OXFORD GENETICS

GENETICS, 2023, 225(1), iyad127
https://doi.org/10.1093/genetics/iyad127
Advance Access Publication Date: 12 July 2023
Investigation

Inferring sparse structure in genotype–phenotype maps

Samantha Petti,[1,†] Gautam Reddy,[1,2,3,†] Michael M. Desai[4,*]

[1]NSF-Simons Center for the Mathematical and Statistical Analysis of Biology, Harvard University, Cambridge, MA 02138, USA
[2]Physics & Informatics Laboratories, NTT Research, Inc., Sunnyvale, CA 94085, USA
[3]Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[4]Department of Organismic and Evolutionary Biology and Department of Physics, Harvard University, Cambridge, MA 02138, USA

*Corresponding author: Department of Organismic and Evolutionary Biology and Department of Physics, Harvard University, 52 Oxford Street, Cambridge, MA 02138, USA. Email: desai@oeb.harvard.edu
[†]These authors contributed equally to this work.

Abstract

Correlation among multiple phenotypes across related individuals may reflect some pattern of shared genetic architecture: individual genetic loci affect multiple phenotypes (an effect known as pleiotropy), creating observable relationships between phenotypes. A natural hypothesis is that pleiotropic effects reflect a relatively small set of common "core" cellular processes: each genetic locus affects one or a few core processes, and these core processes in turn determine the observed phenotypes. Here, we propose a method to infer such structure in genotype–phenotype data. Our approach, *sparse structure discovery* (SSD) is based on a penalized matrix decomposition designed to identify latent structure that is low-dimensional (many fewer core processes than phenotypes and genetic loci), locus-sparse (each locus affects few core processes), and/or phenotype-sparse (each phenotype is influenced by few core processes). Our use of sparsity as a guide in the matrix decomposition is motivated by the results of a novel empirical test indicating evidence of sparse structure in several recent genotype–phenotype datasets. First, we use synthetic data to show that our SSD approach can accurately recover core processes if each genetic locus affects few core processes or if each phenotype is affected by few core processes. Next, we apply the method to three datasets spanning adaptive mutations in yeast, genotoxin robustness assay in human cell lines, and genetic loci identified from a yeast cross, and evaluate the biological plausibility of the core process identified. More generally, we propose sparsity as a guiding prior for resolving latent structure in empirical genotype–phenotype maps.

Keywords: genotype–phenotype map, penalized matrix decomposition, structure discovery, sparsity

Introduction

A central goal of quantitative genetics is to exploit observed correlations between genotype and phenotype to infer the structure of the genotype–phenotype map (Rockman 2008; Wagner and Zhang 2011; Paaby and Rockman 2013; Solovieff *et al.* 2013; Davey Smith and Hemani 2014; Haworth *et al.* 2019). That is, we aim to build models describing how variation in genotype influences variation in phenotype. However, the choice of phenotypes quantitative geneticists choose to analyze is inherently subjective: we typically focus on phenotypes that are practical to measure and/or that are in some sense "important" (e.g. because they are plausibly related to key functions or diseases). These phenotypes are often correlated, presumably because multiple complex traits are often influenced by the same set of core cellular processes. For example, cellular growth rates across a range of different stressful conditions may be determined by a common set of processes such as metabolism, cell wall biosynthesis, DNA repair, and heat or osmotic stress response. This leads to apparent widespread pleiotropy, where individual genetic loci influence many observed phenotypes, presumably because these loci influence one or more core processes that are broadly important across multiple phenotypes.

This perspective suggests that the structure of the correlations between the subjective phenotypes that we choose to measure should contain signatures of the underlying biologically relevant core processes. That is, if we could measure a large and diverse enough set of phenotypes across a sufficiently diverse range of genotypes, the observed phenotypic variation should have a lower-dimensional latent structure that reflects the space of actual core processes. Inferring this lower-dimensional latent structure thus offers the promise of explaining the biological basis of pleiotropy, by identifying the core biological processes and inferring how individual loci influence these core processes to generate the observed phenotypic variation.

Of course, we can only hope to identify core processes which generate variation across the phenotypes we choose to measure, so the core processes we infer will always be limited by this choice. For example, imagine that we measure a set of phenotypes that correspond to the growth rates of yeast cells across a temperature gradient. We might expect that these phenotypes exhibit a correlation structure that reflects three core processes: heat-shock response, cold tolerance, and all other temperature-independent factors relevant to the common growth medium. We could then hope to infer the extent to which each genetic locus influences each of the core processes, as well as the mapping between these three core processes and the observed phenotypes. However, if we were to measure additional phenotypes corresponding to growth rates across (for example) different nutrient concentrations, we

Received: May 24, 2023. **Accepted:** June 21, 2023

might find that this splits the temperature-independent core process into additional processes that explain the variation in the new phenotypes.

In this manuscript, we describe how to infer this lower-dimensional latent structure of phenotype space using a penalized matrix decomposition framework (Witten *et al.* 2009). We assume that we have data that describes the map between genotype and some set of measured phenotypes. In general, this genotype–phenotype map can involve nonlinear effects such as interactions between multiple genetic loci (epistasis). However, we focus here on analyzing a standard linear approximation of this map, in which each locus is assumed to have an additive effect on each of the phenotypes, and the observed phenotype is simply a sum of the additive effects of all the relevant loci. This linear map can be represented as an $E \times L$ matrix, $\mathbf{F}$, which has columns corresponding to each of the L loci and rows corresponding to the effect of these loci on the E measured phenotypes. We note that inferring $\mathbf{F}$ from data on genotypes and corresponding phenotypes can be a complex problem, which we address for one example dataset below, but the core of our analysis in this paper assumes that $\mathbf{F}$ is given and focuses on analyzing the latent structure in this matrix.

In this framework, our problem reduces to inferring lower-dimensional structure in the matrix $\mathbf{F}$. While in principle this structure could be nonlinear, we restrict ourselves to inferring a lower-dimensional subspace that can be expressed as a matrix decomposition of $\mathbf{F}$. Specifically, we wish to approximate $\mathbf{F}$ as the product of two matrices, $\mathbf{F} \approx \mathbf{WM} + \mathbf{b}$, where $\mathbf{M}$ is a $K \times L$ matrix that describes the additive effect of each genetic locus on each of K putative core processes, and $\mathbf{W}$ is an $E \times K$ matrix that describes how each core process affects each measured phenotype. In addition, we include a L dimensional vector $\mathbf{b}$ which represents locus-specific effects on all other processes that contribute equally to the phenotypes measured (and hence cannot be disentangled). For ease of notation, we take $\mathbf{WM} + \mathbf{b}$ to mean that $\mathbf{b}$ is added to each row of $\mathbf{WM}$, i.e. $\mathbf{WM} + \mathbb{1}\mathbf{b}^T$, where $\mathbb{1}$ is the all ones vector of length E. For $K < E, L$, this represents an approximation to $\mathbf{F}$ in terms of a lower-dimensional subspace of K core processes. This structure is illustrated in Fig. 1a. We emphasize that this decomposition assumes that the map between loci and core processes and the map between core processes and measured phenotypes are both linear, which may not be true in general. We return to this caveat in the "*Discussion*".

Unfortunately, this matrix decomposition problem is underdetermined in general, meaning that for any choice of K there are many different pairs of matrices $\mathbf{W}$ and $\mathbf{M}$ that approximate $\mathbf{F}$ equally well. Thus, the fact that a given decomposition gives a good approximation for $\mathbf{F}$ does not necessarily imply that there is any biological meaning to the core processes inferred. This problem is widely recognized in a variety of fields where matrix decomposition is used to infer lower-dimensional structure in high-dimensional data. To make lower-dimensional structure interpretable, domain-specific knowledge must therefore be used to guide the choice of additional constraints, and optimization algorithms must be developed to efficiently find decompositions that obey the constraints. For example, earlier work has used sparsity (Olshausen and Field 1997; Zhang *et al.* 2015), nonnegativity (Paatero and Tapper 1994; Lee and Seung 2000; Wang and Zhang 2012) and non-Gaussianity assumptions (Jutten and Herault 1991; Comon 1994; Hyvärinen and Oja 2000) to construct powerful methods for identifying meaningful latent structure in specific contexts where those constraints are appropriate. The success of these approaches motivates our attempt here to find appropriate constraints that enable the efficient and interpretable reconstruction of a lower-dimensional set of core processes from empirical genotype–phenotype maps. Such constraints can be thought of as incorporating a biological "prior" on the features we expect the data to exhibit.

Recently, Kinsler *et al.* (2020) identified lower-dimensional structure in a dataset describing the effects of a set of yeast mutations on fitness in different environments. Their approach used singular value decomposition (SVD) (Golub and Reinsch 1971) to find a decomposition with $K < E, L$ that approximates the $\mathbf{F}$ well. However, while SVD finds the K-dimensional subspace that explains the most variation for a given K, the specific $\mathbf{W}$ and $\mathbf{M}$ are selected subject to the constraints that the core processes must be orthogonal and that the first j core processes describe the j-dimensional subspace that best approximates $\mathbf{F}$. It is not clear that these constraints lead to putative core processes with biological meaning. More recently, Pan *et al.* (2022) introduced an alternative matrix decomposition method, Webster, which is based on regularized dictionary learning (Yankelevsky and Elad 2016), and apply it to a dataset describing the fitness of cells exhibiting gene knockouts in the presence of various genotoxins (Olivieri *et al.* 2020). This method enforces a hard constraint that each genetic locus affects at most two core processes, which limits the possibility that different loci exhibit different degrees of pleiotropy.

Here, we present a matrix decomposition approach based on the biologically motivated intuition that the lower-dimensional structure of the genotype–phenotype map may be sparse. Specifically, our *sparse structure discovery* (SSD) method finds decompositions where each genetic locus affects a small subset of the core processes (locus-sparsity) and/or each observed phenotype is influenced by a small subset of core processes (phenotype-sparsity). SSD is an example of penalized matrix decomposition, a broad class of matrix decomposition methods that encourage the matrix factors to exhibit particular properties (e.g. sparsity) through hard constraints or regularization (Witten *et al.* 2009). In our case, SSD has separate tunable regularization parameters to control the extent to which locus-sparsity and phenotype-sparsity are encouraged.

The sparsity assumption is consistent with various notions of modularity which have been proposed to explain the evolvability of complex traits (Altenberg 2005; Wagner *et al.* 2007; Crombach and Hogeweg 2008; Hintze and Adami 2008; Wagner and Zhang 2011; Clune *et al.* 2013), and with large-scale studies of pairwise gene deletions in yeast, which find that genes cluster together based on their interaction profiles, suggesting their involvement in a small set of common core processes (Costanzo *et al.* 2010). However, given a matrix $\mathbf{F}$ representing a genotype–phenotype map, it is not a priori clear the extent to which the data exhibits locus-sparsity and/or phenotype-sparsity, making it difficult to justify a specific choice of regularization parameters. To address this, we run our SSD method across a grid of regularization values to obtain an array of decompositions with varying errors, locus-sparsities, and phenotype-sparsities and allow the user to manually select a decomposition of interest for further investigation. Moreover, we have developed two empirical tests to independently validate the extent to which the lower-dimensional structure in an effects matrix $\mathbf{F}$ exhibits locus-sparsity or phenotype-sparsity, the results of which can then be used to guide the selection of SSD decompositions (Fig. 1c). Using these tests, we find evidence of locus-sparsity and phenotype-sparsity across three datasets, motivating the use of these sparsity-enforcing penalties in our SSD method. Further, we show that SSD accurately recovers synthetically generated maps if at least one of the true $\mathbf{W}$ or $\mathbf{M}$ is sparse.

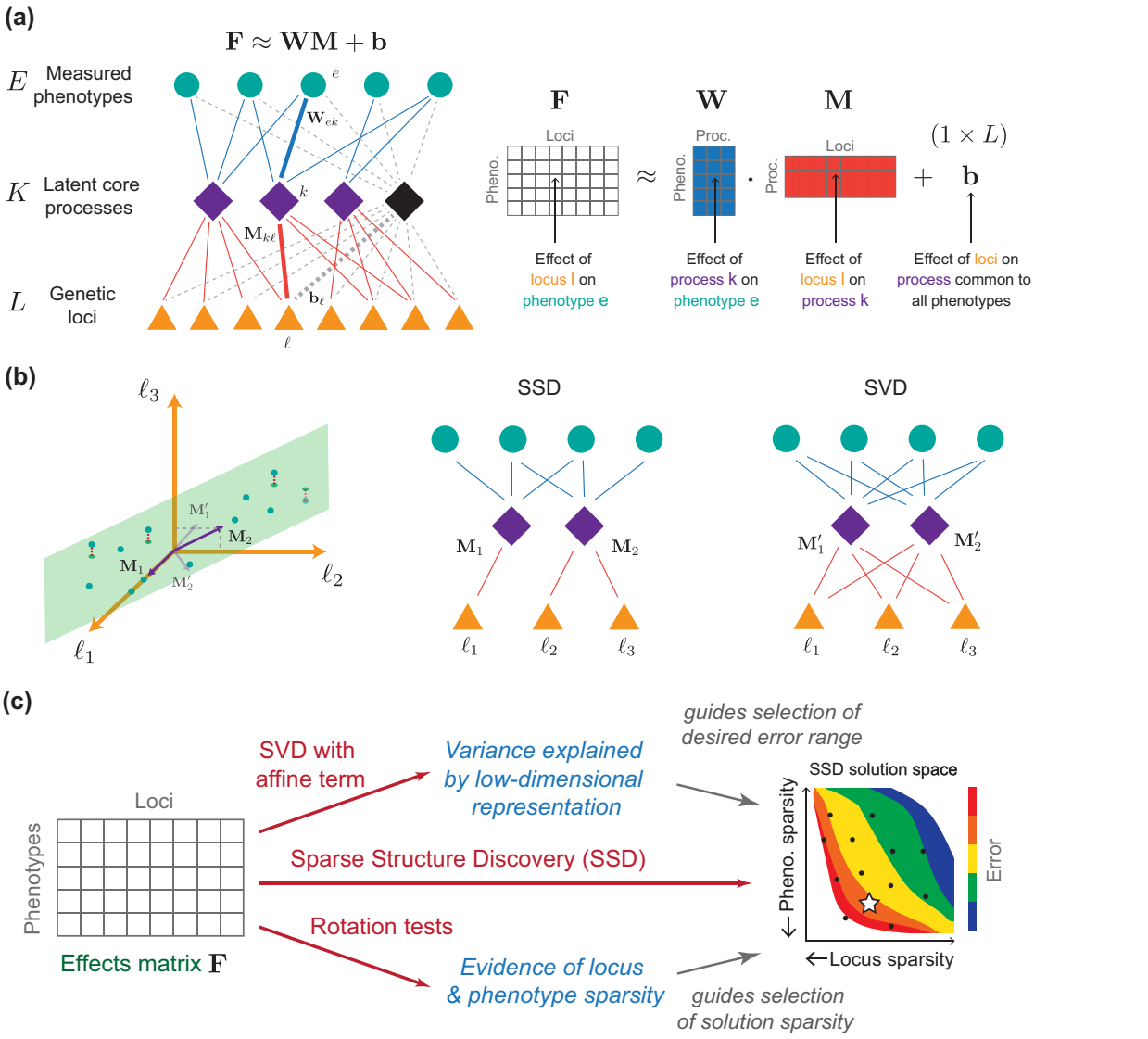

Fig. 1. Overview and geometric interpretation of sparse structure discovery (SSD). a) SSD finds a sparse, low-rank approximation for the effects matrix **F** containing the phenotypic effects of L loci on E phenotypes. b) Each phenotype (row of **F**) can be viewed as a point in locus-space. The core processes (rows of **M**) can be viewed as vectors that span a lower-dimensional subspace, illustrated by the plane. The distances between each phenotype point and the subspace determine the reconstruction error. Since the error is a function of the subspace and there are many matrices **M** which generate the same subspace, many decompositions yield the same error. SSD applied to these phenotypes would favor a sparse decomposition, for example, the core processes $\mathbf{M}_1$, $\mathbf{M}_2$ which here are sparse combinations of (ℓ_1), (ℓ_2, ℓ_3) respectively. Singular value decomposition (SVD) applied to the same phenotypes would yield a decomposition with core processes $\mathbf{M}'_1$, $\mathbf{M}'_2$ that incur the least error but which are unlikely to be sparse. c) In our analysis pipeline, we first apply SSD to find a range of decompositions $\mathbf{F} \approx \mathbf{WM} + \mathbf{b}$ with varying errors and sparsities. The reconstruction error of the SVD solution is used to determine a tolerable error range for SSD solutions. The rotation tests are used to guide the selection of an SSD solution with appropriate levels of sparsity in phenotypes (each phenotype is described by few core processes) and in the loci (each locus is part of few core processes).

The structure of the paper is as follows. In "*Sparse structure discovery*," we describe the SSD method, explain our empirical tests for sparsity, and demonstrate that SSD accurately recovers core processes in synthetic data. In "*Applications to empirical data*," we apply our method to three datasets that measure cellular fitness across environments as a function of three different forms of genetic variability. First, we apply SSD to the Kinsler *et al.* (2020) dataset describing fitness effects of adaptive mutations identified during a laboratory yeast evolution experiment and compare SSD to the SVD-based analysis presented in Kinsler *et al.* (2020). Second, we apply SSD to data describing how single gene knockouts in human cell lines affect fitness in the presence of genotoxic agents (Olivieri *et al.* 2020). We find that, compared to the Webster analysis of the same dataset (Pan *et al.* 2022), SSD solutions exhibit lower error with comparable average sparsity, a more interpretable process-phenotypes map, and a broad range of pleiotropy across loci. Third, we analyze a large-scale quantitative trait locus (QTL) mapping experiment (Ba *et al.* 2022), which measured 18 growth rate phenotypes in about 100,000 F1 offspring of a cross between two related budding yeast strains. For this data, we first develop a joint mapping approach to arrive at an additive effects matrix **F**, which we do using a pipeline based on $\ell_{2,1}$-penalized regression (see Supplementary Material).

Sparse structure discovery

As described above, our method assumes we begin with an empirical linear genotype–phenotype map, represented as an $E \times L$

matrix $\mathbf{F}$ which describes the additive effect of each of the L genetic loci on each of the E measured phenotypes. Our goal is to find latent structure in this genotype–phenotype map of the form $\mathbf{F} \approx \mathbf{WM} + \mathbf{b}$. Note that since we will generally assume that $K < E, L$, the matrices $\mathbf{W}$ and $\mathbf{M}$ contain fewer total parameters than $\mathbf{F}$ (i.e. this is a simpler description of the data). Thus, this factorization will in general only be an approximation, both because there is presumably error in the estimation of $\mathbf{F}$ and because the division into K core processes is a simplifying assumption that will inevitably neglect some aspects of the full complexity underlying each measured phenotype.

Given that the factorization of $\mathbf{F}$ is approximate, a natural goal would be to find matrices $\mathbf{W}$ and $\mathbf{M}$ that minimize the error in this approximation. This is the motivation underlying SVD, which finds a factorization of $\mathbf{F}$ that minimizes the squared Frobenius reconstruction error (i.e. lowest squared error $\|\mathbf{F} - \mathbf{WM}\|_2^2$). However, this error minimization alone is not sufficient to uniquely determine the factorization. Instead, any factorization that describes the same lower-dimensional subspace will perform equally well, as illustrated in Fig. 1b. This is a general problem: for any set of core processes, represented by the rows of $\mathbf{M}$, that achieve a given reconstruction error, there are infinitely many sets of other processes that achieve the same error (obtained by changing the basis of the subspace, e.g. by rotating the rows of $\mathbf{M}$ in the subspace they generate). SVD chooses a particular unique solution to resolve this degeneracy by defining the first core process to be the one-dimensional subspace that minimizes the error for $K = 1$, the second core process to be orthogonal to the first and minimize the error for $K = 2$, the third to be orthogonal to the first two and minimize the error for $K = 3$, and so on. While this is a reasonable and well-defined procedure, there is no reason to believe that the core processes defined in this way will be biologically meaningful.

Here, we define an alternative method for matrix decomposition. Like SVD, our approach attempts to minimize the Frobenius reconstruction error. However, we add two additional constraints based on *sparsity*. Specifically, we aim to find a locus to core process map $\mathbf{M}$ in which each locus participates in only a few processes (i.e. most entries in this matrix are 0). We refer to this as locus-sparsity. Analogously, we aim to find a core process to phenotype map $\mathbf{W}$ in which each phenotype is affected by only a few core processes (i.e. most entries in this matrix are also 0). We refer to this as phenotype-sparsity.

We do not necessarily assume that both types of sparsity exist in a given dataset. Instead, our framework allows us to impose constraints on either or both types with a tunable stringency (and below we describe how the choice of this stringency can be guided by empirical validation tests). To be precise, our *SSD* method aims to find the matrix decomposition $\mathbf{F} \approx \mathbf{WM} + \mathbf{b}$ that minimizes

$$\mathcal{C}(\mathbf{W}, \mathbf{M}, \mathbf{b}) = \|\mathbf{F} - (\mathbf{WM} + \mathbf{b})\|_2^2 + \lambda_W \|\mathbf{W}\|_1 + \lambda_M \|\mathbf{M}\|_1 \quad \text{such that } \|\mathbf{M}_{k,:}\|_2 = 1 \text{ for all } 1 \le k \le K_{\max}, \tag{1}$$

where $\|\mathbf{F} - (\mathbf{WM} + \mathbf{b})\|_2^2$ is the squared Frobenius error, $\|\mathbf{W}\|_1$ is an ℓ_1-norm measure of the phenotype-sparsity, and $\|\mathbf{M}\|_1$ is an ℓ_1-norm measure of the locus-sparsity. Equation (1) is a variant of the penalized matrix decomposition formulations studied in Witten *et al.* (2009). The parameters λ_W and λ_M determine the relative weighting of the accuracy, phenotype-sparsity, and locus-sparsity objectives (higher λ_W will yield solutions that are more phenotype-sparse, and higher λ_M will yield solutions that are more locus-sparse). We note that when these regularization parameters λ_W and λ_M are sufficiently large, the method will assign no loci to some of the core processes, thereby automatically picking a number of core processes K smaller than the input upper bound $K_{\max}$. We include the constraint requiring that the Euclidean norm of each core process is one to ensure that the core processes are all of the same scale; the magnitude of the vectors of $\mathbf{W}$ may vary, reflecting that some phenotypes are more sensitive to the core processes than others.

For fixed values λ_W, λ_M, and $K_{\max}$, SSD will yield a unique set of $\mathbf{W}$, $\mathbf{M}$, and $\mathbf{b}$. However, a key challenge is to choose values of these parameters to determine an appropriate weighting of the accuracy, phenotype-sparsity, and locus-sparsity objectives that will produce a decomposition with plausible biological meaning. To do so, we first apply our method for a range of values λ_W and λ_M to produce a variety of decompositions that vary in reconstruction error, number of core processes, locus-sparsity, and phenotype-sparsity. In every case, the SSD decomposition will have higher reconstruction error than the SVD decomposition with the same number of processes because of the additional constraints. We therefore use the SVD error as a guide to select a desired reconstruction error range, and select sparse decompositions of interest that fall within this range. The choice between these can then be guided by the empirical test described below, which we developed to determine the extent to which an input matrix $\mathbf{F}$ exhibits a low-dimensional structure with locus-sparsity or phenotype-sparsity. Fig. 1c illustrates the pipeline.

Empirical validation of sparsity constraints using rotation tests

To validate our choice of sparsity assumptions, we designed heuristic tests to determine whether a given dataset $\mathbf{F}$ exhibits signatures of locus-sparsity or phenotype-sparsity. We do not assume that the linear term $\mathbf{b}$, which describes the effects of loci on processes that do not vary across the phenotypes, is necessarily sparse. For the purposes of this test, we therefore first subtract the mean effect across phenotypes for each locus from $\mathbf{F}$, as an approximation of $\mathbf{b}$. To test for locus-sparsity, we then apply a random orthogonal matrix $\mathbf{O}$ to the empirical genotype–phenotype map $\mathbf{F}$ to produce a matrix $\mathbf{F}' = \mathbf{FO}$. This rotation conserves low-dimensional structure in $\mathbf{F}$ and leads to the same SVD error but disrupts any potential locus-sparsity. We then apply our SSD method with a range of weights on the locus-sparsity objective to obtain a range of decompositions for $\mathbf{F}$ and $\mathbf{F}'$ that exhibit varying locus-sparsities and reconstruction errors. If the input matrix $\mathbf{F}$ truly has locus-sparsity, our method will consistently find sparser solutions for $\mathbf{F}$ than for $\mathbf{F}'$ across a range of reconstruction errors. If so, we consider this to be evidence of locus-sparsity in $\mathbf{F}$. In practice, we use the shape of the sparsity-error curve for F as a heuristic for selecting a sparse solution of interest. For example, if the curve exhibits an elbow shape (i.e. below some sparsity s the error drops quickly with decreasing sparsity and above s the error increases slowly with increasing sparsity), this may indicate that a reasonable choice of sparsity is s.

To gain intuition for this test, consider an example with five loci and two core processes, with loci ℓ_1 and ℓ_2 both affecting core process 1 (with equal weight) and loci ℓ_3, ℓ_4 and ℓ_5 all affecting core process 2 (also with equal weight). The rows of the matrix $\mathbf{F}$ will each have the form $(\alpha, \alpha, \beta, \beta, \beta)$, where α and β describe the effect of the first and second processes on the phenotype corresponding to that row, respectively. In other words, the phenotype values lie on a 2D plane in 5D space. This plane contains the sparse vectors (1, 1, 0, 0, 0) and (0, 0, 1, 1, 1), which describe the two core processes, and every point on the plane can be written as a weighted

sum of these vectors. Now, imagine that we randomly rotate $\mathbf{F}$, producing a matrix $\mathbf{F}'$ which has rows that lie on a rotation of the 2D plane containing the rows of $\mathbf{F}$ and columns that correspond to random linear combinations of the actual genetic loci. Since the rotation was random, the 2D plane containing the rows of $\mathbf{F}'$ is a random 2D plane in 5D. Most 2D planes in 5D are not spanned by two sparse basis vectors. Therefore, while it is still possible to find two vectors such that each row of $\mathbf{F}'$ can be written as the weighted sum of these vectors (the low-dimensional structure is preserved), the two vectors almost certainly will not be sparse.

To test for phenotype-sparsity, we use an analogous method, except that we rotate the columns of $\mathbf{F}$ to obtain $\mathbf{F}' = \mathbf{OF}$ and vary the phenotype-sparsity objective in SSD to test whether SSD consistently finds sparser solutions for $\mathbf{F}$ than $\mathbf{F}'$ across a range of reconstruction errors.

Sparse structure recovery on synthetic data

To validate our method, we constructed synthetic genotype–phenotype maps with lower-dimensional latent structure of varying sparsity. That is, for a given E, L, and K, we construct simulated data matrices $\mathbf{F} = \mathbf{WM} + \eta$ by randomly choosing $\mathbf{M}$ and $\mathbf{W}$ as described below. The noise η in each element is drawn independently with scale 0.3 times the standard deviation of the entries in $\mathbf{WM}$. We construct simulated $\mathbf{F}$ matrices across a range of sparsities in $\mathbf{M}$ and $\mathbf{W}$. Specifically, for $\mathbf{M}$-sparsity p, entries are nonzero with probability p, and if nonzero, the entry is drawn from a standard normal. We then normalize $\mathbf{M}$ so that each row is a unit vector. We generate $\mathbf{W}$ analogously with $\mathbf{W}$-sparsity q, but without normalization. By not normalizing $\mathbf{W}$, we allow for the possibility that some phenotypes are influenced more strongly by the core processes than others.

We begin by constructing four sets of simulated data: one with both locus-sparsity and phenotype-sparsity ($p = 0.2$, $q = 0.2$), one each with only one type of sparsity ($p = 0.2$, $q = 1$ and $p = 1$, $q = 0.2$), and one with neither ($p = 1$, $q = 1$). For each set, we first applied the locus and phenotype rotation tests. The results are presented in the left column of Fig. 2. Note that the presence of the gap between the error curves for $\mathbf{F}$ and rotated $\mathbf{F}'$ in the locus (phenotype) rotation test depends on whether $\mathbf{M}$ ($\mathbf{W}$) is sparse. Repeating this test across a range of locus-sparsities and phenotype-sparsities, we show that the size of the gap grows continuously with sparsity (Supplementary Fig. 1).

Next, we evaluated whether SSD can accurately reconstruct the true $\mathbf{M}$ and $\mathbf{W}$ matrices. We applied our SSD method to each dataset across a range of locus-sparsity (λ_M) and phenotype-sparsity (λ_W) constraints and selected one decomposition using the SVD error and rotation tests as guides. The reconstruction error of the SVD decomposition on each dataset is in the range 0.047–0.049. Keeping in mind that any SSD solution will necessarily have higher error, we focus on "low-error" decompositions with error up to 0.85, illustrated by dark green, teal, and blue in the space of SSD decompositions (Fig. 2, center column). When either $\mathbf{M}$ or $\mathbf{W}$ is sparse, the space of solutions below a certain error is rectangular (Fig. 2). In this case, we select a decomposition that exhibits the most sparsity for the chosen error criterion, that is, the bottom left vertex of the rectangular contour (indicated by a white star in Fig. 2). When both $\mathbf{M}$ and $\mathbf{W}$ are not sparse, we pick a solution with similar degrees of locus-sparsity and phenotype-sparsity that satisfies our error criterion.

Finally, we compared the $\mathbf{M}$ and $\mathbf{W}$ of the selected SSD solutions to the true $\mathbf{M}$ and $\mathbf{W}$ matrices using a cosine error metric described in the Supplementary Material (third column of Fig. 2). We find that exhibiting sparsity in either $\mathbf{W}$ or $\mathbf{M}$ (first three rows) suffices for SSD to accurately reconstruct both $\mathbf{W}$ and $\mathbf{M}$. Given a non-redundant set of core processes $\mathbf{M}$, there is a unique set of phenotype weights $\mathbf{W}$ that best reconstruct $\mathbf{F}$ (and vice versa for $\mathbf{W}$). In contrast to SSD, the SVD decompositions are unable to accurately reconstruct $\mathbf{M}$ and $\mathbf{W}$, despite lower reconstruction errors when reconstructing $\mathbf{F}$.

The phenotypes constructed as described in this section are correlated in so far as each is a random linear combination of a common set of core processes. However, empirical studies may measure phenotypes with nontrivial structure, e.g. fitness measurements where the same environmental perturbations are added to various growth mediums. To validate the rotation tests and SSD in such a setting, we generated synthetic data with a hub-and-spoke structure. Specifically, we introduce "hub" phenotypes (representing the growth mediums) whose effects are a random linear combination of a common set of core processes and "spoke" phenotypes (each representing a growth medium with a perturbation) whose effects are a linear combination of the corresponding hub phenotype and one core process representing the perturbation (Supplementary Fig. 2a). See Supplementary Material for further details.

Next, we apply the rotation tests to the hub-and-spoke synthetic data and find evidence of both locus-sparsity and phenotype-sparsity (Supplementary Fig. 3). We find that a selected SSD solution exhibiting both types of sparsity accurately recovers the initially described generative structure. If we instead ignore evidence of locus-sparsity and select an SSD solution that exhibits a greater degree of phenotype-sparsity and little locus-sparsity, the decomposition resembles an alternate generative structure where each hub phenotype is instead described by a single core process (Supplementary Fig. 2b). In contrast, SVD finds a solution with lower reconstruction error but with matrices $\mathbf{M}$ and $\mathbf{W}$ that lack any clear relationship to the core processes that generated the synthetic data.

Applications to empirical data

Fitness effects of adaptive mutations in yeast

To illustrate the applicability of our framework, we first analyze data from a recent study by Kinsler *et al.* (2020). This study attempted to infer a lower-dimensional latent structure of phenotype space by measuring the fitness effects of a set of specific yeast mutations across a range of environmental perturbations. Specifically, they isolated 292 yeast strains from an earlier laboratory evolution experiment, each of which contains one or a few putatively adaptive mutations. They measured the fitness of each of these strains across a set of 45 environments (Fig. 3a). Based on these measurements, they divided the 45 environments into 25 "subtle" perturbations (in which fitness effects of mutations vary only slightly) and 20 "strong" perturbations. Applying SVD on the data from the subtle perturbations, they identified an 8-dimensional subspace that explains most of the variation in the data across these perturbations. They then showed that this latent structure can also predict the fitness effects of the mutations across the 20 "strong" perturbations, which they interpret as evidence that the subtle perturbations reveal a "local" modularity that is able to predict the global pleiotropic effects of adaptation in this system.

We sought to investigate whether our SSD method can recover an alternative sparse lower-dimensional structure in the Kinsler *et al.* data. Rather than divide environments into "subtle" and "strong" perturbations, we took the entire mutational effects

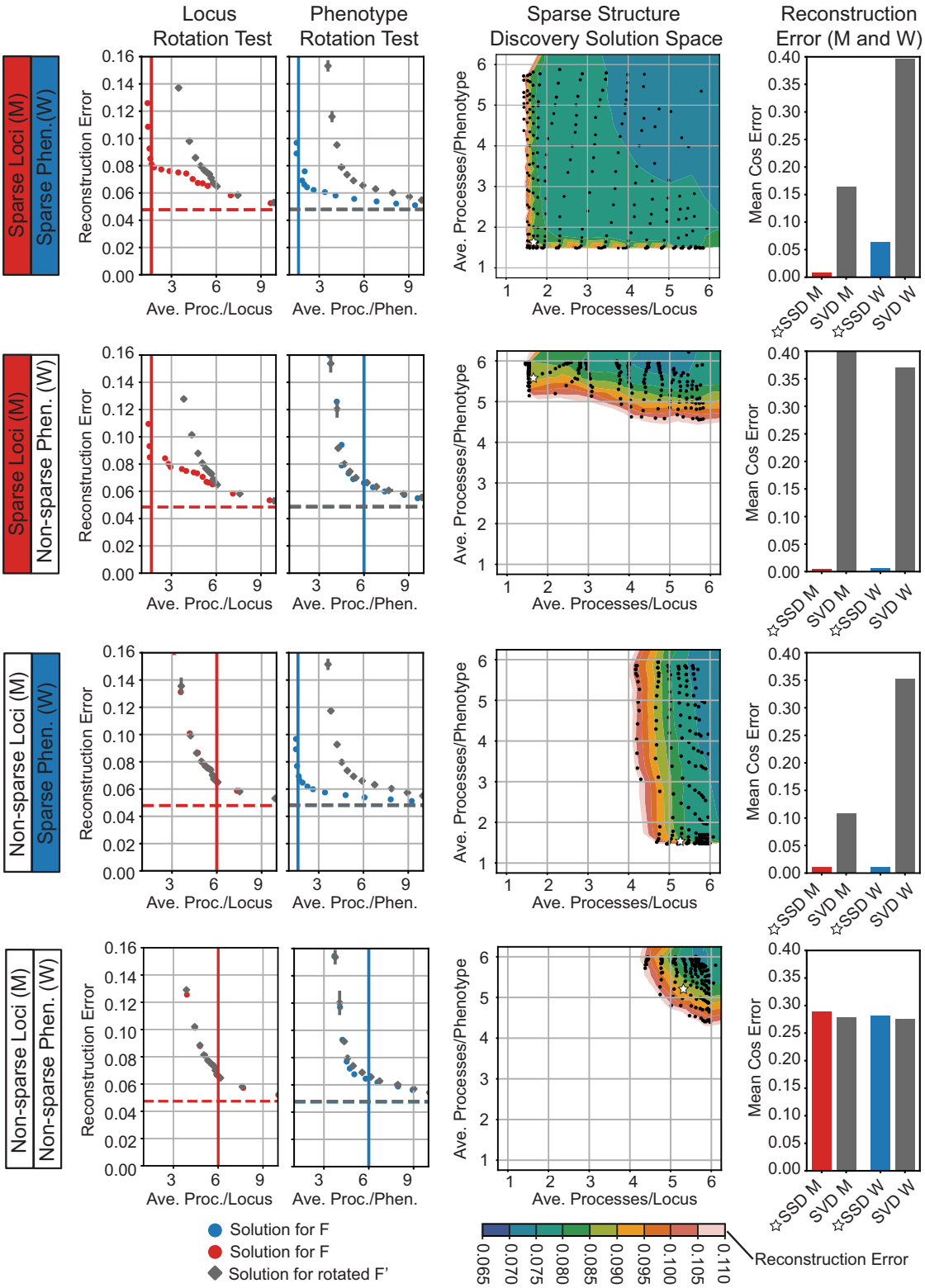

Fig. 2. SSD on synthetic data. Each row corresponds to a synthetic additive effects matrix $\mathbf{F} = \mathbf{WM} + \eta$ generated with different sparsities in $\mathbf{M}$ and $\mathbf{W}$. All examples have $E = 96$ phenotypes, $L = 200$ loci and $K = 6$ true core processes. First column: Locus (phenotype) rotation test illustrates that when $\mathbf{F}$ is generated with sparsity in $\mathbf{M}$ ($\mathbf{W}$), there is a gap between the sparsity of the SSD solutions for $\mathbf{F}$ in colored circles and rotated $\mathbf{F}' = \mathbf{FO}$ ($\mathbf{F}' = \mathbf{OF}$) in gray diamonds. Each scatter point corresponds to a solution for a particular value of the regularization parameter λ_M (λ_W). Error bars are over three random rotations; error bars were often so small that they were not visible over the scatter point. The red and blue horizontal dashed lines indicate the 6-component SVD reconstruction error. The gray dashed line indicates the average 6-component SVD reconstruction error of $\mathbf{OF}$. Note that the SVD error for $\mathbf{FO}$ is equal to that for $\mathbf{F}$. The vertical line indicates the average processes per locus (phenotype) for the true $\mathbf{M}$ ($\mathbf{W}$). Second column: Each scatter point depicts the average processes per locus/phenotype of an SSD solution. The colored background illustrates the interpolated reconstruction errors of the solutions. The solutions selected for further investigation are marked by a star. Third column: The mean cosine error between each row of the inferred $\mathbf{M}$ (column of inferred $\mathbf{W}$) for the selected SSD solution and for the 6-component SVD solution and the true $\mathbf{M}$ ($\mathbf{W}$). The error in $\mathbf{W}$ in the first row is almost exclusively due to 4 phenotypes that use no processes, but are assigned very small weights in some processes by SSD.

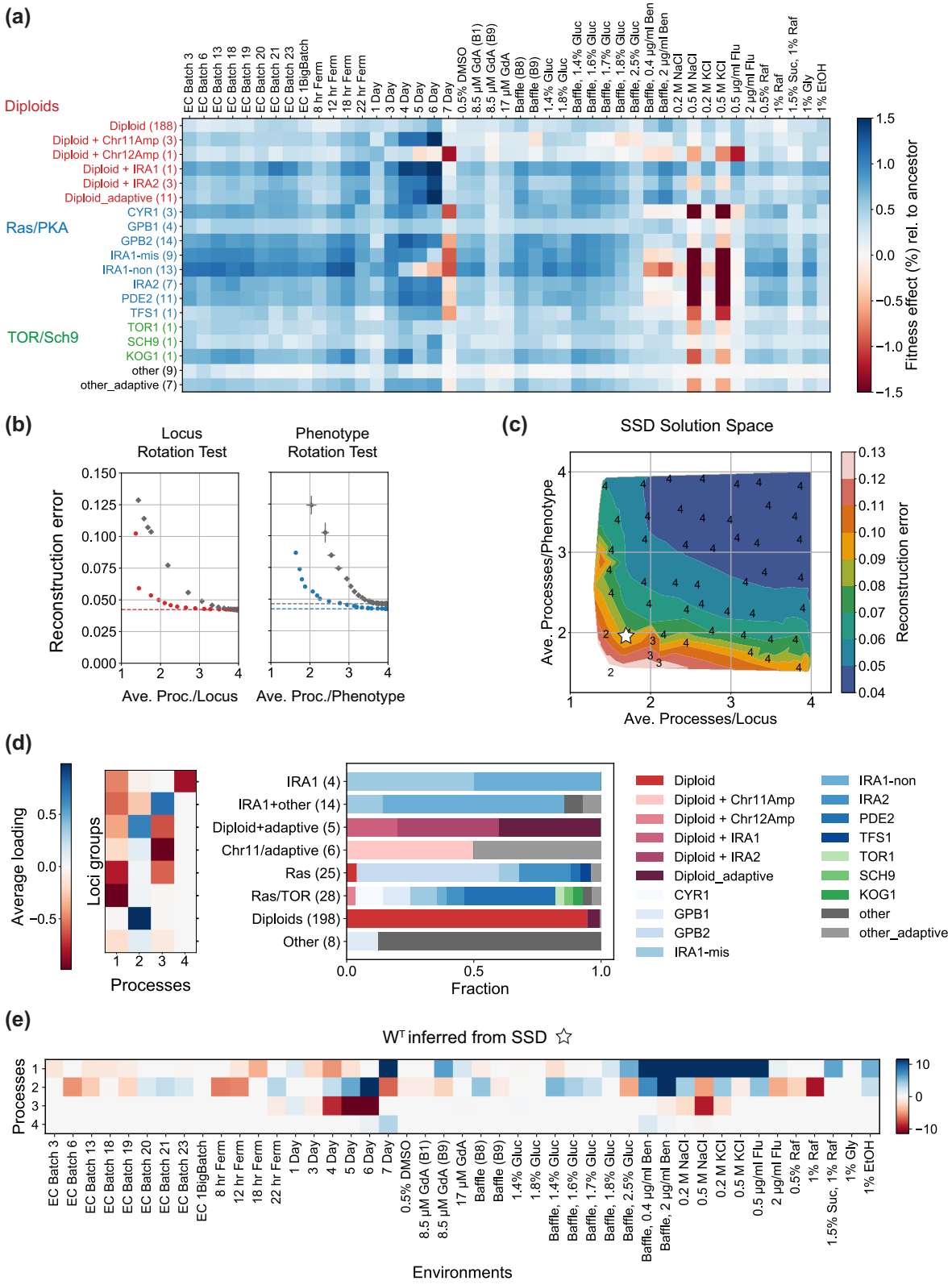

Fig. 3. SSD applied to pleiotropic fitness effects of adaptive mutations in yeast. a) A reduced representation of the effects matrix **F** (45(*E*) × 288(*L*)) where the effects of mutations with common annotations are grouped together. The number of mutations with each annotation is shown in parenthesis. b) The locus and phenotype rotation tests show extensive sparsity in both the process-phenotype and locus-process maps. c) The solution space illustrating highly sparse solutions with low reconstruction error. The integers indicate the number of processes in the solution. The chosen solution with 8.5% error is marked with a white star. d) The **M** matrix with loci clustered into 8 groups based on linkage clustering of loci with a modified cosine similarity metric (see Supplementary Material). On the right, the fraction of loci types in each of the 8 groups is shown. The number of loci in each group is shown in parenthesis next to its label. e) The process-phenotype map **W**.

matrix representing 288 strains across 45 environments as our input **F** (we use 288 instead of the original 292 due to a minor difference in a pre-processing step, see Supplementary Material). We then applied our locus and phenotype rotation tests (Fig. 3b), which confirm that there is strong evidence for sparsity in both the process-phenotype map (**W**) and the locus-process map (**M**). Note however that removing most diploids from this data (one key type of mutation that represents 188 of the 288 mutations studied) eliminates sparsity in **M** but not in **W** (Supplementary Fig. 4). Further analysis (discussed below) finds that the diploids predominantly affect one core process and thus the locus-sparsity indicated by the rotation test can be explained by the large number of diploids in the data. This is not an issue for applying SSD, as SSD requires sparsity in only one of **W** and **M**.

We find that SSD can identify a sparse, 4-dimensional approximation of **F** that incurs less than 8% error in reconstructing the original **F** (Fig. 3c). For concreteness, we focus here on the sparse solution indicated by the white star in Fig. 3c. We selected this solution because the locus-sparsity and phenotype-sparsity levels (an average of 1.5 processes per locus and 2 processes per environment) lie within the region of the rotation test sparsity-error curves where the error goes from dropping quickly to dropping slowing as a function of the sparsity (Fig. 3b). In Supplementary Fig. 5, we highlight the differences between the SVD and SSD solutions. By construction, the SSD solution has a higher reconstruction error than the corresponding SVD solution (7.5% error for the sparse SSD solution, compared to 4% error for the 4-dimensional SVD solution). We find that the SVD solution on a training set also shows lower error in predicting the fitness effects in held-out environments (the 20 strong perturbations or a random subset of 9 environments) compared to the SSD solutions of equal rank (Supplementary Fig. 5). This suggests that SVD tends to find a better low-rank approximation, even when it fails to find meaningful (and potentially sparse) basis vectors (see "*Discussion*"). To highlight this point, if SVD finds the locus-process and process-phenotype maps $\mathbf{M}_{SVD}$, $\mathbf{W}_{SVD}$ on the training set, it can be mathematically shown that the maps $\mathbf{M}' = \mathbf{O}\mathbf{M}_{SVD}$, $\mathbf{W}' = \mathbf{W}_{SVD}\mathbf{O}^T$ for any arbitrary orthogonal matrix **O** will match SVD's generalization error. In contrast, the SSD solution is significantly sparser than the SVD solution (Supplementary Fig. 5) at the expense of a larger generalization error. Thus, while SVD by construction finds the subspace with the lowest reconstruction error, the SSD approach is able to identify sparse basis vectors, capturing the sparsity in the genotype–phenotype map suggested by the rotation tests.

To examine if loci with similar effects on core processes identified by SSD align with existing annotations, we further clustered loci into 8 groups by comparing the columns of the **M** matrix with a modified cosine metric (Supplementary Material). We observe that core process 1 is enriched for mutations in genes involved in the Ras and TOR pathways (Fig. 3d). Missense and nonsense mutations in IRA1 (also involved in the Ras pathway) clustered in the "IRA1+other" group have additional pleiotropic effects on core process 3, which has a large influence on fitness in environments with an extended stationary phase (4, 5, and 6 day environments in Fig. 3e). Diploids are primarily enriched in core process 2, which has broad pleiotropic effects across environments. Diploids with additional mutations in IRA1/2 (clustered in the "Diploid + adaptive" group) exhibit effects that combine the effects shown independently by IRA1/2 in the Ras cluster and the Diploids cluster. Thus, the core processes identified by SSD do appear to have some correspondence with our prior expectations. To ensure that the many diploids do not significantly bias our results, we repeated this analysis on a reduced dataset which excludes a random subset of 168 of the 188 diploids, finding similar features in the **W** and **M** maps despite lower average sparsity in **M** (Supplementary Fig. 4).

Finally, it is easier to read off hypotheses from a sparse SSD decomposition than from a dense SVD decomposition (Supplementary Fig. 5b). For example, since SSD core process 3 almost exclusively impacts environments with an extended stationary phase (4, 5, and 6 days), it is reasonable to hypothesize that loci involved in this core process influence a pathway relevant in stationary phase. In contrast, each SVD core process affects most environments (Supplementary Fig. 5c), thereby confounding an analogous interpretation. The SSD solution further suggests that diploidy primarily contributes to core process 2, and the contribution of this process across environments is a succinct summary of its effect. For the SVD solution, the diploids do not form a single cluster (Supplementary Fig. 5c and d), and no such summary is apparent.

Robustness of gene knockouts to genotoxins in human cell lines

Next, we apply our SSD method to the genotoxic fitness screen collected in Olivieri *et al.* (2020) and curated in Pan *et al.* (2022) (Fig. 4a). This dataset was constructed by performing CRISPR-Cas9 knockouts on an immortalized human cell line (RPE1-hTERT) and subjecting each knockout variant to 31 genotoxic stressors. We show that the core processes described by our SSD decomposition are enriched for particular gene annotations and compare our decomposition to one identified by Webster (Pan *et al.* 2022).

Our rotation tests find evidence of both locus-sparsity and phenotype-sparsity in this genotoxin data (Fig. 4b). Phenotype-sparsity is not assumed by Webster (Pan *et al.* 2022), suggesting that SSD may lead to a more interpretable process-phenotype map. In order to compare directly to the Webster decomposition analyzed in Pan *et al.* (2022), we restrict our attention to SSD solutions that have the same number of core processes ($K = 10$). Note that alternate SSD solutions with fewer core processes incur more error. We observe that at approximately 2.5 processes per locus the rotation test sparsity-error curve (Fig. 4b, right) transitions from sharply decreasing to slowly decreasing, suggesting this level of sparsity in the process-phenotype map. Guided by these constraints and observations, we select a solution that is sparse in both loci and phenotypes (3.3 average processes per locus, 2.5 average processes per genotoxin), indicated by the white star in Fig. 4c.

First, we evaluate whether the core processes described by our solution are enriched for loci with particular functional effects. We organize the locus-process map **M** by the loci annotations compiled in Olivieri *et al.* (2020) and observe that core processes 1, 2, and 7 are enriched for loci involved with the repair of interstrand cross-links (ICLs) by Fanconi Anemia (FA) proteins, nucleotide excision repair (NER), and DNA replication fork quality control (FORK QC), respectively (Fig. 4d). Loci involved with end joining are primarily split between core processes 6 and 10. Finally, core process 5 is enriched for loci involved with base excision repair and single-strand break repair as well as homologous recombination. The functional meaning of the other four processs are not immediately clear from the annotations so we leave them unlabeled; investigating the loci with the strongest effects could elucidate their meaning, as was done by Pan *et al.* (2022). Figure 4e illustrates the process-genotoxin map **W**; the sparsity indicates that a small number of core processes explain the effect of each genotoxic stressor.

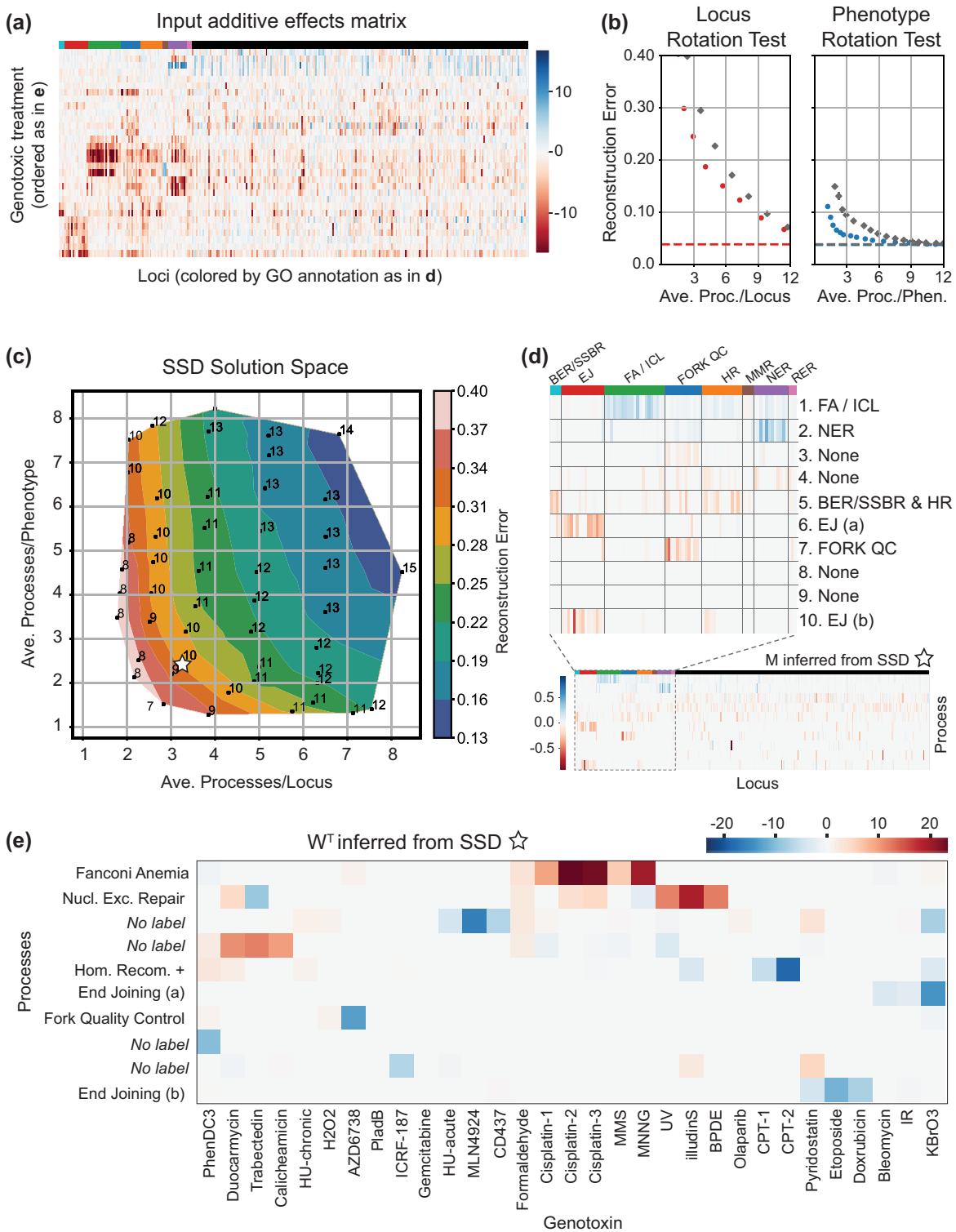

Fig. 4. SSD applied to dataset of human cell responses to gene knockouts under genotoxic stressors. a) The input additive effects matrix **F** generated by Olivieri *et al.* (2020) and curated by Pan *et al.* (2022). b) The locus and phenotype rotation test indicate there is both locus-sparsity and phenotype-sparsity. c) The space of solutions found by SSD. The integers indicate the number of processes in the solution. The white star indicates the solution that we illustrate in d) and e). d) Sorting the loci by GO annotation in the locus-process map **M** reveals that certain processes are enriched for particular annotated functions. e) The process-phenotype map **W** demonstrates that the response to each genotoxin can be explained by a small number of core processes.

In the Supplementary Material, we further describe the differences between Webster and SSD and compare the decompositions of this dataset found by each method. Our SSD method more accurately reconstructs the additive effects matrix while exhibiting more phenotype-sparsity and only slightly less locus-sparsity. Moreover, our SSD decomposition exhibits variation in the degree of pleiotropy across loci, measured by the number of processes each locus participates in (Supplementary Fig. 6).

The genotype–phenotype map of a yeast cross

Next, we analyze data from a recent study (Ba *et al.* 2022) analyzing genotypes and phenotypes of $N \approx 100,000$ F1 haploid yeast offspring (segregants) of a cross between RM (a European wine strain) and BY (a standard lab strain). These two parental strains differ by $S \approx 42,000$ single-nucleotide-polymorphisms (SNPs), leading to a highly diverse set of genotypes in the segregant pool. This earlier work measured the fitness (growth rate relative to the parental BY strain) of each of the segregants in $E = 18$ environments using a bulk barcode-based phenotyping assay.

The base condition for most of these environments is propagation in batch culture with 1:128 dilutions every 24 hours in rich laboratory media [yeast extract-peptone-dextrose (YPD)] at optimal temperature (30°C). We refer to this as the 30°C environment. Other environments are then constructed by adding stressors to this base condition (e.g. lithium, 4-nitroquinoline oxide, ethanol), by varying the temperature (23–37°C), by using defined media with various carbon sources (glucose, mannose, raffinose) instead of YPD, and by using complex natural media (molasses).

To apply SSD to this data, we must first infer the genotype–phenotype map for each of these 18 environments (i.e. we must infer **F**). This is a complex problem; Ba *et al.* (2022) includes an extensive discussion of the challenges associated with this inference and introduces a modified stepwise forward search procedure for this purpose. A particular difficulty is that this mapping is typically not able to precisely pinpoint specific loci that affect each phenotype. Because our goal is to use the genotype–phenotype map across these different environments to infer lower-dimensional latent structure, we adopt a simpler approach here. Instead of identifying putative causal loci separately for each phenotype, we use a penalized regression approach to jointly identify a sparse set of loci that explain the fitness across environments (see Supplementary Material). Then, we use a statistical test to establish a confidence interval for the location of each putative causal locus. This procedure identifies 1,089 genomic regions containing putative loci and their fitness effects in the 18 environments. We use this $18 \times 1,089$ matrix as the effects matrix **F** for SSD, represented schematically in Fig. 5a.

We next apply the loci and phenotype rotation tests (Fig. 5b), finding evidence for extensive sparsity in the process-phenotype map **W** and moderate levels of sparsity in the locus-processes map **M**. The SSD solution space shows an error landscape that favors low-rank ($K \approx 6$–9) approximations to **F** which are sparse in **W** (Fig. 5c). We focus here on the $K = 8$ solution indicated by the white star in Fig. 5c, which represents a tradeoff between achieving high sparsity in **W** and moderate sparsity in **M** while retaining relatively low reconstruction error. The phenotype-sparsity of this solution is approximately 1.5 processes per locus, which corresponds to the sparsity value in the rotation test sparsity-error curve (Fig. 5b, right) where the function transitions from sharply decreasing to slowly decreasing. We verified that this solution explains a fraction of variance on a test set of genotypes comparable to that explained by the full **F** and the 8-component SVD solution (Supplementary Fig. 7a). Other reasonable choices of solutions lead to qualitatively similar results (Supplementary Fig. 7b).

In Fig. 5d, we show the resulting inferred **W**. We find that this matrix is sparse and has some intuitive features. First, we note that the term **b** in our SSD decomposition represents a constant effect of each locus on all of the measured phenotypes (i.e. the aspect of the genotype–phenotype map that is constant across all the environments). The inferred **W** then represents how the loci in a given process produce deviations from these constant effects across the different environments. We find that none of the inferred processes have substantial weight in **W** for our 30°C environment, indicating that **b** fully captures the genotype–phenotype map for this environment. This is intuitive, given that this environment is the basis for all other conditions. The environments which represent this same condition at slightly lower temperatures are also largely captured by **b**, though processes 4 and 8 do become slightly more important as we decrease the temperature. As we increase temperature, we find that process 7 becomes important, suggesting that this process is associated with high temperature response. Several processes are specific to given environments (e.g. process 1 primarily affects fitness in guanidinium chloride (gu), process 2 affects fitness in lithium (li), process 5 in suloctidil (suloc), and 6 in molasses (mol)). Some of these processes, such as processes 2 and 6, contain a largely nonoverlapping set of loci that affect their respective environments (li and mol) in addition to the constant effects captured by **b**. Finally, processes 3, 4, and 8 reflect processes that influence a few conditions, including some observed tradeoffs (e.g. between fitness in raffinose and ynb or mannose).

In Supplementary Table 1, we provide a list of the ORFs localized to each putative causal locus, GO annotations and descriptions from the Saccharomyces Genome Database (Cherry *et al.* 2012), and their influence on each core process (i.e. value in **M**). In Fig. 5e, we show a Sankey figure that illustrates **W** and the most prominent features of **M**. This figure shows both how a number of key loci affect each of the processes (i.e. features of **M**), and how these processes in turn affect fitness in each of the environments (i.e. **W**). For example, we see that the genes ENA1 and ENA5 are the primary contributions to process 2, and that this process primarily influences fitness in lithium. This is consistent with prior expectations, as the ENA cluster is involved in salt tolerance and is known to be important for lithium tolerance (Wieland *et al.* 1995). Similarly, we see that BUL2, known to affect heat-shock element mediated gene expression (see Supplementary Table 1), is the primary contributor to process 7, which influences fitness in the high temperature environments. In addition, some loci which are known to have large effects on fitness across these conditions (e.g. MKT1, IRA2) are also represented in **M**. There are also many other loci (some of unknown function and other unannotated genes) that play a role, and the rationale for these patterns is unclear. Additional experiments measuring fitness across a larger set of environments may help further disentangle structure in this genotype–phenotype map, and help resolve additional processes.

Discussion

Extensive work in quantitative genetics has aimed to develop models that explain the relationship between genotype and a variety of different phenotypes. This work often finds widespread pleiotropy, where specific genetic variants affect multiple phenotypes, creating a complex pattern of correlations between phenotypes. Using these patterns to infer a lower-dimensional structure in the map between genotype and multiple phenotypes is an important goal, which offers the promise of identifying a biologically meaningful explanation for observed patterns of pleiotropy.

A central challenge in achieving this goal is that discovering lower-dimensional structure in high-dimensional data is fundamentally underdetermined. Thus, we must always choose some set of objective functions and/or constraints as the basis for any such decomposition. This choice is inherently somewhat arbitrary, and it is not immediately clear how to select objectives

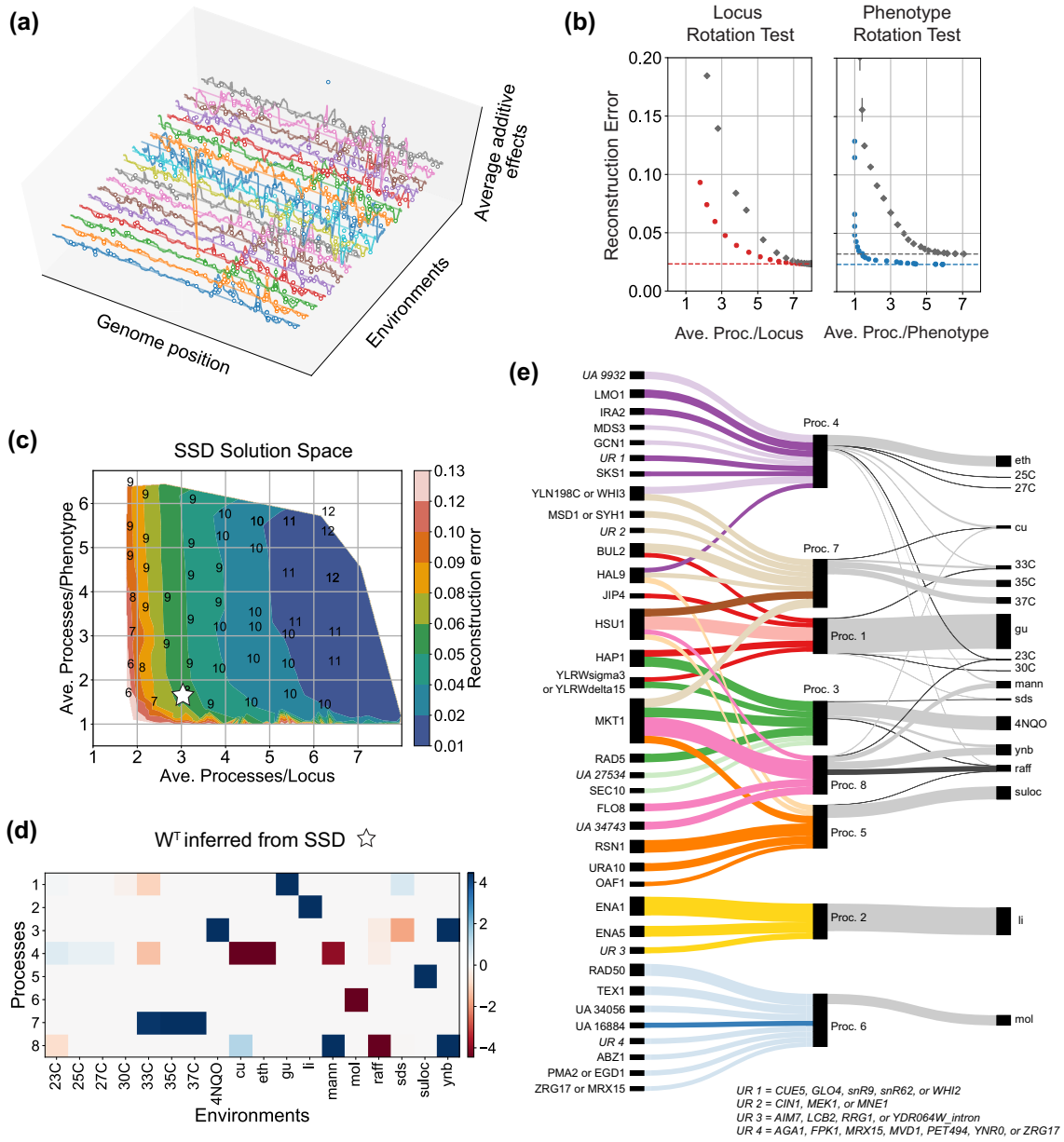

Fig. 5. SSD on genotype–phenotype data from a yeast cross. a) The average additive effects of $S \approx 42,000$ genetic loci, estimated using unpenalized linear regression for each of the 18 environments independently. The environments are arranged from bottom to top as arranged in panel d from left to right. Note that the correlations in average additive effects across neighboring loci due to linkage. b) The loci and phenotype rotation tests, showing extensive sparsity in the process-phenotype map and moderate sparsity in the locus-process map. c) The solution space has a landscape reflecting the sparsity in the process-phenotype map. The integers indicate the number of processes in the solution. The solution picked for downstream analysis is starred. d) The process-phenotype map, $\mathbf{W}$. e) A Sankey figure illustrating the locus-process map $\mathbf{M}$ for the large effect loci in each core process and the process-phenotype map $\mathbf{W}$. The width of each line is proportional to the magnitude of the value in $\mathbf{M}$ or $\mathbf{W}$. In $\mathbf{M}$, the lighter (darker) shade of each color indicates that the RM (BY) allele contributes positively to the process. In $\mathbf{W}$, light and dark gray indicate positive and negative contributions to the phenotypic measurement, respectively. The signs of the core processes are adjusted so that they impact most phenotypic values positively.

and constraints that will lead to solutions that reflect biologically meaningful structure in the data.

In this paper, we address this challenge using a penalized matrix decomposition framework, *SSD*, which allows us to identify a low-dimensional set of "core processes" that concisely explains the observed patterns of pleiotropy in genotype–phenotype data. The method uses sparsity as a key constraint to decompose a model for how genotype influences multiple phenotypes into two linear sparse lower-dimensional maps: a map between the genetic loci and the set of putative core biological processes they affect, and a map explaining how these core processes determine the observed phenotypes. Using simulated data, we demonstrate that SSD can accurately recover the true locus-process and process-phenotype maps as long as at least one of them is sparse. We then apply the method to three empirical datasets, which include the fitness effects of adaptive mutations in different growth conditions, robustness of gene knockouts to a set of genotoxic agents, and the fitness effects of QTLs identified in a yeast cross.

SSD is a flexible method which offers a range of solutions that correspond to different strengths of the sparsity constraints on the

locus-process and process-phenotype maps (formally, one unique solution per choice of the hyperparameters that enforce sparsity). This choice could be made based on some prior biological expectations, or by using standard statistical approaches such as cross-validation to find the set of hyperparameters that minimizes generalization error. However, since our goal is to identify biologically meaningful low-dimensional structure rather than minimize generalization error, we explore the space of solutions found by SSD across a range of hyperparameters, and use the reconstruction error landscape and proposed rotation tests to guide the examination of specific solutions. We do not prescribe a method for selecting a single solution. Instead, by exploring solutions with different levels of sparsity, we can examine features of the solutions which are robust to the choice of specific hyperparameters.

Of course, the use of sparsity as the guiding constraint in our SSD method is a choice, and it would certainly be possible to identify alternative lower-dimensional decompositions of a given dataset by choosing a different set of objectives and constraints. Our choice of sparsity is guided by two main factors. First, because we can use rotation tests to provide evidence for sparsity, we can demonstrate whether or not this constraint is appropriate directly from empirical data (and in cases where there is no evidence for sparsity, SSD should not be used). Second, intuitive notions of modularity in biological systems suggest that sparsity in $\mathbf{M}$ and $\mathbf{W}$ may reflect characteristic features of biological organization. For example, sparsity in the locus-process map may reflect a situation where each gene participates in one or a few biological "modules" with specific defined functions, and each such module relies primarily on a relatively small fraction of all possible genes. Sparsity in the process-phenotype map may hold less generally, but could reflect scenarios where any observed phenotype typically depends primarily on a subset of all possible modules. We also note that our method only requires sparsity in one of these two maps, so it could be useful in scenarios where $\mathbf{W}$ is sparse and $\mathbf{M}$ is not, or vice versa.

Naturally, even in scenarios where a biological system has a modular structure and sparsity seems intuitively appropriate, all biological processes are inherently coupled at some level. For example, the "omnigenic" model recently introduced by Boyle *et al.* (2017) suggests that most loci affect almost every complex trait. The omnigenic model reflects the observation that large numbers of small-effect loci often dominate the heritability of complex traits. This is not inconsistent with the sparsity-inducing ℓ_1 constraint used in SSD. Formally, the ℓ_1 constraint reflects a prior assumption about the distribution (i.e. the spread) of effect sizes, namely, that a small subset of loci have much larger effect sizes than most other loci that affect each process. In contrast, an ℓ_2 constraint, for example, imposes a prior with a tighter spread of effect sizes. This constraint will instead lead to a dense (and non-unique) set of solutions. The sparsity assumption thus remains valid as long as the effects of mutations in the core genes of a pathway are significantly larger than the small effects of the genes outside the pathway, even if there are so many such small-effect genes that they dominate the heritability of the trait.

By using sparsity as a key constraint, our approach produces a different lower-dimensional latent structure in the data than SVD, a commonly used method which finds the subspace of a chosen dimensionality that achieves the lowest error in reconstructing the effects matrix (without any additional constraints). By construction, SVD produces a set of processes (formally, basis vectors that span this subspace) which are orthogonal and which are ordered monotonically based on the variation explained by each process. Previous work Kinsler *et al.* (2020) has shown that SVD applied to a subset of mutations and similar environments generalizes to a held-out set of mutations and dissimilar environments, which suggests that SVD can be fruitfully used to identify an appropriate low-dimensional subspace of processes. However, any set of independent basis vectors which span the subspace will lead to the same generalization error. That is, even though SVD achieves good generalization performance by finding the optimal lower-dimensional decomposition of the genotype–phenotype map, it does not necessarily lead to a unique set of biologically meaningful processes.

Our approach is similar in spirit to Webster, a method based on graph-based dictionary learning introduced recently by Pan *et al.* (2022). Like SSD, Webster relies on a penalized matrix decomposition framework to identify the locus-process and process-phenotype maps. However, Webster imposes a hard constraint that each locus affects at most two processes and imposes no sparsity constraint on the process-phenotype map. In contrast to Webster, SSD finds sparser solutions with an equivalent reconstruction error, and variable degrees of pleiotropy across loci.

We emphasize that the processes identified by SSD or any other method are fundamentally constrained by the genotypes we study and the phenotypes we choose to measure. We cannot hope to resolve any effects of loci that do not vary across the genotypes we analyze. Thus, it is important to consider the nature of the genetic variation in a given study in interpreting the results of an SSD decomposition: if a given type of variant is not represented, we may fail to identify core processes which depend on those variants. Moreover, it is important to note that expanding a dataset by including additional genotypes can in principle change the inferred structure.

Similarly, the constant effects of loci on all the measured phenotypes are represented by the $\mathbf{b}$ term in SSD. This reflects the effects of loci on phenotypes that cannot be resolved by the variation in the measured phenotypes. For example, if some core process influences a given type of stress response and we did not measure any phenotypes that depend on that particular type of stress, we would expect the effects of this core process to be absorbed into $\mathbf{b}$ along with all other processes whose effects do not vary across the measured phenotypes. By measuring additional phenotypes, we could hope to begin to resolve these processes, though our success in doing so would depend on the phenotypes chosen.

We note that by using a matrix decomposition framework, we have implicitly made several important assumptions about the structure of the genotype–phenotype map. First, we have ignored the effects of interactions between loci on the core processes. In other words, we assume that the effect of each locus on each core process does not depend on other loci. Second, the process-phenotype map is assumed to be a linear function of the core processes. Nonlinear structure in the locus-process and process-phenotype maps will lead to structured epistasis between loci in the genotype–phenotype data. This structure is in principle resolvable by measuring epistatic effects between loci for different phenotypes. However, we have focused here on the additive effects matrix, because this is both simpler and can be more reliably estimated given the scope of current datasets.

Finally, our study and others Kinsler *et al.* (2020), Pan *et al.* (2022) assume a strictly hierarchical genotype to process to phenotype map. That is, we assume that the genotype determines the core processes, which in turn determine the observed phenotypes. This structure has some intuitive appeal, and it is central to any latent structure discovery method of this type. However, it may not always hold in reality. For example, one can imagine a scenario where the effects of mutations on one core process depend on the state of another core process (in other words, core

processes affect mutational effects in addition to phenotypes). Our method (along with other matrix decomposition approaches such as SVD) is fundamentally unsuited to describe such scenarios, and developing methods to infer the structure of this and other more general types of genotype–phenotypes maps is an important goal for future work.

Data availability

Our code, including a tutorial, is available at https://github.com/spetti/sparse-structure-discovery. The three previously published data sets used in this work are accessible from refs. Kinsler *et al.* (2020), Olivieri *et al.* (2020), and Ba *et al.* (2022).

Supplementary material is available at GENETICS online.

Acknowledgments

We thank Andrew Murray and members of the Desai lab for useful discussions. We thank Eliot Fenton for giving us access to his scripts to help process the data from the yeast cross experiments.

Funding

SP and GR were partially supported by the NSF-Simons Center for Mathematical & Statistical Analysis of Biology at Harvard (award number #1764269) and the Harvard Quantitative Biology Initiative. MMD acknowledges support from the Simons Foundation (grant 376196), NSF grant PHY-1914916, and NIH grant R01GM104239.

Conflicts of interest

The author(s) declare no conflict of interest.

Literature cited

Altenberg L. Modularity in Evolution: Some Low-level Questions. Cambridge (MA): MIT Press; 2005. p. 99–128.

Ba ANN, Lawrence KR, Rego-Costa A, Gopalakrishnan S, Temko D, Michor F, Desai MM. Barcoded bulk QTL mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. Elife. 2022;11:e73983.

Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–1186. doi:10.1016/j.cell.2017.05.038

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, *et al.* Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res. 2012;40:D700–D705. doi:10.1093/nar/gkr1029

Clune J, Mouret JB, Lipson H. The evolutionary origins of modularity. Proc R Soc B: Biol Sci. 2013;280:20122863. doi:10.1098/rspb.2012.2863

Comon P. Independent component analysis, a new concept? Signal Processing. 1994;36:287–314. doi:10.1016/0165-1684(94)90029-9

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, *et al.* The genetic landscape of a cell. Science. 2010;327:425–431. doi:10.1126/science.1180823

Crombach A, Hogeweg P. Evolution of evolvability in gene regulatory networks. PLoS Comput Biol. 2008;4:e1000112. doi:10.1371/journal.pcbi.1000112

Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23:R89–R98. doi:10.1093/hmg/ddu328

Golub GH, Reinsch C. Singular Value Decomposition and Least Squares Solutions. *Handbook for Automatic Computation: Volume II: Linear Algebra*. Springer Berlin Heidelberg; 1971. p. 134–151.

Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, Carslake D, Hemani G, Paternoster L, Smith GD, *et al.* Apparent latent structure within the uk biobank sample has implications for epidemiological analysis. Nat Commun. 2019;10:1–9.

Hintze A, Adami C. Evolution of complex modular biological networks. PLoS Comput Biol. 2008;4:e23. doi:10.1371/journal.pcbi.0040023

Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000;13:411–430.

Jutten C, Herault J. Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Signal Process. 1991;24:1–10. doi:10.1016/0165-1684(91)90079-X

Kinsler G, Geiler-Samerotte K, Petrov DA. Fitness variation across subtle environmental perturbations reveals local modularity and global pleiotropy of adaptation. Elife. 2020;9:e61271. doi:10.7554/eLife.61271

Lee D, Seung HS. Algorithms for non-negative matrix factorization. Adv Neural Inf Process Syst. 2000;13.

Olivieri M, Cho T, Álvarez-Quilón A, Li K, Schellenberg MJ, Zimmermann M, Hustedt N, Rossi SE, Adam S, Melo H, *et al.* A genetic map of the response to dna damage in human cells. Cell. 2020;182:481–496. doi:10.1016/j.cell.2020.05.040

Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res. 1997;37:3311–3325. doi:10.1016/S0042-6989(97)00169-7

Paaby AB, Rockman MV. The many faces of pleiotropy. Trends Genet. 2013;29:66–73. doi:10.1016/j.tig.2012.10.010

Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics. 1994;5:111–126. doi:10.1002/env.3170050203

Pan J, Kwon JJ, Talamas JA, Borah AA, Vazquez F, Boehm JS, Tsherniak A, Zitnik M, McFarland JM, Hahn WC. Sparse dictionary learning recovers pleiotropy from human cell fitness screens. Cell Syst. 2022;13:286–303. doi:10.1016/j.cels.2021.12.005

Rockman MV. Reverse engineering the genotype–phenotype map with natural genetic variation. Nature. 2008;456:738–744. doi:10.1038/nature07633

Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013;14:483–495. doi:10.1038/nrg3461

Wagner GP, Pavlicev M, Cheverud JM. The road to modularity. Nat Rev Genet. 2007;8:921–931. doi:10.1038/nrg2267

Wagner GP, Zhang J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat Rev Genet. 2011;12:204–213. doi:10.1038/nrg2949

Wang YX, Zhang YJ. Nonnegative matrix factorization: a comprehensive review. IEEE Trans Knowl Data Eng. 2012;25:1336–1353. doi:10.1109/TKDE.2012.51

Wieland J, Nitsche AM, Strayle J, Steiner H, Rudolph HK. The PMR2 gene cluster encodes functionally distinct isoforms of a putative Na+ pump in the yeast plasma membrane. EMBO J. 1995;14:3870–3882. doi:10.1002/j.1460-2075.1995.tb00059.x

Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and

canonical correlation analysis. Biostatistics. 2009;10:515–534. doi:10.1093/biostatistics/kxp008

Yankelevsky Y, Elad M. Dual graph regularized dictionary learning. IEEE Trans Signal Inf Process Netw. 2016;2:611–624.

Zhang Z, Xu Y, Yang J, Li X, Zhang D. A survey of sparse representation: algorithms and applications. IEEE Access. 2015;3:490–530. doi:10.1109/ACCESS.2015.2430359

Editor: H. Huang