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.

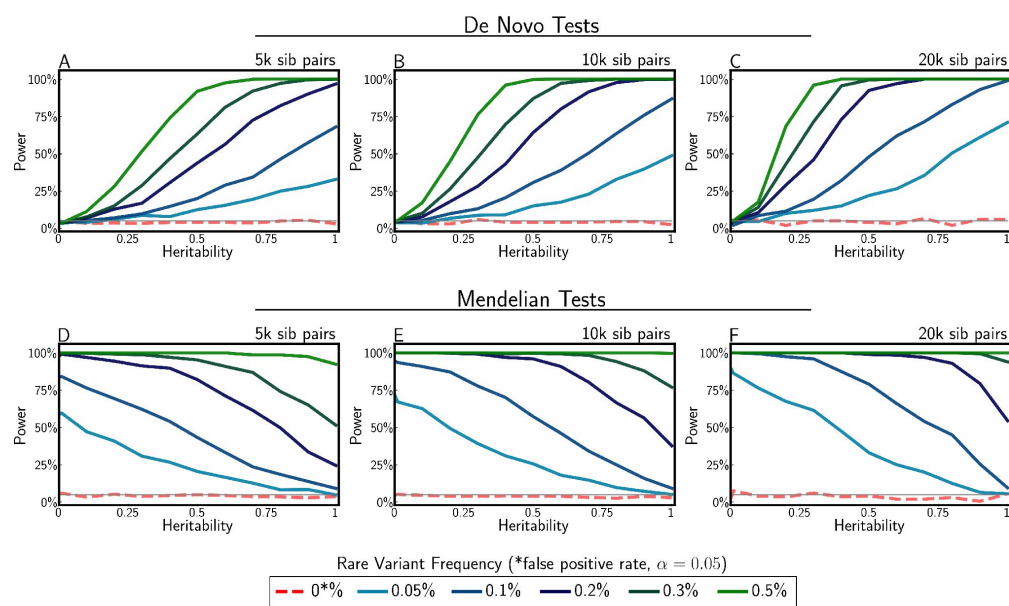


Fig. 5:

Power 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.

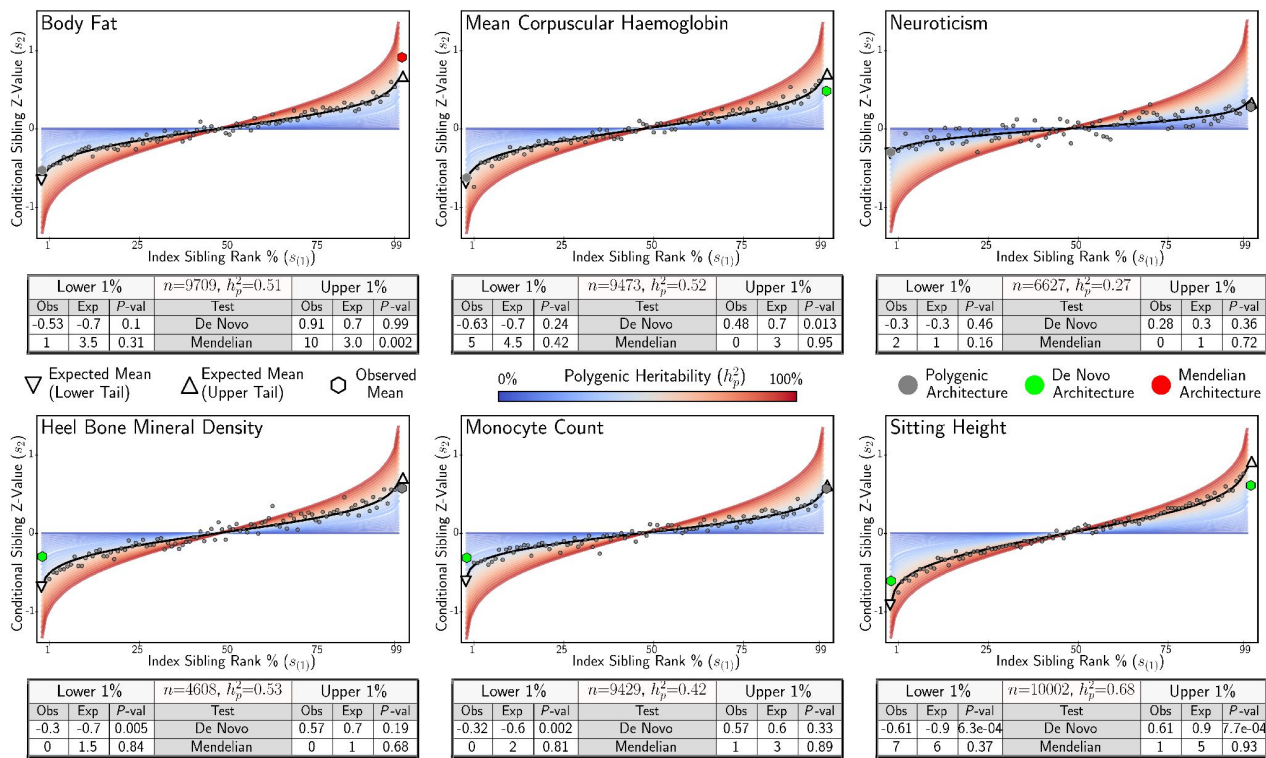


Fig. 6:

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 in which heritability is estimated by index siblings lying between the 5th and 95th percentiles. Statistical tests were applied to conditional siblings with index sibling 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.

Identifying Complex Tail Architecture in UK Biobank Data

We applied our two statistical tests of complex tail architecture to sibling-pair data on six traits from the UK Biobank (26) to illustrate the performance of our tests on real data. For each of the six traits, we tested the trait distribution for normality, estimated the (polygenic) heritability in the body of the trait distribution via Equation 5 (computed between the 5th and 95th percentiles), and performed tests to identify *de novo* and Mendelian architecture in the lower and upper tails of the distribution of each trait. We identified lower tail *de novo* architecture in *Heel Bone Mineral Density* and *Monocyte Count*, upper tail *de novo* architecture in *Mean Corpuscular Haemoglobin* and *de novo* architecture at both ends of the distribution in *Sitting Height*. Upper tail Mendelian architecture was identified in *Body Fat* and no complex tail architecture was detected for *Neuroticism*. These results support evidence from deep sequencing studies that indicate that rare variants play a substantial role in the genetic aetiology for *Sitting Height* (27, 28) and *Heel Bone Mineral Density* (29).

Discussion

In this paper, we present a novel approach to infer the genetic architecture of continuous traits, specifically in the tails of their 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. 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 high-impact alleles are present in the population on the order of 1 out of 1000 individuals. We apply our tests to six traits using UK Biobank data and find evidence for *de novo* architecture in the distribution tails of heel bone mineral density, monocyte count, mean corpuscular hemoglobin and sitting height, as well as Mendelian architecture in the upper tail of body fat.

There are several areas in which our work could have short-term utility. Firstly, those individuals inferred as having rare variants of high-impact 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 (30) or else perform more targeted sequencing of individuals in the trait tails with the goal of optimising power per cost (31). 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 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. For this reason, rejections 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 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 high-impact contribute to complex trait architecture, then 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 trait tails relative to the rest of the trait distribution. Moreover, a large meta-analysis of heritability estimates from twin studies (32) 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 (33), 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 midparents 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.

Acknowledgements

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. Peter Visscher for highlighting some important references.

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, z-value, case-status, or rank of one or more index siblings relative to the population. 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 (18, 34, 35).

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:

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

Then using the following identity (36):

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

We calculate the conditional sibling distribution similarly:

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

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, 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 (37) 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 $\sigma_g^2 = h^2$ and $\sigma_e^2 = 1 - h^2$, respectively. Similar to the previous case we begin by calculating the conditional midparent liability from Bayes theorem:

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

Then, we again use [Equation 11](#) to derive the distribution conditional sibling distribution:

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

Case 3) Multiple Index Trait Values, Continuous Trait: $P(S_3 | S_1 = s_1, S_2 = s_2)$

The previous case 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 (38), 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:

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

Thus, given the joint distribution for three siblings:

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

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 two sibling using [Equation 15](#)

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

Case 4) Binary Trait: $P(S_2 | S_1 = \text{Affected})$

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 (39), is equivalent to conditioning on the event where than index sibling's trait value is above or below a z-value threshold T :

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

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 calculated integrating over the normal distribution truncated at T :

$$p(S_2 | S_1 > T) = \int_T^\infty p(S_2 | S_1 = s_1) p(S_1 = s_1) ds_1 \quad (17)$$

The first two moments of the this truncated normal (40) are:

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

Approximating the index sibling distribution using a normal whose moments are taken from this truncated distribution, [Equation 17](#) becomes:

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

which can be solved using the identity given in [Equation 11](#):

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

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

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

which is equivalent to Reich's (41) correction to Falconer's (42) approximation where the relationship between the relatives is 0.5 for siblings.

The probability that of discordance given an *Unaffected* sibling can be calculated from Bayes Theorem:

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

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](#):

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

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

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

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 sibs relative to the polygenic null using the conditional distribution of a sib given index sib from (Equation 14). Since *de novo* mutations typically result in trait values in the tail of the distribution the test conditions on those index sibs in a specified upper quantile q of the distribution, i.e. those sib 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 sibs with trait values closer to the mean, giving a log-likelihood:

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

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

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

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

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

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

Statistical Test for Mendelian Architecture

Here we test for excess concordance between sibs 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 sib is in the tail $S_1 > \Phi^{-1}(q)$, where q is the quantile of interest. Denoting the set of index sibs in the tail by A_q and the size of the set by n , under the null of polygenicity 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:

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

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 $H_0 : \pi = \pi_0$:

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

Appendix 3 Model Evaluation

Here we compare our theoretical derivations ([Equation 4](#)) that rely on the *infinitesimal polygenic model* (18) 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 (43) for two traits to produce parent and offspring genotype and genetic liability (equivalent to trait value when $h^2 = 1$) ([Figure 8](#)).

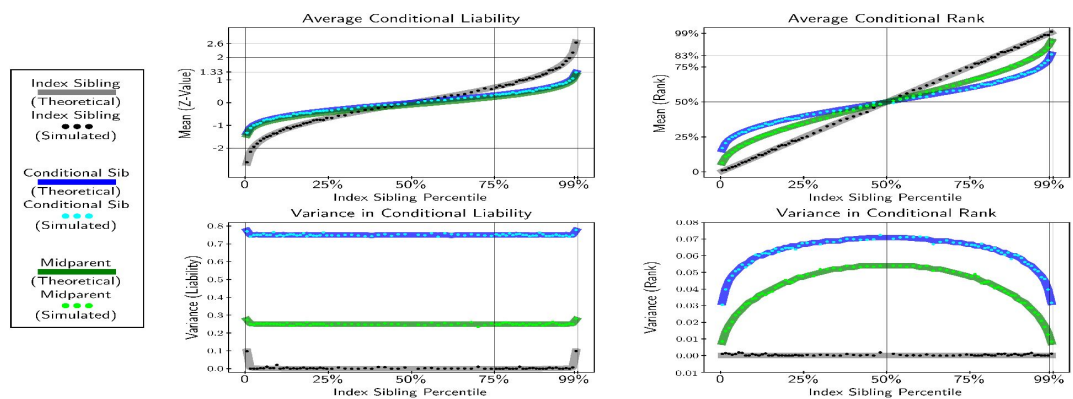


Fig. 7:

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.

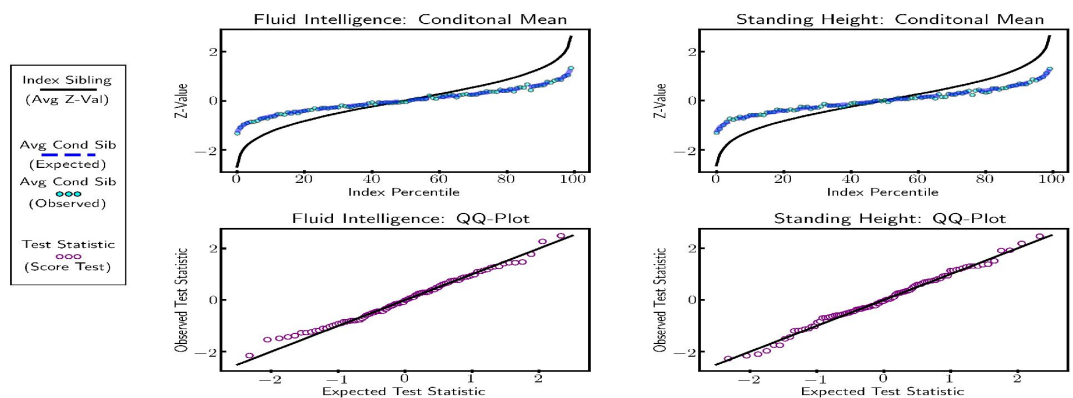


Fig. 8:

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.

References

1. Fisher Ronald A (1919) **Xv.—the correlation between relatives on the supposition of mendelian inheritance** *Earth and Envi-ronmental Science Transactions of the Royal Society of Edinburgh* **52**:399–433
2. Lush Jay L (1940) **Intra-sire correlations or regressions of offspring on dam as a method of estimating heritability of charac-teristics** *Journal of animal science* **1940**:293–301
3. Selzam Saskia , McAdams Tom A , Coleman Jonathan RI , Carnell Susan , O'Reilly Paul F , Plomin Robert , Llewellyn Clare H (2018) **Evidence for gene-environment correlation in child feeding: Links between common genetic variation for bmi in children and parental feeding practices** *PLoS genetics* **14**
4. Yengo Loïc , Vedantam Sailaja , Marouli Eirini , Sidorenko Julia , Bartell Eric , Sakaue Saori , Graff Marielisa , Eliassen Anders U , Jiang Yunxuan , Raghavan Sridharan , et al. (2022) **A saturated map of common genetic variants associated with human height** *Nature* **610**:704–712
5. Selvaraj Margaret Sunitha , Li Xihao , Li Zilin , Pampana Akhil , Zhang David Y , Park Joseph , Aslibekyan Stella , Bis Joshua C , Brody Jennifer A , Cade Brian E , et al. (2022) **Whole genome sequence analysis of blood lipid levels in> 66,000 individuals** *Nature communications* **13**
6. Singh Tarjinder , Poterba Timothy , Curtis David , Akil Huda , Al Eissa Mariam , Barchas Jack D , Bass Nicholas , Bigdeli Tim B , Breen Gerome , Bromet Evelyn J , et al. (2022) **Rare coding variants in ten genes confer substantial risk for schizophre-nia** *Nature* **604**:509–516
7. Iyegbe Conrad , O'Reilly Paul (2022) **Genetic origins of schizophrenia find common ground** *Nature* **604**
<https://doi.org/10.1038/d41586-022-00773-5>
8. Gorlov Ivan P , Gorlova Olga Y , Frazier Marsha L , Spitz Margaret R , Amos Christopher I (2011) **Evolutionary evidence of the effect of rare variants on disease etiology** *Clinical genetics* **79**:199–206
9. Maher M Cyrus , Uricchio Lawrence H , Torgerson Dara G , Hernandez Ryan D (2012) **Population genetics of rare variants and complex diseases** *Human heredity* **74**:118–128
10. Uricchio Lawrence H (2020) **Evolutionary perspectives on polygenic selection, missing heritability, and gwas** *Human genetics* **139**:5–21
11. Momozawa Yukihide , Mizukami Keijiro (2021) **Unique roles of rare variants in the genetics of complex diseases in humans** *Journal of human genetics* **66**:11–23

12. Weiner Daniel J , Wigdor Emilie M , Ripke Stephan , Walters Raymond K , Kosmicki Jack A , Grove Jakob , Samocha Kaitlin E , Goldstein Jacqueline I , Okbay Aysu , Bybjerg-Grauholm Jonas , et al. (2017) **Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders** *Nature genetics* **49**:978–985
13. Shakeshaft Nicholas G , Trzaskowski Maciej , McMillan Andrew , Krapohl Eva , Simpson Michael A , Reichenberg Avi , Cederlöf Martin , Larsson Henrik , Lichtenstein Paul , Plomin Robert (2015) **Thinking positively: the genetics of high intelligence** *Intelligence* **48**:123–132
14. Reichenberg Abraham , Cederlöf Martin , McMillan Andrew , Trzaskowski Maciej , Kapra Ori , Fruchter Eyal , Ginat Karen , Davidson Michael , Weiser Mark , Larsson Henrik , et al. (2016) **Discontinuity in the genetic and environmental causes of the intellectual disability spectrum** *Proceedings of the National Academy of sciences* **113**:1098–1103
15. Hur Yoon-Mi , Kim Jong Woo , Chung Kee Wha , Shin Joong Sik , Jeong Hoe-Uk , Auta Emmanuel (2013) **The nigerian twin and sibling registry** *Twin Research and Human Genetics* **16**:282–284
16. Douglas Scott Falconer and Trudy FC Mackay (1983) **Quantitative genetics**
17. Alberts Bruce , Johnson Alexander , Lewis Julian , Raff Martin , Roberts Keith , Walter Peter (2002) **Meiosis** *In Molecular Biology of the Cell*
18. Barton Nicholas H , Etheridge Alison M , Véber Amandine (2017) **The infinitesimal model: Definition, derivation, and implications** *Theoretical population biology* **118**:50–73
19. Bulmer Michael (1998) **Galton's law of ancestral heredity** *Heredity* **81**:579–585
20. Johnson Toby , Barton Nick (2005) **Theoretical models of selection and mutation on quantitative traits** *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**:1411–1425
21. Baselmans Bart ML , Yengo Loïc , van Rheenen Wouter , Wray Naomi R (2021) **Risk in relatives, heritability, snp-based heritability, and genetic correlations in psychiatric disorders: a review** *Biological Psychiatry* **89**:11–19
22. Ni Guiyan , Moser Gerhard , Ripke Stephan , Neale Benjamin M , Corvin Aiden , Walters James TR , Farh Kai-How , Holmans Peter A , Lee Phil , Bulik-Sullivan Brendan , et al. (2018) **Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood** *The American Journal of Human Genetics* **102**:1185–1194
23. Cheesman Rosa , Coleman Jonathan , Rayner Christopher , Purves KL , Morneau-Vaillancourt Genevieve , Glanville Kylie , Choi Shing W , Breen Gerome , Eley Thalia C (2020) **Familial influences on neuroticism and education in the uk biobank** *Behavior genetics* **50**:84–93

24. Bycroft Clare , Freeman Colin , Petkova Desislava , Band Gavin , Elliott Lloyd T , Sharp Kevin , Motyer Allan , Vukcevic Damjan , Delaneau Olivier , Jared O'Connell , et al. (2018) **The uk biobank resource with deep phenotyping and genomic data** *Nature* **562**:203–209
25. Johan Ludwig William Valdemar Jensen (1906) **Sur les fonctions convexes et les inégalités entre les valeurs moyennes** *Acta mathematica* **30**:175–193
26. Sudlow Cathie , Gallacher John , Allen Naomi , Beral Valerie , Burton Paul , Danesh John , Downey Paul , Elliott Paul , Green Jane , Landray Martin , et al. (2015) **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**
27. Tcheandjieu Catherine , Aguirre Matthew , Gustafsson Stefan , Saha Priyanka , Potiny Praneetha , Haendel Melissa , Ingelsson Erik , Rivas Manuel A , Priest James R (2020) **A phenome-wide association study of 26 mendelian genes reveals phenotypic expressivity of common and rare variants within the general population** *PLoS genetics* **16**
28. Bjornsdottir Gyda , Stefansdottir Lilja , Thorleifsson Gudmar , Sulem Patrick , Norland Kristjan , Ferkingstad Egil , Oddsson Asmundur , Zink Florian , Lund Sigrun H , Nawaz Muhammad S , et al. (2022) **Rare slc13a1 variants associate with intervertebral disc disorder highlighting role of sulfate in disc pathology** *Nature communications* **13**:1–13
29. Kemp John P , Morris John A , Medina-Gomez Carolina , Forgetta Vincenzo , Warrington Nicole M , Youtten Scott E , Zheng Jie , Gregson Celia L , Grundberg Elin , Trajanoska Katerina , et al. (2017) **Identification of 153 new loci associated with heel bone mineral density and functional involvement of gpc6 in osteoporosis** *Nature genetics* **49**:1468–1475
30. Uffelmann Emil , Huang Qin Qin , Munung Nchangwi Syntia , De Vries Jantina , Okada Yukinori , Martin Alicia R , Martin Hilary C , Lappalainen Tuuli , Posthuma Danielle (2021) **Genome-wide association studies** *Nature Reviews Methods Primers* **1**
31. Yang Jinliang , Jiang Haiying , Yeh Cheng-Ting , Yu Jianming , Jeddloh Jeffrey A , Nettleton Dan , Schnable Patrick S (2015) **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**:587–596
32. Polderman Tinca JC , Benyamin Beben , De Leeuw Christiaan A , Sullivan Patrick F , Van Bochoven Arjen , Visscher Peter M , Posthuma Danielle (2015) **Meta-analysis of the heritability of human traits based on fifty years of twin studies** *Nature genetics* **47**:702–709
33. Wray Naomi R , Wijmenga Cisca , Sullivan Patrick F , Yang Jian , Visscher Peter M (2018) **Common disease is more complex than implied by the core gene omnigenic model** *Cell* **173**:1573–1580
34. Barton Nicholas H , Keightley Peter D (2002) **Understanding quantitative genetic variation** *Nature Reviews Genetics* **3**:11–21
35. Barton Nicholas H (2022) **The “new synthesis”** *Proceedings of the National Academy of Sciences* **119**

36. Barndorff-Nielsen Ole , Kent John , Sørensen Michael (1982) **Normal variance-mean mixtures and z distributions** *International Statistical Review/Revue Internationale de Statistique* :145–159
37. McCaw Zachary R , Lane Jacqueline M , Saxena Richa , Redline Susan , Lin Xihong (2020) **Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies** *Biometrics* **76**:1262–1272
38. Bernardo J. M. , Smith A. F. M. (1994) **Bayesian Theory**
39. Falconer DS (1967) **The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus** *Annals of human genetics* **31**:1–20
40. Johnson Norman L , Kotz Samuel , Balakrishnan Narayanaswamy (1995) **Continuous univariate distributions, volume 2**
41. Reich Ti , James JW , Morris CA (1972) **The use of multiple thresholds in determining the mode of transmission of semicontinuous traits** *Annals of human genetics* **36**:163–184
42. Falconer Douglas S (1965) **The inheritance of liability to certain diseases, estimated from the incidence among relatives** *Annals of human genetics* **29**:51–76
43. Benjamin Neale (2018) **Benjamin Neale. Neale lab data: <http://www.nealelab.is/uk-biobank/>, 2018.**

Author information

Tade Souaiaia

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

For correspondence: tade.souaiaia@downstate.edu

ORCID iD: [0000-0003-3922-1372](https://orcid.org/0000-0003-3922-1372)

Hei Man Wu

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

Clive Hoggart

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

ORCID iD: [0000-0002-6875-6993](https://orcid.org/0000-0002-6875-6993)

Paul O'Reilly

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

For correspondence: paul.oreilly@mssm.edu

ORCID iD: [0000-0001-7515-0845](https://orcid.org/0000-0001-7515-0845)

Editors

Reviewing Editor

Alexander Young

University of California, Los Angeles, United States of America

Senior Editor

Molly Przeworski

Columbia University, United States of America

Reviewer #1 (Public Review):

The authors sought to craft a method, applicable to biobank-scale data but without necessarily using genotyping or sequencing, to detect the presence of de novo mutations and rare variants that stand out from the polygenic background of a given trait. Their method depends essentially on sibling pairs where one sibling is in an extreme tail of the phenotypic distribution and whether the other sibling's regression to the mean shows a systematic deviation from what is expected under a simple polygenic architecture.

Their method is successful in that it builds on a compelling intuition, rests on a rigorous derivation, and seems to show reasonable statistical power in the UK Biobank. (More biobanks of this size will probably become available in the near future.) It is somewhat unsuccessful in that rejection of the null hypothesis does not necessarily point to the favored hypothesis of de novo or rare variants. The authors discuss the alternative possibility of rare environmental events of large effect. Maybe attention should be drawn to this in the abstract or the introduction of the paper. Nevertheless, since either of these possibilities is interesting, the method remains valuable.

Reviewer #2 (Public Review):

Souaiaia et al. attempt to use sibling phenotype data to infer aspects of genetic architecture affecting the extremes of the trait distribution. They do this by considering deviations from the expected joint distribution of siblings' phenotypes under the standard additive genetic model, which forms their null model. They ascribe excess similarity compared to the null as due to rare variants shared between siblings (which they term 'Mendelian') and excess dissimilarity as due to de-novo variants. While this is a nice idea, there can be many explanations for rejection of their null model, which clouds interpretation of Souaiaia et al.'s empirical results.

The authors present their method as detecting aspects of genetic architecture affecting the extremes of the trait distribution. However, I think it would be better to characterize the method as detecting whether siblings are more or less likely to be aggregated in the extremes of the phenotype distribution than would be predicted under a common variant, additive genetic model.

Exactly how the rareness and penetrance of a genetic variant influence the conditional sibling phenotype distribution at the extremes is not made clear. The contrast between de-novo and 'Mendelian' architectures is somewhat odd since these are highly related phenomena: a 'Mendelian' architecture could be due to a de-novo variant of the previous generation. The fact that these two phenomena are surmised to give opposing signatures in

the authors' statistical tests seems suboptimal to me: would it not be better to specify a parameter that characterizes the degree of sharing between siblings of rare factors of large effect? This could be related to the mixture components in the bimodal distribution displayed in Fig 1. In fact, won't the extremes of all phenotypes be influenced by all three types of variants (common, rare, de-novo) to greater or lesser degree? By framing the problem as a hypothesis testing problem, I think the authors are obscuring the fact that the extremes of real phenotypes likely reflect a mixture of causes: common, de-novo, and rare variants (and shared and non-shared environmental factors).

To better enable interpretation of the results of this method, a more comprehensive set of simulations is needed. Factors that may influence the conditional distribution of siblings' phenotypes beyond those considered include: non-normal distribution, assortative mating, shared environment, interactions between genetic and shared environmental factors, and genetic interactions.

In summary, I think this is a promising method that is revealing something interesting about extreme values of phenotypes. Determining exactly what is being revealed is going to take a lot more work, however.

Author Response:

The major criticism from the reviewers is that factors other than high-impact rare variants – such as environmental factors or epistasis – could have produced the complex tail architecture that we test for and detect. While we did explain this point in the Discussion, we agree with the reviewers that this should have been emphasized more and earlier in the manuscript.

Regarding suggestions for more complex simulations and methods, we absolutely agree that much more work is needed here to produce optimised inference of all the causes of complex tail architecture. We are performing multiple projects at various stages of completion that we hope will contribute to this, but we felt that this was a good stopping-point in this project to publish what we had completed so far, in order to: (1) introduce the idea of inferring complex genetic architecture from siblings without requiring genetic data, (2) outline an initial theoretical framework for inferring complex tail architecture from sibling data, (3) provide simple tests powered to identify enrichments of de novo or 'Mendelian' variants in the tails (albeit tests that make several strong simplifying assumptions), (4) enable others interested in the topic to build upon this work now. However, we plan to expand our simulations and analyses in a revised manuscript based on reviewer feedback.

We thank the reviewers for their comments about the value of our work, its mathematical robustness and the promise of our method.