**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

5. Subtle environmental perturbations reveal the number of fitness-relevant phenotypes of adapting populations and the existence of context-dependent pleiotropic costs

Something more relevant to pleiotropy?

**Authors:**

Grant Kinsler, Dmitri Petrov, Kerry Geiler-Samerotte

**Affiliations:**

**Abstract:**

**Rough Outline of Introduction:**

# **Paragraph One: Adaptation is awesome, but it shouldn’t be.**

# Single punchy (less technical) sentence that captures main idea of paper:

# Organisms have an amazing ability to adapt to diverse challenges, yet recent observations suggest most mutations affect many traits and most of these effects are deleterious. How is adaptation possible if mutants that influence one trait in a beneficial way influence many other traits in a deleterious way?

* 1. We know it is indeed possible: examples of adaptation’s awesome power
     1. Ben Good showing that microbes never run out of adaptive mutations
     2. Which other examples?
  2. But how is adaptation possible? Conclude by defining the ‘cost of pleiotropy’ and explaining that the more pleiotropy there is, the greater the constraint on adaptation.

1. **Section Two: How much pleiotropy is there and why is it so hard to find out?**
   1. Long history of debate on this topic.
      1. Is pleiotropy omnigenic/universal or modular? If modular, by learning relationships among traits one might be able to predict which suites of traits are jointly influenced by genetic changes.
      2. Though modularity seems reasonable and many studies suggest cellular systems have a modular organization, GWAS studies detect ungodly amounts of pleiotropy suggesting every gene has the potential to influence every trait.
   2. This debate persists because of challenges defining and measuring phenotypes, and integrating information about phenotypes at different levels of biological organization
      1. There are too many phenotypes, potentially millions or an infinite number depending on how creative you are
      2. Phenotypes are often related, effects cascading through levels of biological organization, introduce genotype-phenotype-phenotype map idea
   3. Conclude by explaining the ‘impossible task’: Quantifying all of these traits and enumerating relationships between them to understand which suites of traits (or whether every trait) will be jointly influenced by a mutation is too much work.
2. **Section three: Clever solutions leveraging barcoding**
   1. We have a new approach to understanding the extent of pleiotropy among adaptive mutations that circumvents the problem. We do not measure any of these phenotypes, instead we repeatedly measure fitness.
   2. Briefly explain the idea in general terms, subtle perturbations, FGM
   3. Explain the large data set we will use to test the idea, and how this is only possible because (1) these strains are barcoded, (2) NGS gives us lots of power to measure fitness, and (3) we cleverly use information about fitness to learn something about phenotypes.
3. **Section four: What do we find? Punchlines**
   1. The number of ‘traits’ affected by an adaptive mutation depends on context
      1. Adaptive mutations influence a limited number of ‘traits’ in environments close to the one in which they are adaptive.
      2. As we perturb the environment farther from this one, we reveal additional pleiotropic effects of adaptive mutations
      3. This suggests resolution to the paradox, how pleiotropy can be pervasive and adaptation can still happen
   2. Our method is really cool.
      1. Talk about all the cool things you could learn using this epic approach.

We focus on identifying the fitness-relevant traits of a collection of adaptive mutations that arose in a barcoded evolution experiment [Levy et al]. Despite knowing the genetic basis of adaptation for many of these mutants, with the majority of adaptive mutations being either auto-diploidization or loss-of-function mutations in nutrient-response pathways [Venkataram et al], there remains the question of what phenotypic routes these mutations are taking. Do all the mutations in the nutrient-response pathways represent the same adaptive strategy in this environment? Do particular mutations exhibit specific adaptive strategies?

Li et al. (hidden complexity) began to answer these questions by probing the performance of these adaptive mutations in lag, fermentation, respiration growth stages, all of which were experienced in the evolutionary condition, as well as the extent to which these mutants tradeoff in stationary phase, which was not experienced in the evolutionary condition. The adaptive mutants generally gained an advantage during all three phases experienced in the evolutionary condition, though particular mutants differed in their extent as well as relative importance of the phases in their improvement. This poses a more general question: is performance in these growth phases the only way in which the mutants have gained a benefit? Are there other hidden phenotypes these mutants exhibit that promotes their fitness in this environment or other environments?

**Results**

***Results Section 1: A method to identify fitness-relevant traits***

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.

Figure 1: Conceptual figure, explaining method to identify fitness-relevant “traits”, also explain cross-validation/prediction scheme here? [do we want to use FGM figure or SVD figure? – FGM more intuitive, SVD more accurate to actual inference]

***Results Section 2: Our method makes accurate predictions on simulated data***

To demonstrate that our approach to identify fitness-relevant phenotypes works, we first perform a simulation study to understand if, how, and when our method is able to correctly identify fitness-relevant phenotypes and, perhaps more importantly, understand when our method breaks down. To do this, we first simulate data according to a particular number of phenotypes . When all phenotypes contribute equally

Identifying the number ideal complexity of a model inferred from data is a notoriously hard problem in statistics [citations] – numerous approaches have been devised to balance including enough parameters to capture the important behavior and avoid overfitting measurement noise. We use two independent approaches to identify the number of fitness components to use in our data. To validate our approaches, we simulate data with a known component space and, hence, a known number of components. We can then assess what factors influence our ability to accurately detect the component space.

The first approach takes advantage of known measurement error. For a set of data consisting solely of measurement error, using SVD on this data will reveal some hidden structure with one component explaining the most variation in this random data. This largest component represents the limit of detectability of real signal, as any signal smaller than this largest noise component will be swamped by measurement noise. Thus, to identify the smallest detectable component of noise, we can do SVD on the noise matrix, identify the largest noise component, and set this as our cutoff for detection.

The second approach uses cross validation to select the appropriate number of components to use in the model. Since there are both conditions and mutants to hold out, we use a bi-cross validation scheme [cite owen and perry], where a subset of the mutants and conditions are held out and models for each number of components are fit on the remaining data. Fixing this training model, the held-out conditions have their best location estimated for the original mutants, and vice-versa for the held-out mutants and the original conditions. Finally, we evaluate the fit of the inferred model for the held-out mutants in the held-out conditions to determine how well the model does. Once measurement error begins to become incorporated into the model, the predictive performance in the held-out groups should be reduced, as this measurement error gives errant information.

Both of these approaches are able to identify the correct number of components when measurement noise is sufficiently low. However, as measurement noise increases, the noise swamps out the smallest components of noise and become undetectable by these methods. The limit of detectability is set by the difference between the size of the signal that a component represents and the largest component of noise. As mutants differ less from the ancestor or fewer mutants differ from the ancestor in a given phenotype, the signal represented by this component is reduced, making it less likely to be detected by SVD. Similarly, phenotype detectability is decreased when the phenotype has reduced weighting for particular conditions or the phenotype affects fitness in fewer conditions.

Figure 2: Simulations: our method works in principle on simulated data with a couple of complications – (1) measurement error sets what we can detect, (2) spread of mutants/conditions also affects what we can detect

***Results Section 3: We can accurately predict fitness with 9 fitness-relevant phenotypes***

Now that we have a mathematical framework for identifying fitness-relevant traits for a set of mutants in a range of subtle perturbations, we turn to the collection of adaptive mutations that arose in a barcoded evolution experiment in a glucose-limited condition. We measured the fitness of this collection of mutants in 45 environmental conditions. (see Methods and Venkataram et al for fitness estimation details). After filtering for inclusion in every condition and sufficient coverage to acquire reasonable fitness estimates, we have a collection of 425 mutants for further study. Of these, Y have been sequenced, and D of them are diploids (the result of autodiploidization, see Venkataram et al). Of the remaining mutants, X of them are clearly adaptive (see methods), with various mutations causing their fitness benefit.

In order to capture the phenotypes relevant to the behavior of these mutants in the evolution condition, we focus on the use of subtle environmental perturbations. Of course, in order for these subtle perturbations to be informative, we need to be able to accurately measure the fitness of mutants in these subtle perturbations. The use of DNA barcodes allows us to precisely estimate the fitness effects of many mutants simultaneously. To assess the precision of these measurements, we re-barcoded two mutants, giving us on the order of 10 barcodes labeling the same mutant in the same flask, allowing us to assess the reliability of fitness estimates within flasks, across biological replicates, and across batches and conditions. We are able to precisely measure mutants in the same flask [Figure?] and see similar responses across replicate flasks. However, despite this precision, we see substantially larger variation in fitness across suites of replicates of the same environment done in different batches on different days. These substantial “batch effects” are real and may reflect subtle differences in media, incubator temperature, or other uncontrolled variables.

In addition to 9 batches of the evolution condition (M3), our 45 environmental perturbations include a range of intentional perturbations, where we’ve slightly changed the concentration of glucose, included additional carbon sources, changed the concentration of salt in the media, added drugs, and a suite of other changes. Ranking conditions by the average deviation from the variation across batch conditions for a balanced set of adaptive mutations, we see a range of effect from less than 1 standard deviation, representing a very subtle environmental perturbation from the M3 evolution condition to extremely strong perturbations up to 15 standard deviations away from the evolution condition (Fig. 3A). Of the stronger conditions, many exhibit clear fitness differences for some or all of the recurrent mutations (Fig. 3B). Interestingly, some conditions exhibit interesting behavior for only a subset of the recurrent mutation types. For instance, only GPB2 mutants have a clear fitness difference from the evolution condition in the 1 Day Transfer environment. This demonstrates that there are real phenotypic differences amongst these various adaptive mutants, including between GPB2 and PDE2, despite them both being negative regulators in the RAS/PKA pathway having similar fitness effects in the evolution condition, when we consider more distant environmental perturbations.



**Figure 3. Measuring fitness for a collection of adaptive mutants across many environments has the potential to reveal real biological signal. (A)** Conditions are ordered based on similarity to the average across all batches. Conditions where the balanced recurrent mutations are less than two standard deviations different from the evolved condition are denoted in black and make up the subtle perturbation set. Conditions where the aggregate behavior exceeds two standard deviations are shown in red and make up the strong perturbations. **(B)** For each mutation type, we take the average fitness across the evolution condition batches and the standard deviation of this behavior –Shaded regions represent two standard deviations away from the mean amongst the batch conditions per mutation type. On the right are kernel density estimates for the distribution of each mutation type across all conditions (two standard deviations around batches are shaded).

Because only the phenotypes relevant to fitness in the evolution condition matter to how particular mutants behave in that environment, we are interested in detecting only these fitness-relevant phenotypes. To do this, we rely on the idea that subtle environmental changes will change the relative importance of these fitness-relevant phenotypes but not change which phenotypes are relevant. Thus, it is key that we identify the relevant phenotypes using subtle, rather than strong, environmental perturbations for the collection of adaptive mutants. Using the average Z score for measure of similarity, we partition our data into two sets of conditions: (1) the “subtle perturbation set,” consisting of all M3 conditions and those within 2 standard deviations of the M3 evolution condition average according to the average Z score, and (2) the “strong perturbation set,” which contains the remaining conditions that are more than two standard deviations away.

First, we will construct a phenotype space from the subtle perturbation set, using cross-validation to identify the number of phenotypes to include in the space that gives us maximum explanatory power without overfitting measurement error. Next, using this phenotype space trained only on subtle perturbations, we make predictions about the fitness of mutants in the strong environmental perturbations.

Of course, in addition to the choice of conditions, the phenotype space we construct dependent on the choice of mutants used to construct it. Our method relies on the key assumption that the collection of mutants is sufficiently diverse to reveal a suite of phenotypic responses to subtle perturbations. In the extreme case where adaptive mutations are all of a single type, for example in the case of resistance to drugs targeting particular genes [cite some stuff], our method should only identify a single phenotype. Similarly, if instead of using a balanced collection of adaptive mutants, we use a biased collection, we will also get a biased perspective of the phenotype space. Of course, we do not have knowledge of which mutants have similar or distinct phenotypic signatures *a priori*, so we use putative causal genetic change for each of these adaptive mutants as identified in Venkataram et al. This allows us to partition the mutants into balanced training and testing sets (see Table).

Now armed with our mathematical framework, balanced set of mutants, and these mutants measured in a range of subtle environmental perturbations, we can identify the fitness-relevant phenotypes associated with this collection of mutants in the evolution condition using subtle environmental perturbations. By repeatedly dividing the subtle conditions into a training and test set, we can use bi-cross validation … the best predictive power is with 9 components included in the model. Additionally, because repeatedly dividing the data into these two sets can alter the

We construct a 9-component phenotype space using the entire set of subtle environmental perturbations and the training mutants. Using the phenotypic components of the mutants and weightings of subtle conditions in this space, we explain up to X% of variation in fitness for this set of mutants and conditions, indicating we capture most of the behavior in these subtle perturbations. We can use this subtle perturbation space to make predictions of fitness in strong environmental perturbations by independently fitting the locations of the test mutants and the strong perturbation conditions and then evaluating our ability to predict the fitness of these test mutants in the strong environmental perturbations based on this initial space. Overall, this 9-component model explains the most variation in the data (65% weighted Co. of D.s overall).

For all the test mutants, we generally explain more variance in the fitness values than a model with only a single phenotypic component or the average of permutations that remove the correlational signal of mutants and conditions in the training set, indicating that we are learning about phenotype beyond average fitness across the subtle environmental perturbations, and that these subtle environmental perturbations do in fact contain information about the behavior of these mutants that allows us to make concrete predictions of fitness in other contexts. This ability to make predictions, however, is limited, and our predictive power declines as perturbations become stronger, suggesting that the phenotypic effects observed in the local neighborhood of conditions around the evolution condition may be unable to explain behavior in very different environments and that this predictive power may be limited to an intermediate scale.

*[paragraph detailing interesting, new things we’ve learned, with attention to specifics in fig 4]*

Though it is clear that this phenotypic model can make predictions of aggregate mutant behavior across a variety of conditions an intermediate distance away from the evolution condition, it remains to be shown that we are learning real phenotypic differences and behavior for these mutants. In particular, our nine phenotype model accurately estimates fitness for the cases where recurrently hit mutations exhibited clear fitness differences from the evolution condition (Fig. 4). This includes cases where only a subset of the mutations showed such a difference as in the case of GPB2 in 1 Day environment.

This makes it clear that we are able to measure the effect of these mutations across subtle perturbations well, and that mutant behavior across these conditions represents real biology.



**Figure 4. 9-component model can accurately predict fitness of held-out mutants in strong conditions.** **(A)** Predictions from the 9-component model are typically better than the 1-component mode (open circle) and the average of 1000 permutations (black line, each permutation shown in gray). Gray and white background for eye-guiding purposes only. Comparison of predictions of the 1- and 9- component models for all held-out mutations in Baffle + 1.8% Glucose **(B and C)**, 1 Day **(D and E)**, 1% EtOH **(F and G)**, Baffle + Benomyl **(H and I)**, and 0.5M NaCl **(J and K)** conditions, respectively. Note that less than zero indicates that the prediction is worse than using the mean fitness in that condition and that is weighted according to the number of each mutation type present in the held out data (see Methods for details). Points in B-K colored by the mutation type. For a full set of prediction comparisons see Supplement.

***Results Section 4: The phenotype space itself reveals real biology***

*[what mutants are close to each other in phenotypic space? Do “pathways” cluster together/have similar phenotypic responses? What about genes? Are there interesting, unique members of particular genes?]*

Beyond being able to predict the fitness of these adaptive mutants in other environments, our approach allows us to disentangle differences between these mutants by considering the locations of the mutants in the phenotype space itself.

“Pathway”

Genetically, the adaptive mutants observed tended to fall into four broad classes: autodiploids, RAS/PKA pathway mutations, TOR pathway mutations, and other mutations that did not fall into the other three classes. Do these broad classes of mutation fall out of our subtle phenotype space? To assess this, we use hierarchical clustering to cluster mutants based on their locations in the subtle perturbation space [See figure]. The two broadest clusters pulled out represented a divide between haploids and diploids. Within the diploid cluster, there are three major groups: diploids with a chromosome 11 amplification, diploids with no additional mutations as well as those with additional mutations in RAS/PKA pathways, and a third group of six diploids that have higher that average fitness in the evolution condition but no known additional mutations. This set of six diploids also falls out visually when using t-SNE to visualize the 9-dimenionsal phenotype space [another figure?]. [something about what these diploids are actually doing??] [figure like the genes but by pathway??]

Amongst the haploid cluster, RAS/PKA pathway mutants do not cluster separately from TOR pathway mutations, suggesting there may not be a meaningful difference in the fitness-relevant phenotypes perturbed by mutations in these pathways. However, there do seem to be strong clustering based on the genotype of the mutants. In particular, GPB2, PDE2, GPB1, and IRA1 nonsense mutations cluster closely with each other and are close in phenotype space. This is consistent with our groupings in Figure 3, though our initial decision to include them as meaningful sets of mutations was based in their consistent behavior in M3, which also happens to be reflected in their fitness-relevant phenotypes as identified across a range of subtle perturbation. Because these seem to have clustered responses, we can also study the behavior of these genes by considering looking at the relative distances of the centroid in the 9-component phenotype space and comparing this to our expectations of the similarity of mutations in these genes from their behavior in M3 alone. In particular, we can see that mutations in GPB2 and PDE2 were expected to have similar phenotypic response due to their similar fitness in the evolution condition as well as both being a part of the RAS/PKA pathway. However, we can detect that mutations in these two genes do have very different phenotypic consequences, looking roughly as distant from each other as the average distance between any pair of mutants in the 9-component space [fig 5].

Despite many mutations clustering by gene, there are cases in which mutations in the same gene do not lead to closely-related phenotypic consequences. For instance, IRA1 missense mutations have a large array of phenotypic responses, not clustering with each other nor with IRA1 nonsense mutations [see dendrogram, t-SNE plot]. Interestingly,

We can also study the properties of other adaptive mutations with no known mutations in nutrient-response pathways. Of these adaptive mutations, one is very close to the GPB2 centroid, and is statistically indistinguishable from the distance of the GPB2 mutants from the centroid [need some stats]. We then manually investigated genome wide sequencing data for this mutant and identified a previously uncalled GPB2 mutation, indicating that this similarity was in fact due to this mutation.

Another of the adaptive mutations clusters closely to IRA1 nonsense centroid [again, stats]. This mutant has low coverage in the IRA1 gene, despite otherwise high coverage, suggesting either a deletion of the gene or insufficient coverage to identify a putative causal mutation.

The remaining “other” adaptive mutants…

[Diploid + IRA1 epistasis?]

[IRA2?]

[Are the particular pairs of points that appear to be close but actually have substantially different fitness effects across other conditions?? (for below part)]

[Are we learning anything more than what correlation patterns alone would show us? ]

Despite mutation identity being hidden from the model as it assigns new mutants to locations in the phenotype space, mutations in the same genes tend to cluster near each other in phenotype space (Fig 5) indicating that phenotype space location gives information about the phenotypic identity of our mutants. However, there are cases (IRA1 missense and nonsense) in which the mutation types appear to be more distant from each other than other like types. This represents phenotypic heterogeneity in mutations in these genes, which is also observed from the fitness effects of these particular mutations in even just the evolution condition alone.

Furthermore, our nine phenotype model identifies strong differences between mutation types that are not immediately clear from behavior in a one component model or from fitness alone. In particular, GPB2 and PDE2 mutants have similar fitness in the evolution condition, and from a one component model, the geometric median of the locations of the mutants of these types appear to be located in similar locations in the space (Fig 5A). However, in the full nine component model from the subtle perturbations, the mutants are as different from each other as other pairs of recurrent mutations, suggesting they do have distinct phenotypic effects. This phenotypic prediction for these mutations types, learned from subtle perturbations alone, is observed in the strong perturbations including differences between the genes in the 1 Day transfer condition, and the high salt concentration conditions.

***Results Section 5: Missing predictive power reveals existence of context-dependent pleiotropy***

Figure 6: Inability to detect particular behavior represents unpredictable pleiotropic effects not present/relevant to evolution in M3.

[paragraph: it’s not important that it’s exactly 9 dimensions but rather that it’s not 1 and not infinite, there are a finite number of phenotypes that matter to fitness in this environment] – limit of detection stuff here?

Ok – so now it’s clear that we are learning real biology from this approach. What is going on with our inability to completely detect fitness in all strong environments? Our model is clearly missing some information about these mutants that’s informative to fitness in these more distant environments. Our inability to complete predict fitness in, say, 0.5M NaCl environment is due to additional, undetected behavior of our collection of mutants in this environment. This extra behavior is from components that are undetectable from our suite of subtle perturbations (we know that our limit of detection is X size), reflecting either a large change in one of these undetectable components or an ensemble of these undetectable components adding to a significantly more substantial contribution to fitness in this strong environmental perturbation. [is there a way to tease these apart in simulation??]

One possibility is that we were not exhaustive enough with our set of subtle perturbations – maybe we would have <- is there a way to get at this by making a space with a stronger set? “there exist other subtle things that would have been able to make this prediction…” [something to do with variance explained in the M3??] [i.e. can this all be “hidden” away in the first component?]

For instance, in the 1.5% glucose evolution condition, cells experienced relatively little osmotic stress. However, in a condition with 0.5M KCl or 0.5M NaCl, the cells experience high levels of osmotic stress. Thus, the yeast’s osmotic stress response phenotypes represent undetectable, irrelevant traits for fitness in the evolution condition but suddenly become important phenotypes in these strong perturbations. Similarly, [something about trehalose in stationary phase for 7 day condition].

Of course, selection may be able to detect traits below our limit of detection of measurement, though this suggests that the traits that are (detectably) fitness-relevant in any given environment are relatively few (on the order of 10), providing an avenue for a large fraction mutations to have a net beneficial effect in a given condition.

The importance of subtlety? Is it really that the “meaningful” signal is swamped? Do we no longer learn the important stuff we learned previously?

**Discussion**

*[ this approach is a “new tool” through which we can identify fitness-relevant behavior].*

*Some other possible applications:*

*perturbing ecological communities to identify functional classes of behavior and ecological interactions*

*??*

*Possibility of using this approach to* *identify causal lower-level molecular phenotypes (via RNA-seq data or other phenotypic data) ]*

Comparing adaptive mutants to a collection of “completely random” mutations (i.e. deletion collection or the like) could distinguish between “dimensionality of yeast in this environment” and “dimensionality of adaptation to this environment”. Theory predicts that we should see that these adaptive ones perturb a subset of these conditions. Moreover, we don’t have a complete (nor unbiased) collection of adaptive mutants, so additional routes could exist – these just seem to be the most commonly taken – [something about mu x s].

Additionally, this set of mutants is on average one adaptive mutation from the ancestor. Do mutational routes change over time? Are there particular fitness components that represent easy, accessible routes? How do these accessible routes change over evolutionary time? Do additional mutations on the background of these first steps perturb the same phenotypes or do new phenotypic avenues open up?

What about evolving to multiple, different environments? For instance, if we evolved the same ancestral population to one of the stronger perturbations, say, 0.5M NaCl, do we see very different phenotypic responses? Do we see a similar failure to predict fitness of those mutants in M3?

This approach provides a framework in which to explore fitness-relevant phenotypes and to disentangle fitness-relevant, causal relationships from mere statistical associations. Moreover, our approach provides a way forward out of the gridlock of the modularity vs. universal pleiotropy debate in the context of rapid adaptation. Populations are able to adapt rapidly because a relatively small subset of traits matter to fitness in that environment.

**Methods**

**Estimating dimensionality using noise-only SVD detection.**

To estimate the number of components to include in the model using noisy data,

**Bi-cross validation scheme.**

**Simulating phenotype space.**

**Fitness Measurement details.**

**Calculating Weighted Coefficient of Determination**

**Calculating Weighted Average Z Score**

**References formatted for Cell**