**Title Ideas**

1. Using subtle environmental shifts to probe the number of phenotypes that contribute to fitness

2. Adding subtle environmental shifts to Fisher’s Geometric Model reveals…

3. Subtle environmental perturbations reveal the number of fitness-relevant phenotypes

4. Subtle environmental perturbations reveal the number of fitness-relevant phenotypes in adaptation to glucose limitation

**Authors**

Grant Kinsler\*, Kerry Geiler-Samerotte\*, Dmitri Petrov

**Abstract**

Tackle this later

Hi Grant! Since you are interested in improving your science writing, I tried to give you advice on how to write this paper in line throughout the paper. But my opinions are just that – opinions! So feel free to disagree. The manuscript will evolve with the data, so some of what I advise may become less relevant as we go. I think this is a good start for now. Happy editing!

Are “fitness components” necessarily phenotypes?

**Introduction**: I think the introduction can be thought of as three sections. This is not to say there should be actual sections, or subheadings. It is more a tool to help us write and discuss how to write. There will be transition sentences from one section to the next. The first section is about why our research is important. Here are my ideas for this section, which are not refined (wording sometimes awkward) so please feel free to edit. The second paragraph is critical – it gets to the main question “how many phenotypes are there” and may hint at adaptation, should we decide to include batch effects and focus on adaptation. We will probably want to fine tune the wording of this paragraph once the results section is done so that it corresponds perfectly to what we actually study.

The terms “genotype” and “phenotype” were coined in 1911 by W. Johannsen (Johannsen 1911) to create a distinction between an organism’s characteristics that were inherited and those morphological features that could be observed. Since then, our ability to observe phenotypes has grown by leaps and bounds. We can observe thousands of phenotypes in a single measurement (*e.g.* genome-wide transcriptome or proteome levels [CITE], methylation patterns [CITE], morphometric stuff with mice slicers [CITE Tautz, Frank from Tübingen] and powerful microscopes [CITE Ohya], the location of every gene in nucleus [CITE someone at NW], etc). A major challenge is understanding how all of these phenotypes interact to produce higher-level phenotypes of interest (*e.g.* disease). More specifically, quantitative genetics strives to trace the impact of a genetic change through many levels of molecular-level changes (expression levels, positioning, skeletal) that feedback upon each other in order to ultimately understand how a change at the genotypic level affects organismal fitness.

Complexity at the phenotypic level not only creates challenges for understanding the genetic underpinnings of complex traits, it also highlights questions about the organization of biological systems. Some questions include whether these many interacting phenotypes are organized into modules, such that even though it often appears that genetic changes have effects on many molecular level features (*i.e.* even though pleiotropy appears pervasive), the affected features typically belong to the same module, have levels that influence one another, and can be considered a single unit that orchestrates a single coordinated effect on organismal fitness. In other words, high level questions remain about how many independent phenotypes contribute to organismal fitness. These questions are of practical significance, for example, if many mutations can provide a microbe with drug resistance, each primarily affecting a different gene, but all contributing to drug resistance through the same higher-level phenotypic change, it is a simpler problem to solve than the acquisition of drug resistance through many disparate mechanisms.

Understanding the relationships between phenotypes would also make the study of biology more feasible. Statistical power to identify the genetic variants that contribute to phenotypic variation is diluted when many phenotypes are surveyed. Further, studies that count the number of phenotypes that are affected by a given gene or associated with fitness in a particular environment are upwardly biased if the phenotypes they measure are not independent.

The next section of the introduction is about all the people who have contended with this issue of ‘too many phenotypes’ before us. The point of the section is (1) to convince readers of the importance of our work by revealing the large # of people who tackled this issue, and (2) to synthesize many previous studies into a new understanding of various approaches to this issue.

The point of this section not to describe previous studies. For example, you describe drift load, which is super cool, but our paper is not about drift load. We don’t care if our readers know what drift load is. This section needs to be driven by your opinion of what is important (and perhaps missing) from previous studies, not by the details of the studies themselves. It is very rare to start a sentence and say, “Kinsler et al showed…”. You might want to start a paragraph with an opinion: “The fitness effect of mutations are inherently tied to the genetic backgrounds and environments in which they are measured”. Then expand on this opinion. What is your evidence that this is the case? Why is this the case? Then talk about previous studies in light of this opinion, “previous studies that find conflicting number of phenotypes that contribute to fitness (5 vs. 100) may do so because they focus on different environments”.

It can be hard to synthesize previous literature into a unique perspective. Try thinking about the things previous studies have in common or do differently. Which approaches calculate the correlations across traits and which fit data to a model? Which approaches focus on deleterious mutations or single mutations? Which make assumptions about the distributions of mutational effects? Most importantly, focus on the aspects that are most relevant to your own results that you will present later.

Try not to focus only on descriptive statements like, “Most previous studies examine the phenotypes that contribute to fitness by making large genetic perturbations (cite, cite cite).” Try to, once in a while, think deeper and have statements like, “previous studies that examine fitness-relevant phenotypes across diverse genetic backgrounds (cite cite) might tell us about X, but studies that examine phenotypes across more subtle perturbations will tell us about Y.

Overall, this section should be (1) shorter and (2) more inclusive. Add pareto-optimality and antigenic cartography? Add Brauer & Botstein Mol. Evol. Cell? Here are the paragraphs that you had in this section, plus some extra text and comments that might help. Don’t edit these paragraphs. Rethink based on the comments above and rewrite.

To contend with the complexity of phenotype space, a number of dimensional reduction approaches have been used in previous studies. A very common approach is PCA, or variants of PCA like X, Y, Z. These approaches calculate the correlation structure across traits that are measured to find a smaller number of orthogonal phenotypic components. They assume a linear correlation across traits and are useful when trying to understand correlations between traits in an unstratified data set.

A common framework used to model and understand phenotypic evolution that allows for non-linear relationships between traits is Fisher’s Geometric model (Fisher 1930) (see Fig 1A). In this model, traits are represented by orthogonal axes in a *D*-dimensional space, with the number of dimensions, *D,* indicating the number of traits that have independent contribution to fitness. Organisms are represented by points in this space, with their position determined by the organism’s particular combination of trait values. The fitness of an organism in a particular environment is determined by a function of the organism’s position and the position of the optimum combination of traits in that particular environment (hereafter “the optimum”). Thus, evolution can be seen as proceeding through “adaptive walks” from an ancestral organism that occupies some location in phenotype space. Genetic mutations can cause phenotypic changes, moving the mutant to a new location in this space. If this new location is closer to the optimum, then it may increase in frequency and give rise to additional mutations.

Using this general framework, estimating the number of independent traits represented by mutants is equivalent to estimating the number of dimensions in Fisher’s model. Previous work has aimed to measure the dimensionality of this space for mutations in general (and more specifically deleterious mutations), rather than the dimensionality of adaptation. These studies primarily fall into three categories. One class of methods derives the dimensionality of the space from the distribution of fitness effects (Martin & Lenormand 2006) (Fig 1B). Assuming the ancestor is at the optimum, a Gaussian mutation distribution in phenotype space, and a multivariate Gaussian distribution for the fitness function, Martin and Lenormand (Martin & Lenormand 2006) derive the expected distribution of fitness effects (DFE) taking the form of a gamma distribution. One can then derive the effective dimensionality of the system utilizing the moments of the DFE. Using directed mutation and mutation-accumulation techniques, these methods find that phenotypic dimensionality is relatively low (less than three dimensions) for the model organisms considered.

A second class of methods uses drift load (Poon & Otto 2000; Tenaillon et al. 2007). These also assume that populations are relatively close to the optimum, but instead explicitly use distance from an optimum as a metric to quantify the phenotypic dimensionality. Intuitively, this is accomplished by observing that large populations are able to effectively purge most deleterious mutations. However, small populations are unable to remove mildly deleterious mutations that drift to fixation, thus carrying a “drift load”. Moreover, the number of such deleterious mutations available to a population can be worked out as a function of the number of phenotypic dimensions, assuming some mutation distribution. Combining these relationships, this class of methods can then observe the fitness of populations across a gradient of effective population sizes to estimate the effective number of phenotypic dimensions. From this set of measurements, researchers found that there were relatively many dimensions (more than five) for the organisms observed.

Attempting to reconcile the different measures of phenotypic dimensionally observed by the two classes of methods, Lourenco(Lourenço et al. 2011) conducted analysis that suggests that DFE approaches are downwardly biased by universal pleiotropy assumptions, and that methods using drift load are relatively insensitive to the pleiotropy of mutations. Lourenco proposes a model of “partial pleiotropy” to explain the observed differences, suggesting that single mutations only affect a small subset of traits, but that mutations altogether affect many traits.

A third class of methods utilizes datasets generated to study epistasis (Fig 1C). Assuming a Gaussian mutation distribution around the ancestor and additive effects of mutations in phenotype space, epistasis is expected to be reflected in the nonlinearity of the phenotype to fitness map. Previous studies have used the distribution of angles between pairs of mutations (Weinreich & Knies 2013) and approximate Bayesian computation(Blanquart & Bataillon 2016) on epistasis datasets to understand phenotypic dimensionality.

OK, last we need a section explaining how our method differs from those you just described. The previous section should set this up nicely. Now all you need to say is “Here, we present a model that solves all the issues we just told you about”. But the “issues we just told you about” should be a full paragraph, maybe two, re-visiting the issues and how you solve them. End with a little summary of the results (I know that seems odd, but most introduction sections actually include the conclusion – ever heard this, “tell them what you are going to tell them, then tell them, then tell them what you told them?”).

strict mathematical assumptions to derive tractable mathematical expressions and predictions. Most (if not all) of these assumptions are likely to be violated. It is unlikely that the ancestor is optimally fit, since adaptation is constantly proceeding (Wiser et al. 2013). Moreover, assuming additivity of mutants in phenotype is likely to be wrong. This is particularly clear in the extreme example of pairs of loss of function mutations in the same pathway. Finally, a local Gaussian distribution for mutations in phenotypic space is also unlikely loss of function mutations can drastically change the expression of proteins and thus move a cell’s phenotype very far from the initial set of phenotypes. Additionally, there have been calls for the use of partial pleiotropy in these models to reconcile conflicting conclusions by some of these studies(Lourenço et al. 2011; Bataillon & Bailey 2014).

Something about infinitesimal model? + detectability?

One approach is to instead consider the problem from the other direction – measure changes in fitness across environments to understand which phenotypes are important to fitness in a given environment. There are clear problems with this approach when measuring fitness across very different environments: the environmental contribution to phenotype is likely to have a larger influence and different phenotypes are likely to have very different contributions in these various environments. Instead, consider subtle environmental changes. In such environments, rather than a complete change in the phenotypes important to fitness, there is instead a subtle shift in the relative importance of the phenotypes that influence fitness.

Recent technological advances allow us to measure the fitness of many mutants with high precision and in high-throughput (Levy et al. 2015). In particular, our approach develops the notion of *independent, fitness-relevant phenotypes*, that are affected by genetic variation and have significant and independent contributions to fitness, and uses precise fitness measurements across subtle environmental perturbations to identify these phenotypes (Fig. 1D). These phenotypes are not necessarily measurable traits in the traditional conceptualization of “phenotypes”, rather each fitness-relevant phenotype is likely influenced by complex combinations of measurable (and perhaps currently unmeasurable) features of that organism. We find that our method is able to accurately infer the number of phenotypes relevant to fitness on simulated data. Our ability to detect such phenotypes depends on both the magnitude of measurement error and the relative dispersal of mutants in phenotype space. [Furthermore, we show that we are able to detect particular combinations of mutants and conditions where behavior is inconsistent with the broader collection of mutations. Finally, we apply this method to fitness measurements of a yeast deletion collection, showing *something*.]

Do we focus on adaptation? All of your questions about this paper are coming back to me. If we focus on adaptation, we should include more on that in this last section of the introduction…but you can kind of go back and forth between this section and your results. Once they crystalize, so will the intro. Because of this evolutionary perspective on the importance of phenotype, identifying phenotype that have an effect on fitness in a given environment is an important area of study. Possibly back to evolution…how have things been shaped into modules such that mutations cannot affect some dimensions as easily as others. Lewontin (Lewontin 1974) further stratified these concepts in the light of natural selection and evolution: an organism’s “genotype” characterizes the information passed down to offspring, and an organism’s “phenotype” is the material that natural selection itself acts on. Our analysis re-defines phenotypes not as things we measure, but as things evolution shaped… Sorry, this is some nonsensical stuff. I was toying around with ending where we started, back on the definition of phenotypes.

**So what is this paper about? What is unique about our model? There are actually a lot of things, My list has things like, (1) we use both environment and genetic perturbations to understand phenotypic complexity (do others?), (2) we do not assume mutation distribution, (3) we study tiny perturbations (single mutations, batch effects), (4) we quantify how much power we have to detect phenotypes with smaller effects on fitness (anisotropy thing), (5) we (might) focus on adaptation. First, decide on the most important items in this list. Then, let them drive section 2 and 3 of the introduction.**

We use subtle environmental perturbations (and the associated changes in fitness) to uncover the number of fitness-relevant phenotypes represented by adaptive mutations. We first show that our approach works well on simulated data. Next, we show that for adaptation to a particular environment, despite the presence thousands of independent events that confer a strong selective advantage, only a small number of phenotypes explain the behavior of these mutants. Moreover, these phenotypic components accurately predict the fitness of these mutants in other environments.

**Results**

***A model that predicts the number of phenotypes contributing to fitness***

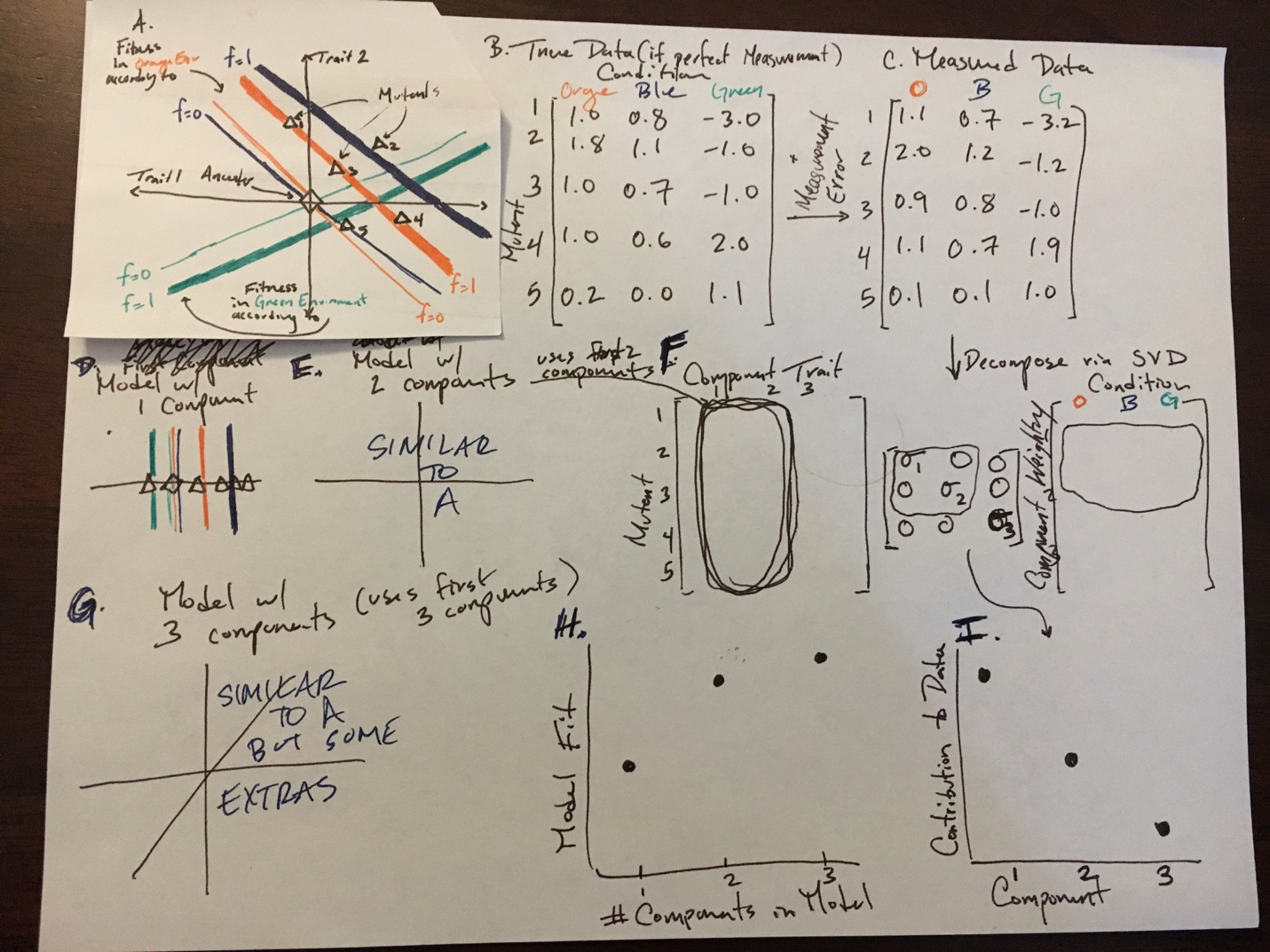
In order to break down the complexity of phenotypes, we imagine a group of single mutants for which we have measured their fitness in many subtly different environments. Each mutant has a phenotype defined by a particular combination of component traits, and a mutant’s fitness in a given environment is determined by the relationship between its component traits and the environment. This mapping from phenotype to fitness is environmentally dependent, with different combinations of traits and different weightings of these traits being more or less important in different environments. Thus, if we subtly change the environment that a mutant is in, then we can subtly change how well-suited each trait is to the environment, and, by extension, how well-adapted the mutant is to this new environment. Consequently, we can identify the number of fitness-relevant traits in the inverse manner, by using a mutant’s variation in fitness across a set of different environmental perturbations to identify how many phenotypes matter to fitness.

One key insight to this process is the use of subtle, rather than strong, environmental perturbations. Subtle perturbations allow us to model mutants as having fixed phenotypes, effectively removing any direct environmental impact on phenotype itself. Moreover, strong perturbations have the potential to reveal new phenotypes that were not relevant to the original environmental condition. Though potentially interesting, such phenotypes do not factor into fitness during the evolution of these strains in a particular condition, and thus, are not informative about the process of adaptation we wish to understand.

We use a particular mathematical representation of this framework, where mutants represented by fixed points in phenotype space based on the specific combination of traits that they represent. The ancestor is placed at the origin (so all other mutant phenotypes are defined relative to that of the ancestor). The fitness of a mutant in a given environment is a linear combination of these trait values, with the particular environmentally-dependent weightings (Figure 1A).

*short overview of inference techniques*

In order to estimate the phenotypic components of the mutants and the weightings of these components in each environment, we decompose the observed fitness data of mutants into two matrices: one representing the trait values of the mutants and another that represents the weightings of these traits for fitness in each environment using singular-value decomposition (SVD) (Figure 1C and F). SVD orders these component traits, such that the first component represents the best possible linear, single-component model for the observed data (Figure 1D and 1H). Subsequently, the first and second component combine to produce the best two-component model (Figure 1E and 1H). This continues until the model that combines all possible components (which is the number of conditions in this case), where the model perfectly captures the observed data (Figure 1G and 1H).



**Figure 1. Singular-value decomposition reveals the number of fitness-relevant phenotypes.** Panel A depicts a model of phenotype space where mutants (triangles) are at locations based on the particular combination of traits that they represent, relative to the trait values of the ancestor (diamond at origin). Fitness in a particular environment is a linear function of this phenotype space, with traits weighted according to the environment (thin and thick colored lines show the trait combinations that yield fitness values of 0 and 1, respectively). Fitness in different environments are characterized by different weightings of traits (as shown by different colors). With perfect information, we can calculate the fitness of each mutant in each condition (panel B), but measured data has some additional noise (panel C). Singular-value decomposition takes the measured data and divides it into 3 matrices: one representing the component traits for each mutant, one representing the weightings of these components for fitness in each environment, and a third that represents the contribution of each of these components to the observed data (panel F). The squared and normalized entries of this third matrix represents this contribution (panel I). Considering only the first SVD component (and its weightings in the environments) gives the best one-trait model that explains the data (panel D). Considering the first 2 (or 3) components give the best two- (three-) trait models as well (panels E and G, respectively). Increasing the number of components always increases the fit of this model to the measured data (panel H), despite there being only 2 true components in the model depicted in panel A. (Note: For visualization purposes, component contributions are collapsed into the component weightings).

***Our model makes accurate predictions about simulated data***

To test our method for estimating the number of phenotypes from fitness data, we perform a simulation study. First, we simulate data by placing an ancestor and mutants in a phenotype space. To generate fitness values for these mutants relative to the ancestor across many environments, we also generate weightings for different environments, and calculate the fitness of these mutants across the environments accordingly. To assess our ability to detect the number of fitness components in the presence of measurement error, we also add simulated measurement error on these fitness values.

We then use SVD to infer the number of fitness-relevant phenotypes on this simulated data – this is equivalent to identifying the appropriate number of components (and its corresponding model) from SVD. One first instinct would be to select the number of components that fully captures the values we observe in the data (the one with the most components possible). However, this approach does not take into consideration that our observations are imperfect due to measurement noise, and there will be some components that are only fitting this noise rather than real biological signal. This phenomenon is known as overfitting, and various methods have been proposed in order to select the model that balances explaining enough real signal and avoids fitting measurement noise.

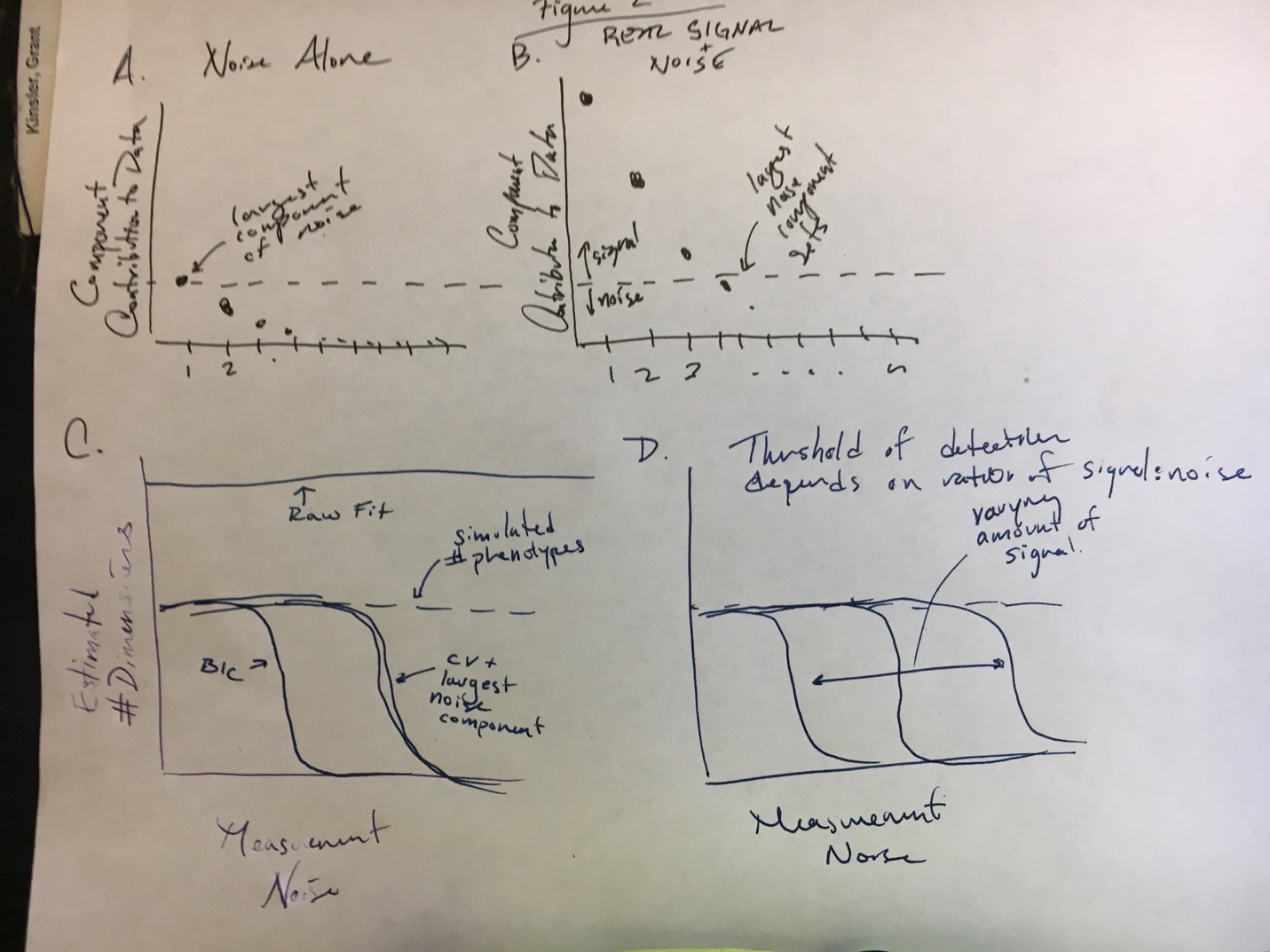
One common approach to solve the overfitting problem is to place an explicit penalty on additional parameters, such that only models the fit sufficiently better than the cost imposed by the additional complexity are preferred. This is referred to as using information criteria for model selection, and several such information criteria have been proposed for cases similar to ours using SVD [CITE paper from Owen and Perry]. These criteria do successfully identify the correct number of fitness-relevant phenotypes when measurement noise is very low, but they substantially underestimate the number of phenotypes at intermediate levels of measurement noise.

Another approach to control for overfitting that uses on cross-validation and model prediction [CITE Owen and Perry] accurately identifies model complexity in simulated data. Here, we divide the data into five separate data sets (for 5-fold cross-validation). One by one, we set aside each subset of the data, fit a model using the remaining four data sets, and evaluate the ability of this model to predict the data of the held-out test set. We then choose the model that shows the best predictive power over the aggregation of these hold-outs. This approach shows much better discriminatory power on our simulated data than either looking at fit alone or imposing an explicit penalty on additional parameters (Fig 2C).

Cross-validation provides a method for identifying the correct model complexity when measurement noise is sufficiently low, and underestimates the model complexity when measurement noise is too high (Fig 2C). This presents a problem for the use of this approach with empirical data, since when cross-validation selects a model, it is unknown whether this model is correct or if measurement noise merely swamps out the signal of the, possibly correct, more complex model. To distinguish between these two possibilities, we need a method to quantify how much of the contribution of each inferred phenotype is composed of real biological signal, rather than measurement noise.

To do this, we rely on the observation that even random measurement noise can have an underlying statistical structure, such that a random matrix, consisting only of data points that are drawn from a common distribution can be decomposed via SVD or other such approaches (Fig 2A). This inferred composition will have some strongest component. This strongest component of random measurement error represents the limit of detection for simulated data that contains a combination of signal and noise, such that any signal components weaker than the strongest noise component are undetectable (Fig 2B). With known measurement, error we can compute the expected size of this largest component by performing SVD on a series of random matrices generated with only measurement error (and no signal) [Sengupta and Mitra in here].

When cross-validation underestimates the number of phenotypic components on measurement data, there is little drop off between our weakest inferred signal component and the strongest noise component, whereas for sufficiently low noise, when we infer the number of phenotypes accurately, there is a noticeable drop-off between these components, indicating that, if any remaining signal remains, it is below the level of this noise component (Fig 2C). This serves as a practical method of evaluating empirical data and places a concrete bound on the size of the signal that is not detected by our inference procedure. This noise threshold varies based on the strength of real signal, such that stronger signals are easier detect, and thus, tolerate more measurement noise (Fig 2D).

**Figure 2. Cross-validation and the largest noise component are well-powered to identify the correct number of fitness-relevant phenotypes.** Using SVD on a matrix consisting only of noise reveals substructure (panel A). The largest component of noise represents the detectability threshold, such that when we decompose a matrix consisting of a combination of signal and noise (panel B), any components larger than the expected contribution of the largest noise component are likely to be real signal and any signal below this threshold is undetectable and indistinguishable from noise. Comparing this method with cross-validation and information criteria (panel C) reveals that information criteria are under-powered to detect the number of fitness-relevant phenotypes, whereas cross-validation and the largest noise component have more power to accurately detect the simulated number of phenotypes. However, this power depends on the ratio of signal to noise (panel D), as the methods are better able to detect the number of fitness-relevant phenotypes with stronger signals.

***Our model makes accurate predictions about real data***

*Now that it’s clear we have methods for identifying the number of phenotypes for simulated data, and we know that the ratio of signal to noise is important for this detection,*

*It’s low! (or not so much)*

[need data]

***We can detect fitness-components down to the limits of detection***

Measurement error limits our ability to detect particular fitness-relevant phenotypes (Fig 2B) – in the extreme case where measurement error is very high (\sigma\_m^2 = 0.1, corresponding to X% of the mean relative fitness value), our method is unable to detect more than one phenotype. This drives the question of how sensitive this method is to measurement error in general, and what aspects of data limit our ability to detect particular phenotypes.

The ability to detect the influence depends on the dispersion of mutants and optima in trait space, precision of measurement, and the combination of the two. If mutants and optima are less dispersed (mutants have similar phenotypes and conditions have subtle differences), then more precise measurements are needed (and vice versa). Because our model assumes that each trait contributes equally to fitness, traits that have lower contribution (in un-rescaled space) will also have the effect of lowering the dispersion of mutants for that trait in rescaled space. In particular, if mutants are uniformly distributed around the ancestor, but traits have varying levels of importance to fitness (Fig. 3A), then the re-scaling of the phenotype space in units of fitness will lower the dispersion of mutants. Similarly, if mutations change some traits more than others, this dispersal could also affect detectability (Fig. 3B).

We assess detectability by simulating cases where the true dimensionality is 3, but the third dimension has either reduced influence on fitness (Fig. 3D) or mutations affect this dimension less strongly (Fig. 3E). Both of these have a similar effect to the accuracy of inference: the less dispersed the third dimension is, the less detectable it is. This effect is exacerbated when measurement error is higher, indicating that the ability for our inference to detect a dimension depends on both measurement error and dispersal. [derive a measure for dispersal, show that this captures everything in Fig. 3F – 3C will be schematic for the measure if needed to explain clearly.]

***Non-subtle environmental perturbations reveal the context dependent mapping from phenotype to fitness***

Our method for inferring the space inherently relies on how the collection of mutants behave over a collection of conditions. How well can we detect the behavior of particular mutants and conditions? That is, is our method only good for inferring the commonalities but not particular behavior, and if so, how can we know that (and which) particular behaviors are not picked up?

*Result 4: Application to Hillenmeyer data?*

**Discussion**

[awesome discussion of this paper + upcoming amazing paper on 1bigbatch]

**Methods**

***Model Details***

We consider an explicit model of phenotypic evolution, where orthogonal traits

We use a global optimization technique to find the parameter values minimize this score for each number of dimensions. To estimate the number of dimensions and avoid overfitting, we use a 5-fold bi-cross validation scheme (Fig. 2) analogous to that used for Singular Value Decomposition(Owen & Perry 2009). We divide our data into 5 distinct sets or “folds” (250 mutants into 5 groups of 50 and 50 conditions into 5 groups of 10). For each fold of the data, we exclude a set of mutants and conditions to use as a test set and first use the computational method to estimate the best space for the remaining set of data (See step 1 of Fig 2A.) for each value of D (the number of dimensions) we are interested in. If we were to only do this, we would expect the fit to continually increase as we increase the number of dimensions, because we will begin to fit measurement noise. Next, we fix this space and use only the locations of the original set of mutants as information to find the best location for the conditions in the test set (step 2a of Fig 2A), and separately, use the locations of the original set of conditions to find the best location for the mutants in the test set (step 2b). Finally, we evaluate the relative fitness values predicted from the estimated locations of the mutants and conditions in the test set that were placed independently (step 3). The model that has the best fit of the test set’s relative fitness values to the measured values is the correct number of dimensions that was not over-fit to the original data. This is repeated for each of the 5 test sets, and we pick the model with the best average predictive ability across all 5 folds.

***Estimation Methods***

We use singular-value decomposition (SVD) to decompose the observed measurement matrix

Information Criteria used are those described in [cite reference] and designed for singular-value decomposition.

***Cross-validation details.*** We use bi-cross-validation as suggested for use with singular-value decomposition by Owen [cite Owen and Perry reference]. This involves first taking the data matrix and randomly dividing it into k subsets of mutants and conditions. Then, iteratively, each of these subsets is removed from the matrix and the remaining data matrix is decomposed using singular-value decomposition.

***Noise component identification details.***

***details.***

[transition paragraph?]

**References**

Bataillon, T. & Bailey, S.F., 2014. Effects of new mutations on fitness: Insights from models and data. *Annals of the New York Academy of Sciences*, 1320(1), pp.76–92.

Blanquart, F. & Bataillon, T., 2016. Epistasis and the structure of fitness landscapes: Are experimental fitness landscapes compatible with fisher’s geometric model? *Genetics*, 203(2), pp.847–862.

Fisher, R.A., 1930. *The Genetical Theory of Natural Selection*, Oxford: Oxford University Press.

Johannsen, W., 1911. The genotype conception of heredity. *The American Naturalist*, 45, pp.129–159.

Levy, S.F. et al., 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, advance on, pp.1–78. Available at: http://dx.doi.org/10.1038/nature14279.

Lewontin, R.C., 1974. *The genetic basis of evolutionary change*,

Lourenço, J.M., Galtier, N. & Glémin, S., 2011. Complexity, Pleiotropy, and the Fitness Effect of Mutations. *Evolution*, 65(6), pp.1559–1571.

Martin, G. & Lenormand, T., 2006. A general multivariate extension of Fisher’s geometrical model and the distribution of mutation fitness effects across species. *Evolution; international journal of organic evolution*, 60(5), pp.893–907. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.0014-3820.2006.tb01169.x/abstract.

Owen, A.B. & Perry, P.O., 2009. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Annals of Applied Statistics*, 3(2), pp.564–594.

Poon, A. & Otto, S.P., 2000. Compensating for our load of mutations: freezing the meltdown of small populations. *Evolution*, 54(5), pp.1467–1479.

Tenaillon, O. et al., 2007. Quantifying organismal complexity using a population genetic approach. *PLoS ONE*, 2(2).

Weinreich, D.M. & Knies, J.L., 2013. Fisher’s Geometric Model Of Adaptation Meets The Functional Synthesis: Data On Pairwise Epistasis For Fitness Yields Insights Into The Shape And Size Of Phenotype Space. *Evolution*, 67(10), pp.2957–2972.

Wiser, M.J., Ribeck, N. & Lenski, R.E., 2013. Long-Term Dynamics of Adaptation in Asexual Populations. *Science*, 342(6164), pp.1364–1367.