The trait coding rule in phenotype space

Jianguo Wang

Sun Yat-Sen University

Xionglei He (✉ hexiongl@mail.sysu.edu.cn)

Sun Yat-Sen University <https://orcid.org/0000-0003-1050-802X>

Article

Keywords:

Posted Date: February 16th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1297947/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Jan 21, 2022

The trait coding rule in phenotype space

Jianguo Wang & Xionglei He

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou
510275, China

Correspondence should be addressed to X.H. (hexiongl@mail.sysu.edu.cn).

Abstract

Is it possible to infer the genetically determined value of a quantitative trait from other traits? To answer the question we need to know how traits are coded in phenotype space (P) which can be partitioned into a subspace determined by genetic factors (P^G) and a subspace affected by non-genetic factors (P^{NG}). Evolutionary theory predicts P^G is composed of limited dimensions while P^{NG} may have infinite dimensions, which suggests a novel dimension decomposition method to separate them. After validating the method using simulation we applied it to a yeast phenotype space comprising 405 traits. The obtained yeast P^G matches the actual genetic components of the yeast traits, explains the broad-sense heritability, and facilitates the mapping of quantitative trait loci, highlighting the success of the subspace separation. A limited number of latent dimensions in the P^G were found to be recurrently used for coding the diverse yeast traits, while dimensions in the P^{NG} tend to be trait specific and increase constantly with trait sampling. Similar results were obtained by analyzing the UK Biobank human brain phenome, which elucidated the genetic versus

non-genetic origins of the left-right asymmetry of brain. In sum, phenotypic traits are coded by a rather small set of genetically determined basic dimensions and numerous trait-specific dimensions that are shaped by non-genetic factors, a rule enabling the identification of genetic components of quantitative traits based solely on phenotype.

Introduction

Although genotype and phenotype are both the themes of genetics, our understanding of phenotype is surprisingly poor ¹. One of the main reasons is the inferior objectivity of phenotype ²⁻⁵. For example, while height and weight appear to be objective characters of humans, the body mass index (BMI), which is the weight divided by the square of height, is artificial ⁶. The inferior objectivity, with the resulting infinite complexity, has rendered phenotype a peripheral topic in conventional genetic studies ¹. Notably, in recent years we witnessed growing interests in phenotype, evidenced by the many proposals for mapping the phenome of a variety of organisms ⁷⁻¹⁰. In addition to producing data, the emerging field of phenomics is facing a fascinating challenge: What are, like the triplet codons in genotype, the coding rules in phenotype? ¹¹

Since a phenotypic trait often varies among individuals, one can study the various traits of a species in a multi-dimensional linear space (P), in which each trait (T) is represented by a vector. Considering that there should be a set of basic independent dimensions in a given P , we propose

$$P = P^G + P^{NG}, \quad (1)$$

where P^G represents a subspace formed by the dimensions that are fully determined by genotype (i.e., G-dimensions), and P^{NG} represents the other subspace formed by dimensions that are

explained by non-genetic factors such as environmental variables, developmental plasticity, measuring errors, human definitions, and so on (i.e., NG-dimensions). Accordingly, each trait (T) can be expressed as a linear function of the G-dimensions and NG-dimensions

$$T = \sum_j a_j G_j + \sum_k b_k NG_k, \quad (2)$$

where G_j represents the j^{th} G-dimension in P^G , NG_k represents the k^{th} NG-dimension in P^{NG} , and a_j and b_k represent the coefficients of G_j and NG_k , respectively. Hence, T is a combination of the vector of genetically determined values (for simplicity, genetic component or T^G) and the vector of the values ascribing to non-genetic factors (for simplicity, non-genetic component or T^{NG}):

$$T = T^G + T^{\text{NG}}, \quad (3)$$

where $T^G = \sum_j a_j G_j$ and $T^{\text{NG}} = \sum_k b_k NG_k$.

The Fisher's geometric model¹², together with the extension by Orr¹³, predicts that the number of G-dimensions should be rather small for extant organisms. This is because a very large number of G-dimensions would hinder the adaptation to new environments, leading to extinction of the organisms, a phenomenon termed as “cost of complexity”¹³. Although the model does not predict exactly how small the number of G-dimensions should be¹⁴, it suggests robust constraints on the dimensionality of P^G ¹⁵. In sharp contrast, the number of NG-dimensions in P^{NG} would be infinite. This is because of the variability of environment, the randomness of developmental plasticity and measuring error, and the arbitrariness of human definition¹⁶. The enormous complexity resulting from the infinite dimensionality of P^{NG} suggests the necessity of separating P^G from P^{NG} before revealing any rules in P .

In this study, we developed a geometric method that is able to separate subspaces with distinct dimensionality. The method offered a phenome-based approach to decomposing the genetic and non-genetic components for hundreds of quantitative traits in yeasts and humans, revealing a rather simple rule on how the traits are coded in phenotype space.

Results

Theory of uncorrelation-based high-dimensional dependence (UBHDD)

Let's first consider a two-dimensional space with three non-parallel and non-orthogonal vectors α , β , and η (Fig. 1). Based on linear algebra, η can always be expressed as a linear function of α and β no matter whether α and β have strong (Fig. 1a) or little (Fig. 1b) marginal correlation to η . This is because the three vectors share the two dimensions (X-axis and Y-axis). In the three-dimensional space shown in Fig. 1c, η has a unique dimension (Z-axis). As a result, η can no longer be expressed by α and β despite the same marginal correlations to α and β as in Fig. 1b. Hence, dimension sharing but not marginal correlation underlies the high-dimensional dependence among vectors.

We derived the probability ($Pr(\Psi)$) of two k -dimensional vectors that share the same dimensions in an N -dimensional space as a function of their marginal correlation (Pearson's R^2 ; $0 < R^2 < 1$) (Methods). Without loss of generality, the probability trajectories of $N = 10, 100$, and $1,000$ are shown, respectively, for two vectors each with two dimensions (Fig. 1d). There are three corollaries: First, the probability converges to one if the two vectors have a strong marginal correlation for any finite N , which is formulated as $Pr(\Psi) \rightarrow 1$ if $R^2 \rightarrow 1$ and $N < N_0$, where N_0

is a finite number. Second, with the decrease of the marginal correlation between the two vectors, the probability converges to zero in a space of very large N , which is formulated as $Pr(\Psi) \rightarrow 0$ if $N \rightarrow \infty$ and $0 < R^2 < R_u^2$, where $R_u^2 \leq 1$. Third, with the decrease of the marginal correlation between the two vectors, the probability remains reasonably high in a space of small N (such as $N=10$), which is formulated as $Pr(\Psi) > Pr_0$ if $N < N_0$ and $0 < R^2 < R_u^2$, where $Pr_0 \geq 0$. Accordingly, given an R_u with a small absolute value, uncorrelated vectors ($R^2 < R_u^2$) would have a high probability of sharing dimensions in a space of small N but little probability in a space of very large N . This suggests a strategy for separating P^G from P^{NG} , the former of which is supposed to have a limited N while the latter an infinitely large N .

Specifically, for a trait T_i in a given P , we can define its uncorrelated traits under certain R_u threshold. Then, T_i can be modelled by its uncorrelated traits as

$$T_i = \sum_{j \in U_i} b_j T_j + \varepsilon_i, \quad (4)$$

where T_j is j^{th} trait in P , b_j is the coefficient, ε_i is the residual vector and U_i contains the indices of T_i 's uncorrelated traits (Methods). With the increase of trait sampling in P and the decrease of R_u , only G-dimensions but no NG-dimensions would be shared between T_i and its uncorrelated traits. As a result, each G-dimension of T_i can be expressed as a linear function of its uncorrelated traits, which is, however, not the case for the NG-dimensions of T_i . Hence, according to Eq. (4), we expect $T_i^g = \sum_{j \in U_i} b_j T_j$ and $T_i^{ng} = T_i - T_i^g$, where T_i^g represents the estimated genetic component (formed by G-dimensions) and T_i^{ng} the estimated non-genetic component. By conducting such

uncorrelation-based decomposition for every trait in P we would achieve the separation of P^G from P^{NG} . We term the method uncorrelation-based high-dimensional dependence (UBHDD).

Validation of UBHDD in simulated phenotype space

To validate UBHDD in separating subspaces with distinct dimensionality, we simulated a phenotype space comprising P^G with a small number of G-dimensions ($N_1=10$) and P^{NG} with a much larger number of NG-dimensions ($N_2=10,000$) (Methods). The G-dimensions and NG-dimensions are generated by standard multivariate normal distribution. Each trait (T) is generated by random linear combination of the G-dimensions and NG-dimensions as given by Eq. (3), with the former representing T^G and the latter representing T^{NG} . A total of 1,000 traits are simulated in a population of 1,000 individuals. Each trait is standardized such that the variance of T^G equals to the broad-sense heritability (H^2). Combining all T^G or all T^{NG} forms the sampled P^G or P^{NG} , respectively.

UBHDD is conducted as follows: two traits are defined as uncorrelated if their Pearson's $R^2 < R_u^2$, where $R_u^2 \approx 0.02$, corresponding to $p = 0.01$ (T-test with Bonferroni correction); with conventional machine learning framework (LASSO) we modeled each trait using its uncorrelated traits, as given by Eq. (4); the resulting matrices containing all T^g and all T^{ng} are called P^g and P^{ng} , representing the estimated P^G and P^{NG} , respectively.

As expected, with the increase of trait sampling the number of sampled dimensions is much more rapidly saturated for P^G than for P^{NG} (Fig. 1e). We noted that the sampled dimensions in P^{NG} would keep increasing if the dimensionality of P^{NG} was infinitely large. In addition, correlated traits share both G-dimensions and NG-dimensions while uncorrelated traits could share G-

dimensions but rarely NG-dimensions (Fig. 1f). This suggests G-dimensions but not NG-dimensions would underlie the success of UBHDD. Indeed, in all cases we find the T^{E} estimated by UBHDD is highly correlated with T^{G} , the actual genetic component of T (Fig. 1g). The variance of T^{E} also matches well the variance of T^{G} , the broad-sense heritability of T (Fig. 1h). We also simulated phenotype spaces with $N_1 = 20, 50$, or 100 (N_2 remains unchanged), and obtained largely the same results (Fig. S1). These analyses validated the capacity of UBHDD in separating P^{G} from P^{NG} .

It is worth noting that UBHDD is a method of dimension decomposition but not dimension reduction. We compared UBHDD with principal component analysis (PCA), a classical dimension reduction method, in a simulated phenotype space with structure. The structured space is simulated as above except that two large clusters with strongly correlated members exist (Fig. S2a; Methods). We found UBHDD remained successful in separating P^{G} from P^{NG} , insensitive to the space structure (Fig. S2b). However, PCA overfits the traits in the two large clusters and underfits the others (Fig. S2c; Methods). The failure of PCA in separating P^{G} from P^{NG} is not surprising because PCA maximizes the explained variance of the top PCs and is therefore sensitive to data structure.

Using UBHDD to separate a yeast phenotype space

We examined a phenotype space comprising 405 morphological traits of the budding yeast *Saccharomyces cerevisiae*¹⁷. The traits are measured in a population of 815 segregants, each of which has two clones/replicates and known genotype¹⁸ (Fig. 2a). The traits are typically about area, distance, angle, and brightness that describe the shape of mother cell and bud, the neck

separating mother cell from bud, the localization of the nuclei in mother cell and bud, and so on, across different cell stages (Fig. 2b). The narrow-sense heritability (h^2) of the traits ranges from 0 to 0.56 with a median of 0.15, and the broad-sense heritability (H^2) ranges from 0 to 0.86 with a median of 0.42 (Fig. S3).

In conventional quantitative genetics a trait T is often composed of a genetic component (T^G) and an environmental component that represents the contribution of all non-genetic factors (T^{NG}). When the genotype of individuals is known, one often uses a linear mixed model (LMM) to separate the T^G from T^{NG} . Meanwhile, the separation could be done by UBHDD, which requires only phenome information according to the above theory and simulation results (Fig. 2c).

We applied UBHDD to the 405 yeast traits and obtain for each of them the T^{g} and T^{ng} (Methods). We also applied LMM to separate T^G from T^{NG} for each trait by taking advantage of the genotype/replicate information (Methods). For most of the traits T^{g} is highly correlated to T^G (Fig. 2d-e). The among-individual variance of T^{g} ranges from 0.03 to 0.98 across the different traits (median=0.53), which is much larger than what a shuffling analysis (serving as negative control) could achieve (maximum=0.013) (Fig. 2f; Methods). Furthermore, the variance of T^{g} is comparable to the variance of T^G , the broad-sense heritability estimated by LMM (Fig. 2g). The results are robust to the R_u threshold used for defining uncorrelation (Fig. S4). Notably, the T^{g} estimation could be improved in a larger population that enables more robust UBHDD modelling; meanwhile, the T^G estimation could be more accurate if there were more than two replicates. As another critical test, we expect T^{g} should have a larger narrow-sense heritability (h^2) than T^{ng} . Indeed, in most case the h^2 of T^{g} is larger than that of T^{ng} , and also more QTLs were detected for

T^{g} than T^{ng} (Fig. 2h-i; Methods). Taken together, these results suggest the overall success of UBHDD in separating the genetic component of the yeast traits.

The separations by UBHDD are robust between two yeast populations

In addition to the segregant population (seg-population), we also examined a yeast gene-deletion population (del-population) which contains ~5,000 *S. cerevisiae* strains each lacking a non-essential gene (Fig. 2j). The same 405 traits were measured for each of the strains in the del-population¹⁰. We conducted UBHDD in the del-population and obtained the T^{g} and T^{ng} for each of the traits (Methods). We then compared the T^{g} functions learned in del-population with the T^{g} functions previously learned in seg-population (Methods). Taking the trait C11.1_A as an example, when the T^{g} function learned in seg-population was applied to del-population, the T^{g} estimations are highly similar to the estimations by the T^{g} function learned in del-population, with an identity score = 0.88 (Fig. 2k; Methods). The identity score of the 405 traits ranges from 0.29 to 0.99 with a median of 0.82 (Fig. 2l), suggesting the genetic subspace obtained by UBHDD be robust between the two yeast populations.

Using UBHDD to separate human brain phenotype space

To test if UBHDD works in a more complex phenotype space, we examined UK Biobank human phenome. We focused on the 675 image-derived phenotypes (IDPs) of brain generated by dMRI in 25,957 white British individuals without kinship and with genotype available (Fig. 3a; Methods)¹⁹. These brain traits represent 9 different measures including fractional anisotropy (FA), intra-

cellular volume fraction (ICVF), isotropic or free water volume fraction (ISOVF), mean diffusivity (MD), diffusion tensor mode (MO), orientation dispersion index (OD) and the three eigenvalues in a diffusion tensor fit (L1, L2 and L3) in up to 75 brain regions.

We applied UBHDD to the 675 brain traits after excluding covariates and obtained for each of them the T^g and T^{ng} (Methods). The variance of T^g ranges from 0.17 to 0.87 with a median of 0.48, which is much larger than what a shuffling analysis could achieve, suggesting shared dimensions between uncorrelated traits captured by UBHDD (Fig. 3b; Methods). The results are robust to the R_u threshold (Fig. S5). Because there are, unlike yeasts, no clones (i.e., monozygotic twins) for most individuals, we couldn't use LMM to estimate the T^G and broad-sense heritability. Instead, we examined narrow-sense heritability. Consistent with the findings in yeast, T^g in general has a larger h^2 than T^{ng} (Fig. 3c; Methods). There are also more QTLs detected in T^g than T^{ng} (Fig. 3d).

The separation of T^g from T^{ng} of the brain traits provides a unique opportunity to resolve a long-standing puzzle, namely, the relative contribution of genetic versus non-genetic factors to the left-right asymmetry of human brain^{20,21}. We examined all 297 trait pairs each representing the same measure in two symmetrical brain regions. We calculated for each trait pair the Pearson's R^2 of T^g and T^{ng} , respectively. In all trait pairs the R^2 of T^g is much larger than that of T^{ng} (Fig. 3e). The R^2 shows differences across measures and brain regions, revealing the applicability of different measures to demonstrating the genetic or non-genetic factors of brain-region-specific asymmetry. This finding indicates non-genetic factors be the major source of the brain asymmetry, highlighting environmental effects on asymmetry associated brain physiology and dysfunction.

Distinct dimensionality of P^g and P^{ng} reveals a trait coding rule

Using UBHDD we estimated the genetic component T^g and non-genetic component T^{ng} for each of the traits examined in yeasts and humans. Combining all T^g of the yeast traits (or human traits) forms P^g , the estimated genetic subspace of the yeast (or human) phenotype space. Similarly, combining all T^{ng} forms P^{ng} , the estimated non-genetic subspace. We then examined the latent dimensions in P^g and P^{ng} , respectively. Using principle component analysis (PCA) we obtained the number of top PCs that explain 85% variance of a subspace. We found that, with the increase of trait sampling, the number of PC dimensions is rapidly saturated for P^g but not for P^{ng} (Fig. 4a-b). The observation of distinct dimensionality between P^g and P^{ng} is consistent with the theory of UBHDD as proposed in Fig. 1. To know how the dimensions of P^g and P^{ng} are used by the traits we calculated the gradient between the number of sampled traits (n_T) and the number of obtained dimensions (n_D), denoted as $\Delta n_T / \Delta n_D$. With the increase of dimensionality the gradient rapidly increases to be large for P^g but remains small for P^{ng} , suggesting the P^g dimensions are recurrently used by the traits while the P^{ng} dimensions tend to be trait-specific (Fig. 4c-d). Consistently, the pairwise correlation of T^g , which measures dimension sharing, is much larger than that of T^{ng} (Fig. 4e-h). Therefore, in both the yeast and human brain phenotype space the traits are coded by a rather small set of common dimensions that are determined by genotype and numerous trait-specific dimensions that are shaped by non-genetic factors.

Discussion

In this study we dissected the geometric substructures of two phenotype spaces, revealing a simple rule on how traits are coded by two types of latent dimensions. We noted that the learned trait coding rule in turn suggests a phenome-based strategy for separating the genetic (both additive and non-additive) component from non-genetic component of a quantitative trait (Fig. 5). The conventional strategy used in quantitative genetics relies on biological replicates (i.e., clones or monozygotic twins) that are, however, often unavailable. There are two additional related strategies for estimating the additive part of genetic component: One is to refer to the trait of parents and the other relies on genotype information. Hence, the phenomic strategy would be a useful complement to the conventional strategies. In particular, it may revive the classical field of evolutionary morphospace^{11,22-24} in which a lot of morphological traits of wild animals or plants are available while the kinship and genotype information often lacks. Meanwhile, the phenomic strategy would be of tremendous value to the emerging field of phenomics²⁵. For example, it would guide the trait sampling strategy in phenomic studies by gauging the obtained genetic and non-genetic dimensions.

There are two implications worth discussion. First, this study highlights the success of a coarse-grained approach to understanding phenotype. By adopting a from-trait-to-trait view, the purpose of finding the genetic dimensions of phenotype space is not to map the underlying genes but to obtain the intrinsic phenotypic components by excluding extrinsic effects. The coarse-grained approach is timely considering the increasing difficulty of the traditional gene-centered reductionism approach in studying complex traits²⁵⁻²⁸. Second, this study suggests a strategy for dealing with the “curse of dimensionality”^{29,30}. It is conceivable that, like phenotype space, many complex systems can be partitioned into a part determined by intrinsic factors and another part shaped by extrinsic factors, the former of which is of rather low dimensionality while the latter is

composed of myriad dimensions. Hence, the UBHDD method could be a generally useful tool to the studies of complex systems.

There are also a few caveats in the study. First, the UBHDD method depends on dense sampling of a phenotype space. We may use a down-sampling strategy to assess the sufficiency of trait sampling. We found the overall performance of UBHDD for the yeast traits is nearly saturated (Fig. S6a); however, the performance for the human brain traits is sensitive to down-sampling (Fig. S6b), suggesting the current sampling of the brain space is still insufficient. Second, the uncorrelation thresholds (R_u) used in this study may not be ideal. In principle, a smaller R_u is always helpful for avoiding the effects of non-genetic dimensions, which, however, would leave too few traits for conducting UBHDD. We found a good assessment of the threshold lies in the learned UBHDD functions. In our cases, the coefficients (Co) in the learned UBHDD function of a focal trait are not explained by the marginal correlations (Mc) of the explanatory variables (traits) to the focal trait (Fig. S7). In other words, the performance of UBHDD does not rely on those traits with stronger marginal correlation to the focal trait. In future, we may optimize R_u threshold trait by trait by considering the number of remaining traits under a given threshold as well as the relationship between Mc and Co . Third, only two phenotype spaces are examined. The generality of the findings should be further tested in more complex phenotype spaces comprising various types of traits (morphology, physiology, behavior, and so on) and with many organs involved.

Methods

Yeast segregant population (seg-population).

We study a panel of segregants of a yeast cross (*S. cerevisiae* strain BY $\times$ strain RM) generated by a previous study¹⁸. A total of 1,008 segregants are available with genotypes, among which 815 were phenotyped¹⁷. The obtained 405 phenotypic traits measure the areas and circumferences, the elliptical approximation, brightness, thickness, axis length, neck width, neck position, bud position, axis ratio, cell size ratio, outline ratio, proportion of budded cells, proportion of small budded cells, segment distances between mother tip, bud tip, middle point of neck, center of mother and bud, nuclear gravity centers, nuclear brightest points, angles between segments, and so on. The phenotyping was conducted for two clones of each segregant, and the trait values are Z-score transformed.

Yeast single-gene-deletion population (del-population).

A previous study¹⁰ has conducted similar phenotyping for 4718 yeast mutants each lacking a non-essential gene (del-population). The 405 traits in the seg-population are also available in the del-population. These traits in the del-population are scaled based on the mean and standard deviation in the seg-population to make models obtained from the two populations comparable.

UK Biobank.

We collect the 870 brain MRI phenotypes and related covariates (measuring center, age, sex, weight, home location and MRI system parameters) in UK Biobank¹⁹, of which 675 image-derived phenotypes (IDPs) measured by dMRI are chosen for UBHDD modelling, heritability estimation and QTL mapping. A few missing values in covariates are imputed by linear regression with brain phenotypes. A total of 25,957 White English without kinship and with genotypes are chosen. We conduct normalization on each of the 675 traits (R package, 'bestNormalize') and exclude the contribution of covariates by linear regression (R, 'lm'). These traits are subject to following

analysis including UBHDD, estimation of narrow-sense heritability and QTL mapping. Exemplar traits are obtained by R package ‘aplcluster’. We also obtain the genotypes for the subjects above. The pipeline is as follows. First, we use the software ‘qctool’ to extract SNPs (imputation score>0.9, MAF>0.01, genotype calling probability>0.7 and biallelic). Second, we use the software PLINK2 to extract subjects with brain traits measured. Third, we use the software ‘PLINK2’ to extract SNPs (MAF>0.01, missing proportion of SNPs or subjects<0.1, Hardy-Weinberg Equilibrium exact test p-value >1e-9, and VIF threshold<5 at a window size of 50 kb and a step size of 5 (variant count)). After the three steps, we finally obtain the SNPs to be used to calculate narrow-sense heritability and conduct QTL mapping.

Typical brain regions: anterior corona radiata (ACR); anterior limb of internal capsule (ALIC); body of corpus callosum (BCC); cerebral peduncle (CP); cingulum cingulate gyrus (CGC); cingulum hippocampus (CGH); corticospinal tract (CST); external capsule (EC); fornix (FX); fornix cres+stria terminalis (Fx/ST); genu of corpus callosum (GCC); inferior cerebellar peduncle (ICP); medial lemniscus (ML); middle cerebellar peduncle (MCP); pontine crossing tract (PCT); posterior corona radiata (PCR); posterior limb of internal capsule (PLIC); posterior thalamic radiation (PTR); retrolenticular part of internal capsule (RLIC); sagittal stratum (SS); splenium of corpus callosum (SCC); superior cerebellar peduncle (SCP); superior corona radiata (SCR); superior fronto-occipital fasciculus (SFO); superior longitudinal fasciculus (SLF); uncinate fasciculus (UNC).

The probability of two traits with the same dimensions

We achieve this estimation by transforming this problem into a geometric model of probability. We will first derive the general expression and then give the closed form in a specific condition.

319 First, consider a general condition where two k -dimensional unit vectors α and β with angle equal
320 to θ share i dimensions in an N -dimensional space, assuming $i \leq k \leq \frac{N+1}{2}$ and $0 < \theta < \frac{\pi}{2}$ without
321 loss of generality. In this condition, the first vector α has a form like $(...a_{l_1}...a_{l_2}...a_{l_3}...a_{l_i}...a_{l_{i+1}}...a_{l_k}...)$
322 which has k terms not equal to zero and the second β has a form like $(...b_{l_1}...b_{l_2}...b_{l_3}...b_{l_i}...b_{s_1}...b_{s_{k-i}}...)$
323 which also has k terms not equal to zero. Both of α and β have i terms not equal to zero in common
324 i dimensions. Thus,

$$325 \quad \cos \theta = \cos \langle \alpha, \beta \rangle = \frac{\alpha \beta}{\|\alpha\| \|\beta\|} = \sum_{j=1}^i a_{l_j} b_{l_j} \quad (2)$$

326 Let $\alpha_i = (a_{l_1}, a_{l_2}, ..., a_{l_i})$, $\beta_i = (b_{l_1}, b_{l_2}, ..., b_{l_i})$, $\|\alpha_i\| = \sqrt{a_{l_1}^2 + ... + a_{l_i}^2} = r_1$ and $\|\beta_i\| = \sqrt{b_{l_1}^2 + ... + b_{l_i}^2} = r_2$.
327 Thus,

$$328 \quad \cos \theta = \alpha_i \beta_i = \|\alpha_i\| \|\beta_i\| \cos \xi = r_1 r_2 \cos \xi, \quad (3)$$

329 where $0 < \xi < \frac{\pi}{2}$ is the angle of α_i and β_i . Due to $0 < \cos \theta < 1$, $0 < r_1 \leq 1$, $0 < r_2 \leq 1$, thus,

$$330 \quad 0 < \cos \theta \leq \cos \xi \leq 1 \quad (4)$$

331 and

$$332 \quad 0 < \cos \theta \leq r_1 \leq 1 \quad (5)$$

333 Set $r_1 = r$, we obtain the geometric distribution of α as

$$334 \quad S_{k,i}(r) = A_i(r) A_{k-i}(\sqrt{1-r^2}), \quad (6)$$

335 where $A_i(r)$ denotes the superficial area of an i -dimensional sphere with radius equal to r ³¹.

336 Because α_i is the projected vector of α in the subspace of β , combining with Eq. (2), the projected

337 vector β' of β on α_i satisfies

$$338 \quad \|\beta'\| = \|\beta_i\| \cos \xi = \frac{\cos \theta}{\|\alpha_i\|} = \frac{\cos \theta}{r} \quad (7)$$

339 and

$$340 \quad \cos \langle \alpha_i, \beta \rangle = \frac{\alpha_i \beta}{\|\alpha_i\| \|\beta\|} = \frac{\|\alpha_i\| \|\beta'\|}{\|\alpha_i\| \|\beta\|} = \frac{\cos \theta}{r} = \|\beta'\|, \quad (8)$$

341 where $\alpha_i \beta = \|\alpha_i\| \|\beta'\|$ and $\|\beta\| = 1$. Therefore, we obtain the geometric distribution of β as

$$342 \quad L_{k,i}(r) = A_{k-1}(\sqrt{1 - \|\beta'\|^2}) = A_{k-1}\left(\sqrt{1 - \frac{\cos^2 \theta}{r^2}}\right) \quad (9)$$

343 Therefore, combining Eq. (5) and Eq. (8), we obtain the geometric estimation for certain α and β

344 given k, i and θ as

$$345 \quad V_{k,i} = \begin{cases} \int_{\cos \theta}^1 S_{k,i}(r) L_{k,i}(r) dr, & i < k \\ A_k(1) A_{k-1}(\sqrt{1 - \cos^2 \theta}), & i = k \end{cases} \quad (10)$$

346 Therefore, the probability of two k -dimensional traits sharing the same dimensions (ψ) in an N -

347 dimensional space given marginal correlation ($\cos \theta$) is estimated as

$$348 \quad Pr_{N,k,\theta}(\Psi) = \frac{C_N^k V_{k,k}}{\sum_{j=k}^{2k-1} C_N^j C_j^{2k-j} C_{2j-2k}^{j-k} V_{k,2k-j}} \quad (11)$$

349 When $\theta \rightarrow 0$, there are $r \rightarrow 1$ and $S_{k,i}(r) \rightarrow 0$. Because $L_{k,i}(r) \leq A_{k-1}(\sqrt{1 - \cos^2 \theta})$, we get

$$350 \quad \frac{V_{k,i}}{V_{k,k}} < \frac{\int_{\cos \theta}^1 S_{k,i}(r) dr}{A_k(1)} \rightarrow 0, (k \neq i, \theta \rightarrow 0) \quad (12)$$

351 Therefore, given N

$$352 \quad Pr_{N,k,\theta}(\Psi) \rightarrow 1, (\theta \rightarrow 0) \quad (13)$$

353 When $N \rightarrow \infty$, there are $\frac{C_N^k}{C_N^j} \rightarrow 0, (j > k)$, therefore, given θ

$$354 \quad Pr_{N,k,\theta}(\Psi) \rightarrow 0, (N \rightarrow \infty) \quad (14)$$

355 Next, consider a specific situation where two 2-dimensional vectors in an N -dimensional space.

356 We can obtain

$$357 \quad Pr_{N,2,R}(\Psi) = \frac{C_N^2(4\pi)}{C_N^2(4\pi) + C_N^3(8 - 8\cos \theta)} = \frac{3\pi}{3\pi + 2(N-2)(1 - \cos \theta)} = \frac{3\pi}{3\pi + 2(N-2)(1 - R)}, \quad (15)$$

358 where $R = \cos \theta$. It is obvious that Eq. (14) still satisfies Eqs. (12-13). Eq. (14) is used to generate

359 the trajectories in Fig. 1d. From (12-13) we can obtain three corollaries:

360 Corollary 1: $Pr(\Psi) \rightarrow 1$ if $R^2 \rightarrow 1$ and $N < N_0$, where N_0 is a finite number.

361 Corollary 2: $Pr(\Psi) \rightarrow 0$ if $N \rightarrow \infty$ and $0 < R^2 < R_u^2$, where $R_u^2 \leq 1$.

362 Corollary 3: $Pr(\Psi) > Pr_0$ if $N < N_0$ and $0 < R^2 < R_u^2$, where $Pr_0 \geq 0$.

Remarks. First, although $Pr_{N,k,\theta}(\Psi)$ represents the probability of sharing the same dimensions, it's hard to achieve linear deduction when few dimension sharing in a very large space for a limited trait sample. Second, the corollary 1 guarantees that linear combinations of G-dimensions can be generated by linear combinations of correlated traits. Corollary 1 combined with corollary 2-3 underlies the success of UBHDD since UBHDD only constrains the correlation between dependent variable and independent variables but not that among independent variables.

Phenotype space simulation

The basic parameters to conduct phenotype space simulation include the number of trait (n), the number of samples or population size ($m=1000$), the number of G-dimensions ($N_1=10, 20, 50, 100$) and the number of NG-dimensions ($N_2=10,000$), the number of G-dimensions per trait ($d_1=N_1/2$), the number of NG-dimensions per trait ($d_2=N_1/2$), broad-sense heritability (H^2) or the variance of genetic component for standardized traits ($H^2 \sim U(0, 1)$). A space can be expressed by a set of bases. When the set of bases are orthogonal, the dimensionality of the space is equal to the number of the set of bases. Assume P^G is independent of P^{NG} . Then, the matrix composed of G-dimensions and NG-dimensions (m by (N_1+N_2)) is randomly generated by standard multivariate normal distribution (R package 'MASS'). Then, the first N_1 column vectors are set to be G-dimensions, the set of orthogonal bases of P^G subspace and the left N_2 column vectors are set to be NG-dimensions, the set of orthogonal bases of P^{NG} subspace. For a focal trait (T_i), the P^G component (T_i^G) is formulated as

$$T_i^G = \sum_j^{N_1} a_j G_j, \quad (16)$$

383 and the P^{NG} component (T_i^{NG}) is formulated as

$$384 \quad T_i^{\text{NG}} = \sum_k^{N_2} b_k NG_k. \quad (17)$$

385 Then, T_i^G and T_i^{NG} are standardized. Finally, the focal trait (T_i) is generated by

$$386 \quad T_i = c_0 T_i^G + (1 - c_0) T_i^{\text{NG}} \quad (18)$$

387 The coefficients a_j and b_k are randomly sampled from normal distribution $N(0, 1)$ and random N_1 -
388 d_1 coefficients of a_j and random N_2 - d_2 coefficients of b_k are set to be zero. The c_0 satisfies

$$389 \quad H_i^2 = \frac{c_0^2}{c_0^2 + (1 - c_0)^2}, \quad (19)$$

390 where H_i^2 is the H^2 of T_i assigned at the parameter setting. For each simulated T_i defined by Eq.
391 (17), we also define a set of correlated traits of T_i as

$$392 \quad T_{i,j} = c_{i,j} T_i^G + (1 - c_{i,j}) T_i^{\text{NG}}, \quad (20)$$

393 where $c_{i,j}$ is the coefficient controlling the j^{th} correlated trait ($T_{i,j}$) of T_i and satisfies

$$394 \quad H_{i,j}^2 = \frac{c_{i,j}^2}{c_{i,j}^2 + (1 - c_{i,j})^2}, \quad (21)$$

395 where $H_{i,j}^2$ is the H^2 corresponding to the j^{th} correlated trait ($T_{i,j}$) of T_i and satisfies

$$396 \quad |H_{i,j}^2 - H_i^2| \leq 0.2 \quad (22)$$

In unstructured population, the number of correlated traits defined for each trait is the same (cluster size equal to 10 in the study). If we set different numbers of correlated traits to different traits, we can simulated a structured population (50 clusters of size 10, a cluster of size 200 and a cluster of size 300 in this study). For $N_1=10, 20, 50$ and 100 , we set $n=1000$ and $m=1000$; for $N_1=100$, we also set $n=2000$ and $m=1000$ and $n=2000$ and $m=2000$ to observe apparent separation between P^G and P^{NG} . Noticeably, d_1 and d_2 are set without loss of generality assuming that either genetic or non-genetic dimensions for a given trait is limited relative to the total non-genetic dimensions ($d_1, d_2 \ll N_2$).

Estimation of broad-sense heritability (H^2) and P^G in yeast

A focal trait is modelled by linear mixed model (LMM) ¹⁸ as

$$T_i = \mu + Z_g u_g + e_g, \quad (23)$$

where T_i is a focal trait, μ is the population mean, Z_g is the design matrix indicating which segregant each replicate belongs to, u_g is a vector of random effect, and e_g is a vector of residuals.

The variance of the focal trait is decomposed into genetic effect (σ_g^2) and environmental effect (σ_e^2). H^2 is then estimated as

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad (24)$$

The random effect estimated for each segregant is defined as P^G . The R package ‘lme4’ is used. Standard error is estimated by Jackknife.

Estimation of narrow-sense heritability (h^2)

416 A focal trait is modelled by LMM¹⁸ as

417
$$T_i = \mu + Z_a u_a + e_a, \quad (25)$$

418 where Z_a is the identity matrix, u_a is a vector of random effect, $u_a \sim N(0, A\sigma_a^2)$, and e_a is a
419 vector of residuals. The variance structure of the trait is formulated as

420
$$\sigma_T^2 = A\sigma_a^2 + I\sigma_e^2, \quad (26)$$

421 where A is estimated as the relatedness matrix, I is the identity matrix, σ_a^2 and σ_e^2 are additive
422 variance and residual variance, respectively. Then, narrow-sense heritability (h^2) is estimated as

423
$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \quad (27)$$

424 The R package ‘rrBLUP’ is used and standard error is estimated by Jackknife for yeast traits. The
425 software GCTA is used for human brain traits.

426 **QTL mapping**

427 In yeast we follow the pipeline used in a previous study³². The association between a focal trait
428 and a focal SNP is calculated as LOD score defined by $-\ln(\ln(1-R^2)/2\ln(10))$, where n is the number
429 of non-missing segregants and R is the Pearson’s correlation. The threshold is determined by 1000
430 times shuffling of segregants. The strongest SNPs larger than the threshold in each chromosome
431 are defined as QTLs. Total two rounds of QTL calling are conducted. The first round is conducted
432 at the original traits and the second round is conducted at the residuals of the original traits. The
433 R package ‘qtl’ is used.

434 In human, we use the widely used software GCTA. The threshold p is set to $2.68\text{e-}8$ equal to
 435 $0.05/n$, where n is the number of total SNPs. Clumping analysis is conducted by PLINK with the
 436 same p .

437 **Uncorrelation-based high-dimensional dependence (UBHDD) modelling**

438 The model is formulated as

$$439 \quad T_i = \sum_{j \in U_i} b_j T_j + \varepsilon_i, \quad (28)$$

440 where T_i is i^{th} trait, b_j is the j^{th} coefficient, ε_i is the residual vector and U_i contains the indices of
 441 uncorrelated traits of T_i . Then, we can obtain the estimated genetic component as

$$442 \quad T_i^g = \sum_{j \in U_i} b_j T_j \quad (29)$$

443 and the estimated non-genetic component as

$$444 \quad T_i^{\text{ng}} = T_i - T_i^g \quad (30)$$

445 In yeast, the 815 samples are divided into a training subset with 715 samples and a testing subset
 446 with 100 samples. Then, a focal trait is modelled in the training subset. The prediction
 447 performance of the learned function was assessed as the R^2 between predicted and observed trait
 448 values in the testing subset. Ten-fold cross validation and LASSO regularization are used to avoid
 449 overfitting (R package ‘glmnet’). Standard error is estimated by 20 repeats. To control for
 450 potential technical bias, we also conduct shuffling analysis for each trait by shuffling T_i before
 451 learning with uncorrelated traits invariable. In brain, the 25957 samples are randomly divided into
 452 10 subsets with equal size, each time a subset is selected as a testing set and the others as a training

set, and a focal trait is modelled in training set. The prediction performance of the learned function is assessed as R^2 between predicted and observed trait values in the testing set. Ten-fold cross validation and LASSO regularization are used to avoid overfitting (Python package ‘glmnet’). Standard error is estimated by ten-fold cross validation. In simulated phenotype space and seg-population, the uncorrelation thresholds are set based on t-test with multiple-testing correction ($p=0.01/(n-1)$ where n is the number of traits). In del-population, the threshold is set the same with that of seg-population to be comparable. In human brain, the threshold is approximately set to 0.15 referring to the threshold in seg-population.

Comparison between UBHDD and PCA in simulated structured population

We first simulated a structured population as above. Then, we apply UBHDD to the structured population and obtain the P^g . Next, we apply PCA to the structured population and keep the top PCs with explained variance up to the mean of those of UBHDD (the mean of variances of T_i^g). Finally, we recalculate the explained variance for each of simulated traits based on the kept top PCs.

Identical Score

To evaluate the consistency of two variables, we can display them in a scatter plot. Then, the variance can be decomposed into two components, one along the straight line $y=x$ and another along the straight line $y=-x$. This is similar to PCA except the transformed coordinate axes are pre-defined. The larger the variance of the component along $y=x$ is, the more consistent the two variables are. The variance of the component along $y=x$ is formulated as

473

$$\sigma_{y=x}^2 = 1 - \frac{\|T_i^{\text{del},f} - T_i^{\text{del},\varphi}\|^2}{2(\|T_i^{\text{del},f}\|^2 + \|T_i^{\text{del},\varphi}\|^2)}, \quad (31)$$

474

where $T_i^{\text{del},f}$ is the genetic component of a focal trait in del-population estimated by the function

475

(f_i) learned in seg-population; $T_i^{\text{del},\varphi}$ is the genetic component of a focal trait in del-population

476

estimated by the function (φ_i) learned in del-population. We name the $\sigma_{y=x}^2$ identity score to

477

evaluate the robustness of P^g obtained by UBHDD between the two distinct yeast populations.

478

Dimensionality estimation by PCA

479

Given a subspace of P , P^g or P^{ng} , we conduct PCA after scaling (Z-score transformation). The

480

number of top PCs with explained variance up to 85% is estimated as the dimensionality of the

481

subspace.

482

483

References

484

1 Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nature reviews. Genetics* **11**, 855-866, doi:10.1038/nrg2897 (2010).

485

486

2 Evangelou, E. *et al.* Impact of phenotype definition on genome-wide association signals: empirical evaluation in human immunodeficiency virus type 1 infection. *American journal of epidemiology* **173**, 1336-1342, doi:10.1093/aje/kwr024 (2011).

487

488

3 Schulze, T. G. & McMahon, F. J. Defining the Phenotype in Human Genetic Studies: Forward Genetics and Reverse Phenotyping. *Human heredity* **58**, 131-138, doi:10.1159/000083539 (2004).

489

490

4 Winawer, M. R. Phenotype definition in epilepsy. *Epilepsy & behavior : E&B* **8**, 462-476, doi:10.1016/j.yebeh.2006.01.012 (2006).

491

492

5 Wojczynski, M. K. & Tiwari, H. K. Definition of Phenotype. *Advances in Genetics* **60**, 75-105, doi:10.1016/s0065-2660(07)00404-x (2008).

493

494

6 Nuttall, F. Q. Body Mass Index: Obesity, BMI, and Health: A Critical Review. *Nutrition today* **50**, 117-128, doi:10.1097/NT.0000000000000092 (2015).

495

496

7 Munoz-Fuentes, V. *et al.* The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conservation genetics* **19**, 995-1005, doi:10.1007/s10592-018-1072-9 (2018).

497

498

499

500

501 8 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*
502 **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

503 9 Seren, U. *et al.* AraPheno: a public database for Arabidopsis thaliana phenotypes. *Nucleic acids*
504 *research* **45**, D1054-D1059, doi:10.1093/nar/gkw986 (2017).

505 10 Ohya, Y. *et al.* High-dimensional and large-scale phenotyping of yeast mutants. *Proceedings of*
506 *the National Academy of Sciences of the United States of America* **102**, 19015-19020,
507 doi:10.1073/pnas.0509436102 (2005).

508 11 Shoal, O. *et al.* Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype
509 space. *Science* **336**, 1157-1160, doi:10.1126/science.1217405 (2012).

510 12 Tenaillon, O. The Utility of Fisher's Geometric Model in Evolutionary Genetics. *Annual review of*
511 *ecology, evolution, and systematics* **45**, 179-201, doi:10.1146/annurev-ecolsys-120213-091846
512 (2014).

513 13 Orr, H. A. Adaptation and the cost of complexity. *Evolution* **54**, 13-20, doi:DOI 10.1111/j.0014-
514 3820.2000.tb00002.x (2000).

515 14 Wang, Z., Liao, B. Y. & Zhang, J. Genomic patterns of pleiotropy and the evolution of complexity.
516 *Proc Natl Acad Sci U S A* **107**, 18034-18039, doi:10.1073/pnas.1004666107 (2010).

517 15 Welch, J. J. & Waxman, D. Modularity and the cost of complexity. *Evolution* **57**, 1723-1734,
518 doi:Doi 10.1554/02-673 (2003).

519 16 Houle, D. Colloquium papers: Numbering the hairs on our heads: the shared challenge and
520 promise of phenomics. *Proceedings of the National Academy of Sciences of the United States of*
521 *America* **107 Suppl 1**, 1793-1799, doi:10.1073/pnas.0906195106 (2010).

522 17 Liu, L. *et al.* The Origin of Additive Genetic Variance Driven by Positive Selection. *Molecular*
523 *biology and evolution* **37**, 2300-2308, doi:10.1093/molbev/msaa085 (2020).

524 18 Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. V. & Kruglyak, L. Finding the sources of
525 missing heritability in a yeast cross. *Nature* **494**, 234-237, doi:10.1038/nature11867 (2013).

526 19 Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective
527 epidemiological study. *Nature neuroscience* **19**, 1523-1536, doi:10.1038/nn.4393 (2016).

528 20 Sha, Z., Schijven, D. & Francks, C. Patterns of brain asymmetry associated with polygenic risks for
529 autism and schizophrenia implicate language and executive functions but not brain
530 masculinization. *Molecular psychiatry*, doi:10.1038/s41380-021-01204-z (2021).

531 21 Sha, Z. *et al.* The genetic architecture of structural left-right asymmetry of the human brain.
532 *Nature human behaviour* **5**, 1226-1239, doi:10.1038/s41562-021-01069-w (2021).

533 22 Goswami, A., Binder, W. J., Meachen, J. & O'Keefe, F. R. The fossil record of phenotypic
534 integration and modularity: A deep-time perspective on developmental and evolutionary
535 dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **112**,
536 4891-4896, doi:10.1073/pnas.1403667112 (2015).

537 23 Klingenberg, C. P. & Marugan-Lobon, J. Evolutionary covariation in geometric morphometric
538 data: analyzing integration, modularity, and allometry in a phylogenetic context. *Systematic*
539 *biology* **62**, 591-610, doi:10.1093/sysbio/syt025 (2013).

540 24 Jones, A. G., Burger, R. & Arnold, S. J. Epistasis and natural selection shape the mutational
541 architecture of complex traits. *Nat Commun* **5**, doi:10.1038/ncomms4709 (2014).

542 25 Williams, E. G. & Auwerx, J. The Convergence of Systems and Reductionist Approaches in
543 Complex Trait Analysis. *Cell* **162**, 23-32, doi:10.1016/j.cell.2015.06.024 (2015).

544 26 Chen, H., Wu, C. I. & He, X. The Genotype-Phenotype Relationships in the Light of Natural
545 Selection. *Mol Biol Evol* **35**, 525-542, doi:10.1093/molbev/msx288 (2018).

- 27 Liu, L., Wang, J., Yang, J. R., Wang, F. & He, X. The expression tractability of biological traits
shaped by natural selection. *J Genet Genomics* **46**, 397-404, doi:10.1016/j.jgg.2019.08.001
(2019).
- 28 Liu, L. *et al.* Decoupling gene functions from knockout effects by evolutionary analyses. *Natl Sci
Rev* **7**, 1169-1180, doi:10.1093/nsr/nwaa079 (2020).
- 29 Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat Methods* **15**, 399-400,
doi:10.1038/s41592-018-0019-x (2018).
- 30 Fan, J., Han, F. & Liu, H. Challenges of Big Data Analysis. *Natl Sci Rev* **1**, 293-314,
doi:10.1093/nsr/nwt032 (2014).
- 31 Li, S. concise formulas for the area and volume of a hyperspherical cap.pdf. *Asian Journal of
Mathematics and Statistics* **4**, 5 (2011).
- 32 Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome
relationship using phenome-wide association studies. *Nature reviews. Genetics* **17**, 129-145,
doi:10.1038/nrg.2015.36 (2016).

Figure legends

Fig. 1. Theory of UBHDD and the supporting simulation

(a,b): In each panel there are three non-parallel and non-perpendicular vectors (α , β , η) shown in a two-dimensional plane defined by the X and Y axes. Vector η can always be expressed as a linear combination of vectors α and β no matter whether the angle θ_1 (θ_2) between η and α (β) is small (i.e., correlated) as in panel a, or close to $\frac{\pi}{2}$ (i.e., uncorrelated) as in panel b, where $0 < \theta_1, \theta_2 < \frac{\pi}{2}$. Note that, with two traits represented by two vectors, the marginal correlation (Pearson's R) between the two traits is equal to the cosine of the angle (θ) between the two vectors (i.e., $R = \cos\theta$).

(c): Three non-parallel and non-perpendicular vectors in a three-dimensional space are shown, in which η is nearly perpendicular to the XY-plane and thus uncorrelated to α and β . In contrast to panel b, here η can no longer be expressed as a linear combination of α and β because η owns a unique dimension (Z-axis).

(d): Uncorrelated vectors have a moderate probability of sharing the same dimensions in a space of low dimensionality. The probability of sharing dimensions depends the dimensionality of space (N), the dimensionality of each trait (k), and the marginal correlation level (Pearson's R^2) of the two vectors. Based on a general geometric deduction (Methods), the probability would converge to one when R^2 converges to one, and to zero when N converges to infinity. Here the probability trajectories as a function of R^2 are shown for $N=10$, 100 and 1000, respectively, with $k=2$ for both vectors. For a very small R^2 the probability approaches zero for $N=1000$ but remains a moderate level for $N=10$.

(e): In the simulated phenotype space the number of obtained dimensions is saturated much more rapidly in P^G ($N_1=10$) than in P^{NG} ($N_2=10,000$) as a function of trait sampling. Error bar represents the 95% confidence interval estimated by 100 replicates of trait sampling.

(f): The probability of sharing P^G dimensions remains constantly high for traits with various levels of marginal correlation. In contrast, the probability of sharing P^{NG} dimensions rapidly converges to zero for traits of low correlation. The threshold $R_u \approx 0.15$ (0.147) is used for defining uncorrelated traits in the simulated phenotype space, which corresponds to $p = 0.01$ after correction for multiple testing. Error bar shows the standard error.

(g): The T^S estimated by UBHDD is highly correlated with the actual genetic component T^G of the simulated traits. As a control, the correlation between T^{ng} and T^G is also shown. A total of 1,000 traits are examined and standard box plots are used to display the data. The p -value is computed by paired t-test.

(h): The variance of T^S in the simulated population is similar to the variance of T^G , the actual broad-sense heritability (H^2) of a trait in the population. Each dot represents a simulated trait and a total of 1,000 traits are shown.

Fig. 2. Separation of T^S from T^{ng} by UBHDD for 405 yeast traits.

610 **(a):** A summary of the yeast phenome data. The yeast segregant population is generated by a cross
611 of two *S. cerevisiae* strains (BY and RM). A total of 405 phenotypic traits are characterized for
612 each of 815 segregants, with two clones examined for each segregant.

613 **(b):** A schematic diagram of a yeast cell with landmarks for describing the shape or position of the
614 cell wall and nuclei of the mother and daughter cells.

615 **(c):** Two strategies used for separating the genetic component from the non-genetic component of
616 a quantitative trait. The genotype-based linear mixed model (LMM) is a classical strategy, and
617 the resulting components are denoted as T^G and T^{NG} . The phenome-based UBHDD, which requires
618 no genotype information, is proposed in this study; the resulting components are denoted as T^S and
619 T^{ng} .

620 **(d-e):** The T^S estimated by UBHDD is highly similar to the T^G derived from LMM. As a control,
621 T^{ng} is often distinct from T^G . The panels d shows the details of a randomly selected trait (C11.1_A),
622 and the panel e shows the summary statistics (Pearson's R^2) of the 405 traits.

623 **(f):** Substantial trait variance is captured by T^S . The inset shows the results after randomly
624 shuffling the segregants. Note that the total variance of a trait is one.

625 **(g):** The variance of T^S is similar to the variance of T^G , the broad-sense heritability estimated by
626 LMM.

627 **(h):** The narrow-sense heritability (h^2) of T^S is generally larger than that of T^{ng} . Each dot represents
628 a trait and 405 yeast traits are examined.

629 **(i):** There are more QTLs detected in T^S than T^{ng} . Each dot represents a trait and 405 traits are
630 examined (with 20 traits on the line of $y=x$). The dots are plotted with jitter for visual
631 distinguishability.

632 **(j):** For a given trait the T^S function learned in the segregant population can be compared with the
633 T^S function learned in another yeast population comprising 4,718 gene deletion strains, which
634 assesses the robustness of UBHDD.

635 **(k):** For a randomly selected trait C11.1_A, the two functions produce highly similar T^S in the gene
636 deletion strains, with an identity score = 0.88. The identity score is defined as the variance
637 component along $y=x$ in the scatter plot. Each dot represents a gene deletion strain and 4,718
638 strains are examined.

639 **(l):** Density distribution of the identity scores of the 405 yeast traits.

640
Fig. 3. Separation of T^S from T^{ng} by UBHDD for 675 human brain traits.

641 **(a):** A summary of the human brain phenome data. Here the typical brain regions in the classical
642 brain region atlas of Johns Hopkins (used in UK biobank) are shown. For example, ACR is short
643 for anterior corona radiate and others are listed in Methods.
644

(b): Substantial trait variance is captured by T^g . The inset shows the results after randomly shuffling the individuals. Note that the total variance of a trait is one.

(c): The h^2 of T^g is generally larger than that of T^{ng} . For saving computational time we examined, instead of all 675 traits, only the 107 exemplar traits obtained using the R package ‘apcluster’. Each dot represents an exemplar trait.

(d): There are more QTLs detected in T^g than T^{ng} . Each dot represents an exemplar trait (with 15 traits on the line of $y=x$).

(e): The left-right similarity is invariably stronger in T^g than T^{ng} , indicating non-genetic factors be the major source of brain left-right asymmetry. The similarity is measured by Pearson’s R^2 between a symmetrical trait pair. A total of 297 trait pairs are examined.

Fig. 4. Distinct dimensionality of P^g and P^{ng} .

(a-b): The number of latent dimensions (n_D) in P^g and P^{ng} , respectively, as a function of the number of sampled traits (n_T) for yeast (a) and human brain (b). The n_D is estimated as the number of top principal components that explain 85% variance of the sampled traits. Error bar represents the 95% confidence interval estimated by 100 replicates of trait sampling.

(c-d): The number of additional traits per additional dimension ($\Delta n_T / \Delta n_D$) is constantly high in P^g but small in P^{ng} for both yeast (c) and human brain (d). This suggests P^g dimensions be recurrently used to code the traits while P^{ng} dimensions tend to be trait-specific.

(e-f): The genetic component (T^g) of the traits often show correlation due to the common P^g dimensions. The Pearson’s R of T^g for all trait pairs is shown for yeast and human brain phenome, respectively. The two vertical red lines mark $-0.1 < R < 0.1$.

(g-h): The non-genetic component (T^{ng}) of the traits shows little correlation, echoing the fact that P^{ng} dimensions are trait-specific. The Pearson’s R of T^{ng} for all trait pairs is shown for yeast and human brain phenome, respectively. The two vertical red lines mark $-0.1 < R < 0.1$.

Fig. 5. A phenome-based strategy (UBHDD) is proposed to decompose the genetic and non-genetic components of quantitative traits, which complements the kinship- or genotype-based conventional strategy.

Fig. 1

Fig. 2

Fig. 3

Fig. 4

Fig. 5

Fig. S1. Decomposition of P^G and P^{NG} by UBHDD in simulated phenotype spaces.

(a) The dimensionality of P^G subspace (N_1) = 20, population size (m) = 1000 and the number of traits (n) = 1000. (b) N_1 = 50, m = 1000 and n = 1000. (c) N_1 = 100, m = 1000 and n = 1000. (d) N_1 = 100, m = 1000 and n = 2000. (e) N_1 = 100, m = 2000 and n = 2000. (c) shows a poor performance but can be improved with increasing number of traits (d) and further achieves higher performance with the increasing of population size (e).

Fig. S2. Comparison between UBHDD and PCA in structured simulated phenotype space.

(a) shows the structure of the simulated phenotype space shown by PCA. Each point represents a trait. Out of the 1000 traits, 500 form two larger correlated clusters (n=300, 200, respectively) and the others form 50 smaller clusters each of which consists of 10 correlated traits. (b) shows the successful estimation of genetic component by UBHDD. (c) shows the failure of PCA to separate the genetic component. The top two PCs with the same 50% variance explained in PCA, overfit the 500 traits in the two larger clusters but underfit the 500 traits in the 50 smaller clusters.

Fig. S3. Broad-sense heritability (H^2) and narrow-sense heritability (h^2) of 405 yeast traits.

Fig. S4. Robust estimation of T^g by UBHDD under different uncorrelation thresholds (R_u) in yeast. The threshold 0.147 corresponds to $p=0.01$ with Bonferroni correction in yeast seg-population. The estimated genetic variance is robust to the threshold used to conduct UBHDD.

Fig. S5. Robust estimation of T^g by UBHDD under different uncorrelation thresholds (R_u) in human brain. The threshold 0.15 is used in human brain phenotype space. The estimated genetic variance is robust to the threshold used to conduct UBHDD.

Fig. S6. Evaluation for the number of traits saturated with G-dimensions in phenotype spaces.

The accurate separation of genetic and non-genetic subspaces not only depends on rational uncorrelation threshold but also enough trait sampling. We conduct the same learning process for different proportions of trait subsets from 10% to 100%, say, a down-sampling strategy. Then, the distribution of UBHDD performance (R^2 , the variance of genetic component estimated by UBHDD) is compared among these trait subsets. (a) shows the distributions of yeast. (b) shows the distributions of human brain.

Fig. S7. Criterion for uncorrelation threshold (R_u).

For a focal trait (T_i), we can learn a linear function built on its uncorrelated traits (T_j): $T_i = \sum b_j T_j$. To judge the uncorrelation threshold (R_u), we provide a statistical test as follows. First, we calculate the square of marginal correlation (MC^2) between the focal trait and each of its uncorrelated traits. Then, we calculate the square of coefficients (Co^2) for each uncorrelated trait in the learned linear function, say, b_j^2 . An optimal threshold is determined if the R^2 between MC^2 and Co^2 is insignificant, meanwhile, taking the number of uncorrelated traits available into account. (a-b) are results of UBHDD model. (a) shows the result under the R_u used in this study for an example trait in yeast. (b) shows the results for all of the 405 traits in yeast. As a contrast, we also learned linear functions based on total traits (Total model) for each of the 405 traits in yeast. (c-d) are results of Total model. (c) shows the same example trait in yeast based on Total model. (d) shows the results for all of the 405 traits in yeast based on Total model. Similarly, the results in human brain are shown for UBHDD model (e-f) and Total model (g-h). The red line denotes the adjusted $p=0.01$ to the number of traits.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [rs.pdf](#)