

ARTICLE



<https://doi.org/10.1038/s41467-020-18694-0>

OPEN

Biophysical ambiguities prevent accurate genetic prediction

Xianghua Li ¹ & Ben Lehner ^{1,2,3}✉

A goal of biology is to predict how mutations combine to alter phenotypes, fitness and disease. It is often assumed that mutations combine additively or with interactions that can be predicted. Here, we show using simulations that, even for the simple example of the lambda phage transcription factor CI repressing a gene, this assumption is incorrect and that perfect measurements of the effects of mutations on a trait and mechanistic understanding can be insufficient to predict what happens when two mutations are combined. This apparent paradox arises because mutations can have different biophysical effects to cause the same change in a phenotype and the outcome in a double mutant depends upon what these hidden biophysical changes actually are. Pleiotropy and non-monotonic functions further confound prediction of how mutations interact. Accurate prediction of phenotypes and disease will sometimes not be possible unless these biophysical ambiguities can be resolved using additional measurements.

¹Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain. ²Universitat Pompeu Fabra (UPF), Barcelona, Spain. ³ICREA, Pg. Luis Companys 23, Barcelona 08010, Spain. ✉email: ben.lehner@crg.eu

A fundamental challenge across diverse fields of biology including human genetics, animal and plant breeding, and evolutionary theory is to predict how changes in genotypes result in changes in phenotypes and fitness. Accurate prediction of phenotypes from sequence entails two sub-challenges: predicting the mutations that individually affect a trait of interest and by how much, and predicting the joint effects when multiple mutations are combined in an individual. Progress is being made in both systematically identifying^{1–3} and predicting^{4–6} the mutations that impact traits of interest. Moreover, the extent to which mutations combine additively or with genetic (epistatic) interactions is being systematically quantified across diverse systems and phenotypes^{7,8}.

However, a more fundamental question remains that is not addressed in any of these studies. Even if we have perfect measurements of the individual effects of a set of mutations on a trait and a very good mechanistic understanding of a system, can we always predict what happens when two mutations are combined?

In this study, we use a simple biophysical system to address this question. We show that, for diverse biological systems, the answer to this question will often be no. The fundamental reason for this is that different combinations of biophysical parameters can give rise to the same phenotypic value⁹.

The phage lambda repressor, CI, is one of the best-understood proteins in biology and a classic model for gene regulation, protein biophysics and systems biology^{10–14}. CI regulates transcription from two divergent promoters with well-established dose–response curves: it represses transcription from the P_R promoter via a monotonic function but induces and then represses transcription from the P_{RM} promoter via a non-monotonic peaked function. The molecular mechanisms that underlie these regulatory responses are well-understood^{10,15,16} and thermodynamic models that incorporate them accurately predict the behaviour of the system^{17–20}. Specifically, Ackers' statistical thermodynamic model predicts the probabilities of the ON and OFF configuration states of the P_R and P_{RM} promoters as a function of the total repressor concentration¹⁷. To predict how mutations that affect the stability of CI combine to affect gene regulation, Ackers' model can be combined with a thermodynamic model of protein folding¹⁹.

Like most proteins²¹, CI is multifunctional: in order to regulate transcription it must fold correctly^{22–25}, form a dimeric complex²⁶, bind to DNA at multiple operator sites^{27,28} and also form a higher-order tetrameric complex^{29,30} on the genome (Fig. 1a). Mutations in CI can affect any of these biophysical activities, making CI a good model for investigating how mutations with different biophysical effects interact to alter cellular phenotypes.

However, mutations in a CI, like mutations in other proteins, can actually affect more than one biophysical parameter at the same time. For example, of 12 mutations that alter the binding affinity of CI to DNA, six (50%) also affected the stability of the protein^{27,31–33}. Such biophysical pleiotropy is common, for example, mutations that alter enzymatic activity often reduce protein stability³⁴. Similarly, mutations that alter protein binding affinities also frequently impact stability^{31,35} and in allosteric proteins changes in the affinity of binding at one site will alter the binding affinity at a second site³⁶.

Here, using gene regulation by the lambda repressor model, we show that, even for a very simple biophysical system, it is often impossible to predict what happens when two mutations are combined even if we have perfect measurements of their effects on a trait. The cause of this apparent paradox is the one-to-many mapping between phenotypes and the underlying biophysical parameter changes that can cause them. When combining mutations, the outcome can be very different depending upon what these unidentified biophysical changes actually are. Our results illustrate how accurate genetic prediction of phenotypes and disease will often not be possible unless additional

measurements are made to resolve the biophysical ambiguities in genotype–phenotype maps.

Results

Combining mutations in a thermodynamic model. To better understand how genetic variants with different biophysical effects combine to alter phenotypes, we investigated how mutations in a model transcription factor, the lambda repressor (CI), alter the expression of two target genes using an extensively validated thermodynamic model (Fig. 1b)^{17–20}. We first considered mutations that affect the folding or stability of CI. Changes in protein stability are one of the most frequent effects of amino acid changes and a major cause of genetic disease^{22–25}. The fraction of a protein in its natively folded state depends on the difference in Gibbs free energy (ΔG) between its folded and unfolded states. Unless they are energetically coupled³⁷, mutations have effects on stability that are additive at the level of free energy but non-additive for changes in protein concentration and expression from the P_R and P_{RM} promoters, which are our two phenotypic traits of interest (Fig. 1c, d)^{19,38,39}.

Genetic prediction for mutations affecting protein stability. If two mutations that only affect protein stability are combined, the change in expression from P_R is often non-additive (i.e. there is substantial epistasis)¹⁹. However, the phenotype of the double mutant can normally be unambiguously predicted from the phenotypes of the two constituent single mutants because the free-energy-phenotype function is monotonic⁴⁰ (Fig. 2a). The exception is when mutations have phenotypes that map to the top or bottom plateaus of the free-energy-phenotype function where the gradient approaches zero (Fig. 1d and Supplementary Fig. 1b–e) and measurement imprecision results in ambiguity in the underlying causal free-energy changes.

For expression from the P_{RM} promoter, however, this is not the case. Combining two mutations with measured effects on P_{RM} expression can result in more than one P_{RM} expression value, depending upon what the hidden underlying free-energy changes are^{19,40}. The cause of this ambiguity in the phenotype of a double mutant is the non-monotonic input–output function of P_{RM} (Fig. 1c, d), which means that many phenotypic values can map to two different underlying changes in the free energy of protein folding (Fig. 1d). Thus, when combining mutations of known phenotypic effect, there can be up to four different valid phenotypic outcomes in the double mutant (Fig. 2e) and these outcomes can differ by almost the entire phenotypic range (Fig. 2e, i). Thus, even if mutations only affect protein folding, non-monotonic input–output functions and plateaus in free-energy-phenotype functions can make it impossible to predict how two mutations of known effect will combine to alter a phenotype.

Mutations with other known biophysical effects. Mutations in proteins can, however, affect more than their stability. For example, mutations in CI can alter the binding affinity of the protein for itself (dimerization)²⁶, its affinity for DNA^{27,28} and the affinity between two dimers to form a tetramer^{29,30}. As for mutations affecting protein stability, mutations causing additive changes in the free energy of these molecular interactions (Fig. 1d) often combine to cause non-additive changes in expression from the two target promoters (Fig. 2b–d), generating substantial epistasis. However, for expression from P_R there is again no ambiguity in the double mutant phenotypes, with the exception of uncertainty created by imprecise measurements at the plateaus of the free-energy-phenotype functions (Fig. 1d and Supplementary Fig. 1b, c). However, as when combining mutations that only affect protein folding, pairs of mutations of known

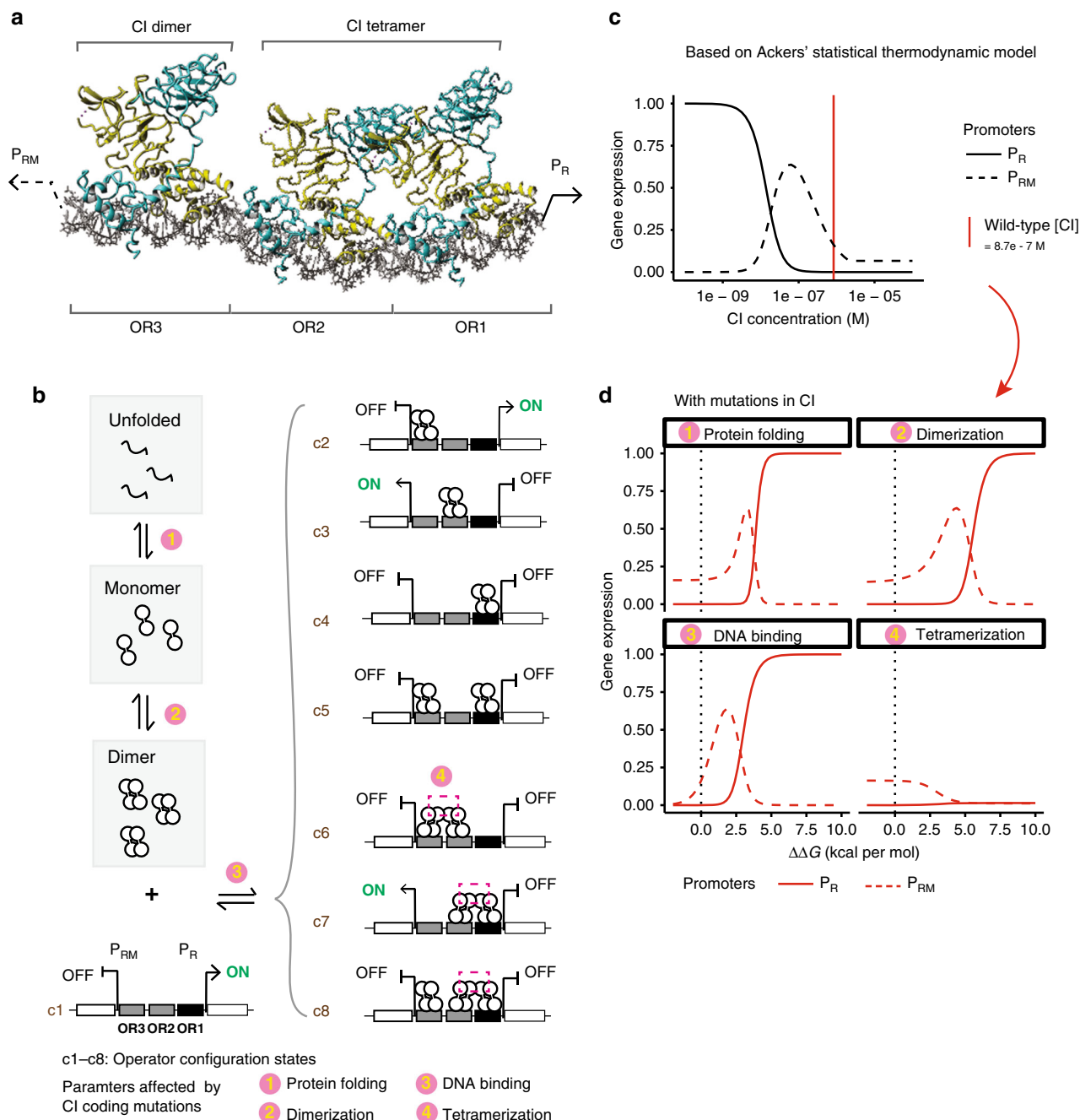


Fig. 1 Genetic interactions in a transcription factor. **a** CI binds three operators as a dimer with two dimers also forming a tetrameric complex. Cyan and yellow distinguish the two monomers of each dimer. **b** Statistical thermodynamic model of gene regulation by the lambda repressor (CI). CI exists as unfolded, folded monomer, free dimer and dimers that are bound to operators. The partitioning of these molecules depends on Gibbs free-energy differences between states. **c** Dose-response curves of the P_R and P_{RM} promoters. **d** Mutations result in additive changes in the free energy of protein folding, dimerization, DNA binding and tetramerization. When only one free-energy term is altered, gene expression is altered by the eight plotted relationships. Dotted vertical black lines denote $\Delta\Delta G = 0$ (wild type). See also Supplementary Fig. 1. Source data are provided as a Source data file.

phenotypic effect that both only affect either dimerization or DNA binding can combine to have up to four different P_{RM} phenotypes as double mutants (Fig. 2f-k, Supplementary Fig. 2). Similar conclusions are obtained if the two mutations individually affect two different (but known) biophysical parameters: P_{RM} expression often cannot be unambiguously predicted, including when one of the mutations affects tetramerization (Supplementary Fig. 2b, c), while P_R expression can always be predictable without ambiguity (Supplementary Fig. 2a).

Prediction for mutations with unknown biophysical effects. So far, we have considered cases where we know the identity of the biophysical parameter affected by each mutation. But normally we actually do not know which biophysical property of a protein is altered by a mutation. For example, any measured change in P_R expression resulting from a mutation in CI could be caused by a mutation that affects folding, DNA binding or dimerization (Fig. 1d, mutations that affect tetramerization have a more limited range of phenotypic outcomes).

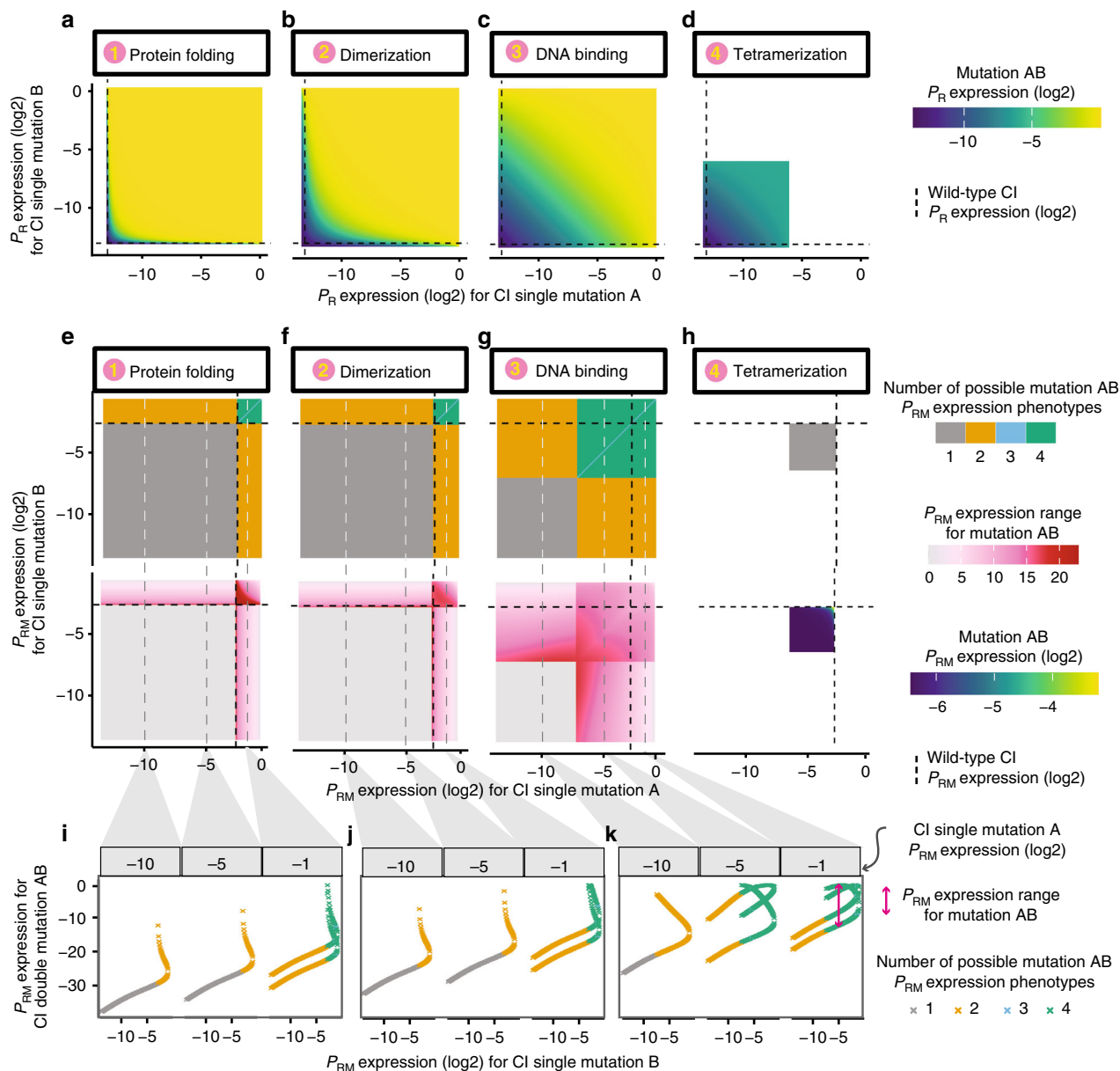


Fig. 2 Non-monotonicity results in ambiguous phenotype prediction. **a–d** Double mutant P_R expression when combining CI mutations both affecting the same biophysical parameter: protein folding (**a**), dimerization (**b**), DNA binding (**c**) or tetramerization (**d**). **e–h** Double mutant P_{RM} expression when combining CI mutations affecting the same biophysical parameter: protein folding (**e**), dimerization (**f**), DNA binding (**g**) or tetramerization (**h**). Top row panels show number of possible P_{RM} expression phenotypes when combining two single mutant phenotypes. Bottom row panels (**e–g**) show the range of possible P_{RM} phenotypes. Bottom row of (**h**) shows P_{RM} expression since there is no ambiguous prediction. **i–k** Examples showing how three mutations with known P_{RM} expression phenotypes combine with second mutations with known phenotypes to result in up to four different expression levels in the double mutant. Source data are provided as a Source data file.

We therefore considered what happens when two mutations combine and each of these mutations might have altered one of two different biophysical parameters, for example either protein stability or DNA-binding affinity. Now, even when considering expression from P_R as the phenotype of interest, there is always ambiguity when predicting the phenotypes of double mutants (Fig. 3a–f and Supplementary Fig. 3a–f). For example, there are now four valid phenotypic outcomes when combining two mutations if each can alter either stability or DNA binding (but not both, Fig. 3a–f). Considering expression from P_{RM} as the phenotype of interest, there are now many valid phenotypes for each double mutant

when combining mutations of known effect (Fig. 3g–l and Supplementary Fig. 3g–l).

If mutations can affect any one of the four biophysical parameters, the number of possible double mutant phenotypes can be very large indeed (Fig. 3m, n and Supplementary Fig. 3m, n). For example, two mutations with known effect on P_{RM} expression can combine to produce up to 15 different double mutant phenotypes if each mutation can affect any one (and only one) of the four possible free-energy terms (Fig. 3n). Thus, when we do not know the biophysical property of a protein that is altered by each mutation, it becomes impossible to predict the phenotypes of double mutants from the phenotypes of single mutants alone.

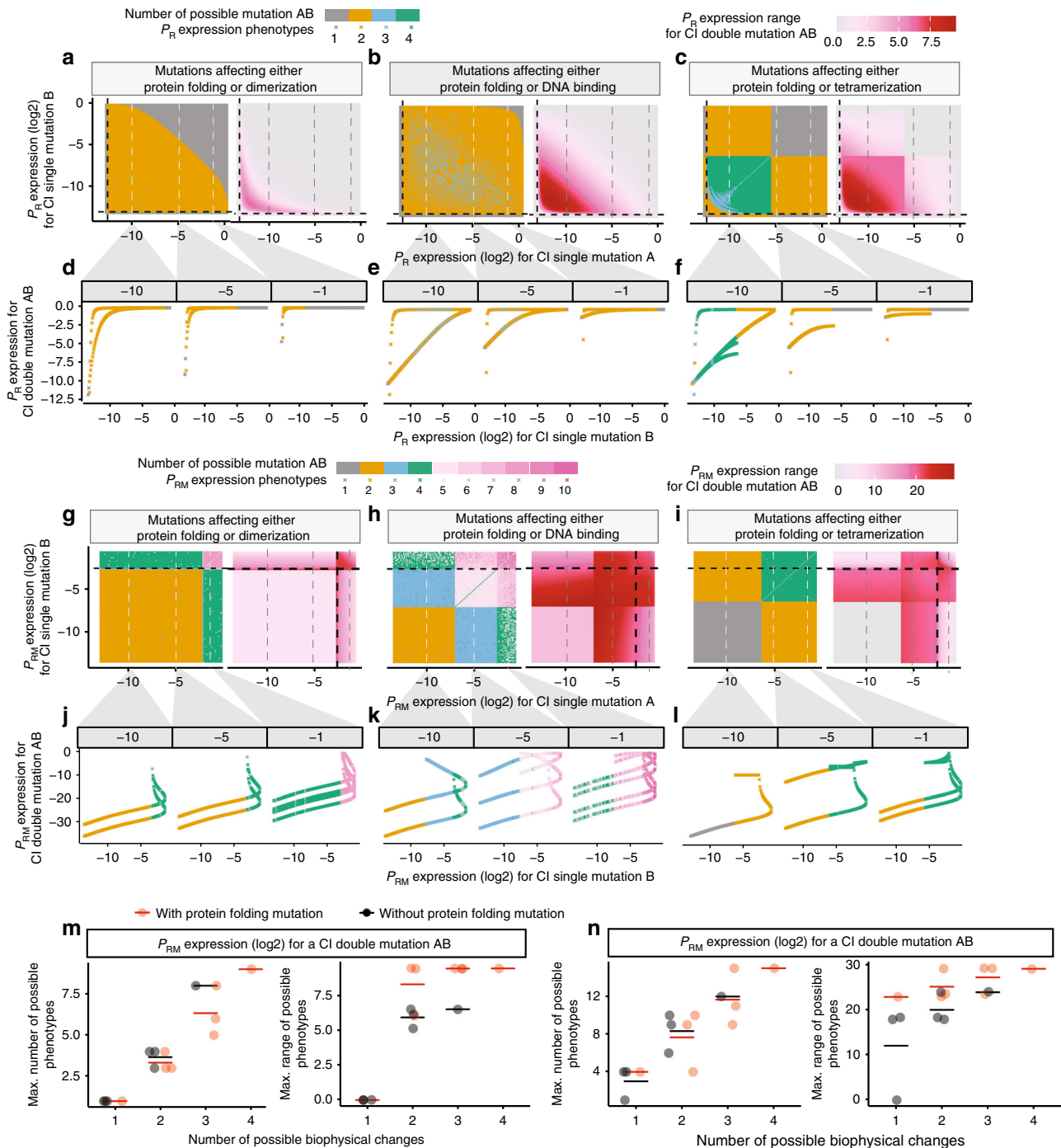


Fig. 3 Biophysical ambiguity prevents phenotype prediction. **a–c** P_R expression when combining two mutations that affect either protein folding (Mutation A) or another biophysical parameter (Mutation B) but not both: dimerization (**a**), DNA binding (**b**) or tetramerization (**c**). Number (left) and range (right) of possible double mutant P_R phenotypes (left). **d–f** Examples showing how a mutation with a known phenotype combines with other mutations, leading to 1 to 4 possible double mutant P_R expression levels. **g–i** Number (left) and range (right) of double mutant P_{RM} expression levels when mutations can affect folding or another biophysical parameter. **j–l** Examples showing how a mutation with a known P_{RM} phenotype can combine with other mutations to result in many different P_{RM} phenotypes. **m, n** Maximum number (left) and range (right) of double mutant phenotypes when two mutations can each affect one of the indicated number of different biophysical properties. Horizontal lines denote the mean of the data points. $n = 4, 6, 4$ and 1 , respectively, for the groups with number of possible biophysical parameters equal to 1, 2, 3 and 4. Source data are provided as a Source data file.

Biophysical pleiotropy further confounds genetic prediction. In reality, the situation can actually be worse than this because mutations can affect more than one biophysical parameter at the same time. For example, of 12 mutations changing the binding affinity of CI to DNA, half also altered the stability of the protein^{27,31–33}. We define these situations when one mutation

influences two or more biophysical parameters as biophysical pleiotropy.

Allowing one (Fig. 4a, b, Supplementary Fig. 4) or both (Fig. 4f, j and Supplementary Fig. 4) mutations in CI to be pleiotropic and to alter two different free-energy terms results in the possible double mutant outcomes now covering a continuous

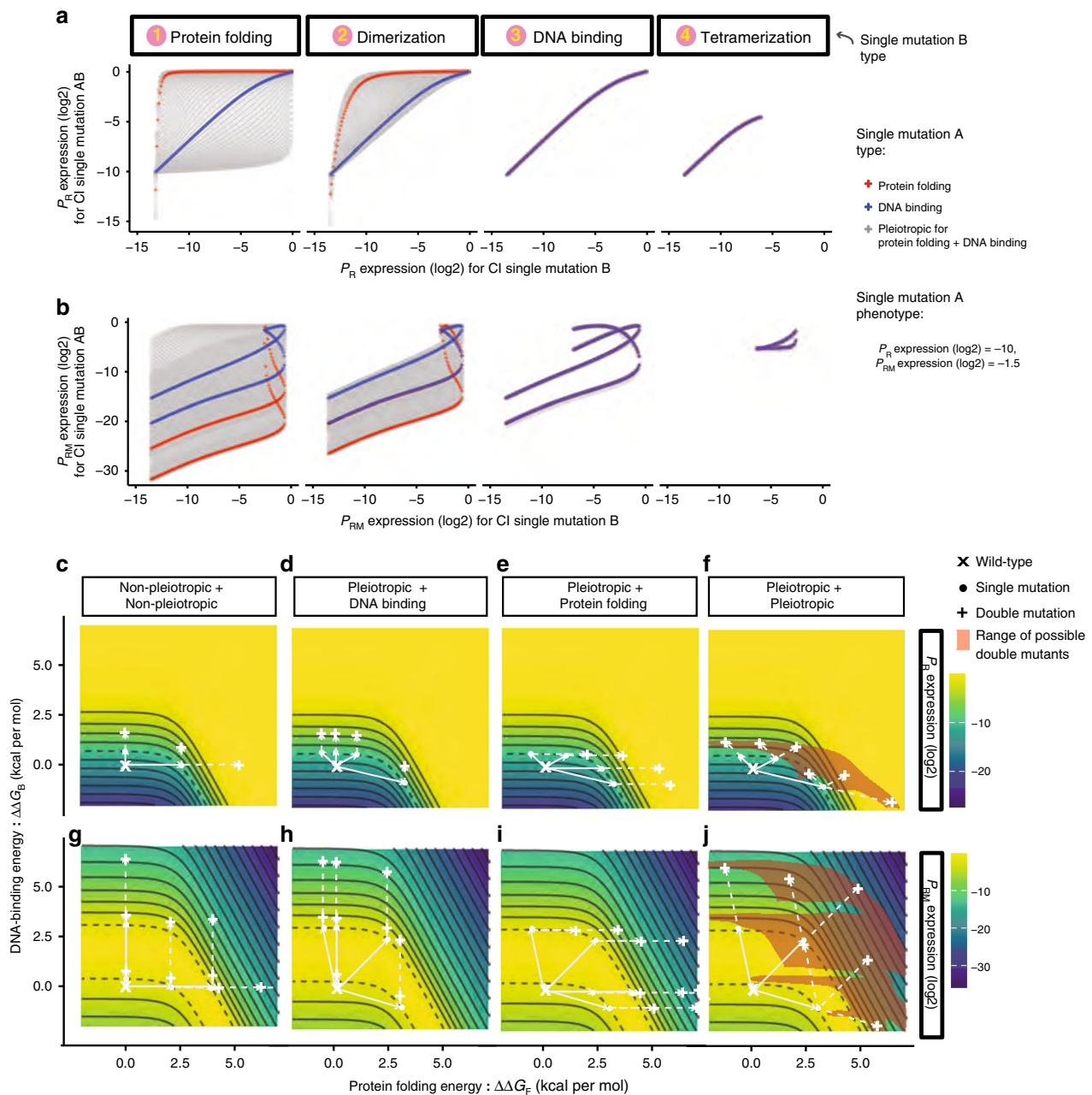


Fig. 4 Biophysical pleiotropy further confounds phenotype prediction. **a, b** Double mutant phenotypes when pleiotropic mutations (Mutation A) variably affecting folding and DNA binding to give P_R expression = -10 (**a**) and P_{RM} expression = -1.5 (**b**) in $\log(2)$ scale are combined with different types of mutations (Mutation B). **c–j** Free-energy-phenotype landscapes for mutations that affect the free energy of folding (x-axis) and/or DNA-binding energy (y-axis). Phenotypic isochores are drawn with an interval of 2 in $\log(2)$ scale. A continuous range of free-energy changes can underlie an observed phenotype (dashed isochore). Combining two mutations with the same effect can result in a range of double mutant phenotypes (red shaded areas in (**f**) and (**j**)). Example double mutant outcomes are shown when neither (**c, g**), one (**d, e, h, i**) or both (**f, j**) mutations are pleiotropic. Source data are provided as a Source data file.

range of values (Fig. 4 and Supplementary Fig. 4). Thus, when mutations are biophysically pleiotropic, we cannot predict the phenotype of a double mutant containing two mutations of precisely measured individual effects.

Biophysical ambiguity confounds genetic prediction. To illustrate how these diverse double mutant phenotypes arise when combining pairs of mutations with identical phenotypic effects, we plot in Fig. 4c–f how the expression from P_R changes as a function of changes in the free energy of folding ($\Delta\Delta G_F$) and DNA binding ($\Delta\Delta G_B$). Non-pleiotropic mutations that only alter

folding are horizontal movements in this space, mutations that only affect DNA binding are vertical movements and pleiotropic mutations are diagonal movements. All of the changes in free energy that result in the same phenotype form a phenotype isochore, for example the grey dashed curves in Fig. 4c–f represent all parameter changes that can produce a 4-fold increase (2 in $\log(2)$ scale) in P_R expression.

When two non-pleiotropic mutations that cause this same phenotypic change (lie on the same phenotype isochore) are combined together there are three possible combinations of free-energy changes (the two mutations alter DNA binding, folding, or

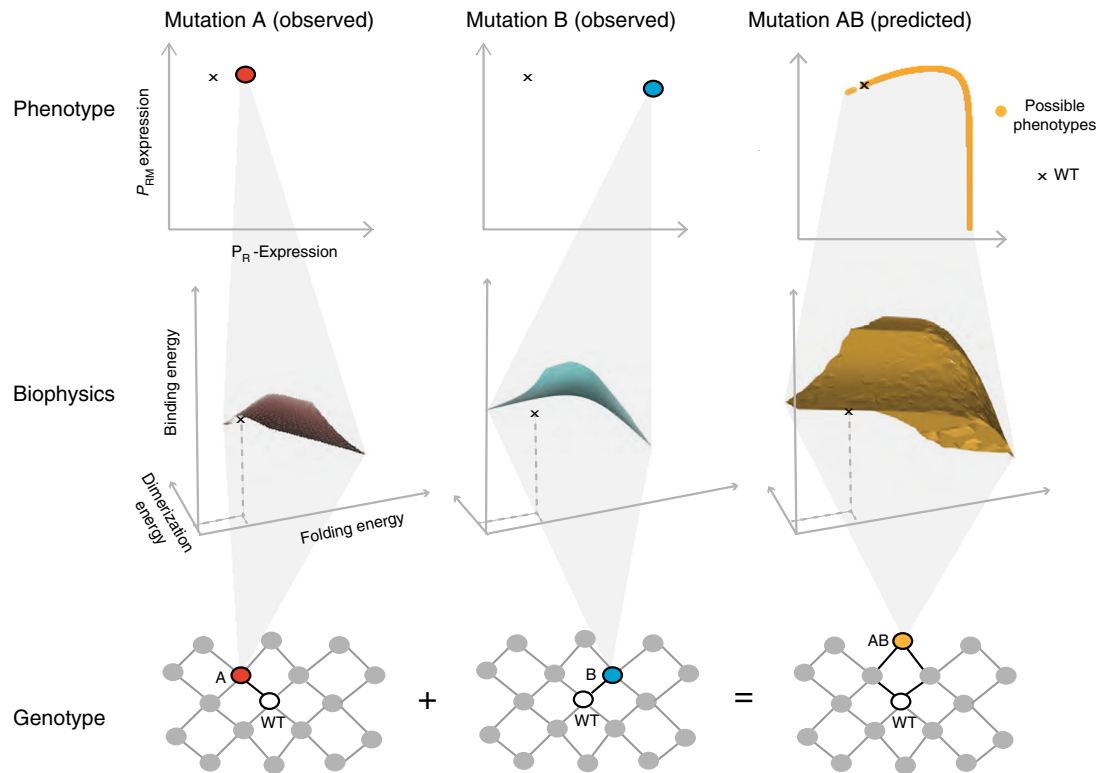


Fig. 5 Biophysical ambiguity as a hidden layer for phenotype prediction. Mutations can alter multiple biophysical properties to cause the same observed change in a phenotype. For example, the P_R and P_{RM} phenotypes of mutations A and B in CI could be caused by many different changes in three free-energy terms. When these two mutations are combined, the double mutant can have a phenotype spanning nearly the entire phenotypic range, depending upon what the hidden parameter changes are in each single mutant. Source data are provided as a Source data file.

one alters folding and the other binding) and two possible resulting double mutant phenotypes (Fig. 4c). When a non-pleiotropic mutation affecting DNA binding is combined with a pleiotropic mutation affecting both free-energy terms, there are many possible combinations of free-energy terms but, because of the topology of the free energy-phenotype landscape, all of the double mutants have very similar phenotypes (Fig. 4d). In contrast, when a non-pleiotropic mutation affecting folding is combined with a pleiotropic mutation, the possible double mutants do not fall on an isochore but now cover a range of possible phenotypes (Fig. 4e). Finally, when two pleiotropic mutations are combined, the possible double mutants are widely spread in the free-energy landscape (red shaded area in Fig. 4f) and take many different phenotypic values (Fig. 4f). The equivalent free-energy-phenotype landscape is plotted for P_{RM} in Fig. 4g–j and for other combinations of free-energy terms in Supplementary Fig. 4. It is both the monotonicity and symmetry of these landscapes that determines the degree of ambiguity when combining mutations.

When mutations can alter three or more free-energy terms, these landscapes become difficult to visualise (Fig. 5). For example, if each mutation in CI can alter stability, DNA binding or dimerization, each mutation with a known phenotype potentially maps to any position on a surface of combinations of causal parameter changes. Combining two mutations with precisely measured phenotypic effects can combine to have phenotypes that span nearly the entire range of possible phenotype values (Fig. 5). This is because, without additional information, the actual parameter changes in the double mutant can take many values within a 3D volume of possibilities. There is now nearly complete ambiguity in the predicted phenotype of the double mutant (Fig. 5).

Biophysical ambiguity in even simpler systems. Finally, although gene regulation by the lambda repressor is a relatively simple biological system, we note that biophysical ambiguity also confounds the prediction of double mutant phenotypes in even simpler systems. For example, consider a protein whose only function is to bind another molecule (a ligand), with the concentration of the bound complex directly proportional to the phenotype of interest (Fig. 6a). In such a minimal system mutations can only alter protein stability or the binding affinity to the ligand. The outcome in a double mutant can still differ depending upon which free-energy terms are individually affected in each single mutant (Fig. 6b, c). Again, allowing pleiotropic mutations further thwarts the ability to predict the phenotypes of double mutants from the phenotypes of single mutants (Fig. 6d, e). Similar conclusions are obtained using a model in which a protein's only function is to bind to itself to form a dimer (Supplementary Fig. 5). Thus, even in these most basic biological systems of a single binding reaction of a macromolecule, it is often impossible to predict what happens when single mutants of known phenotype are combined without additional measurements or inferences.

Discussion

Taken together, our results show that, even for a simple biological system—the regulation of gene expression by a single transcription factor—it is often impossible to unambiguously predict how two mutations of known phenotypic effect will combine together to alter the same phenotype in a double mutant.

The fundamental cause of this uncertainty is the one-to-many relationship between a measured phenotype and the underlying causal changes in biophysical parameters. Mutations can affect multiple biophysical properties of a system—for example, the

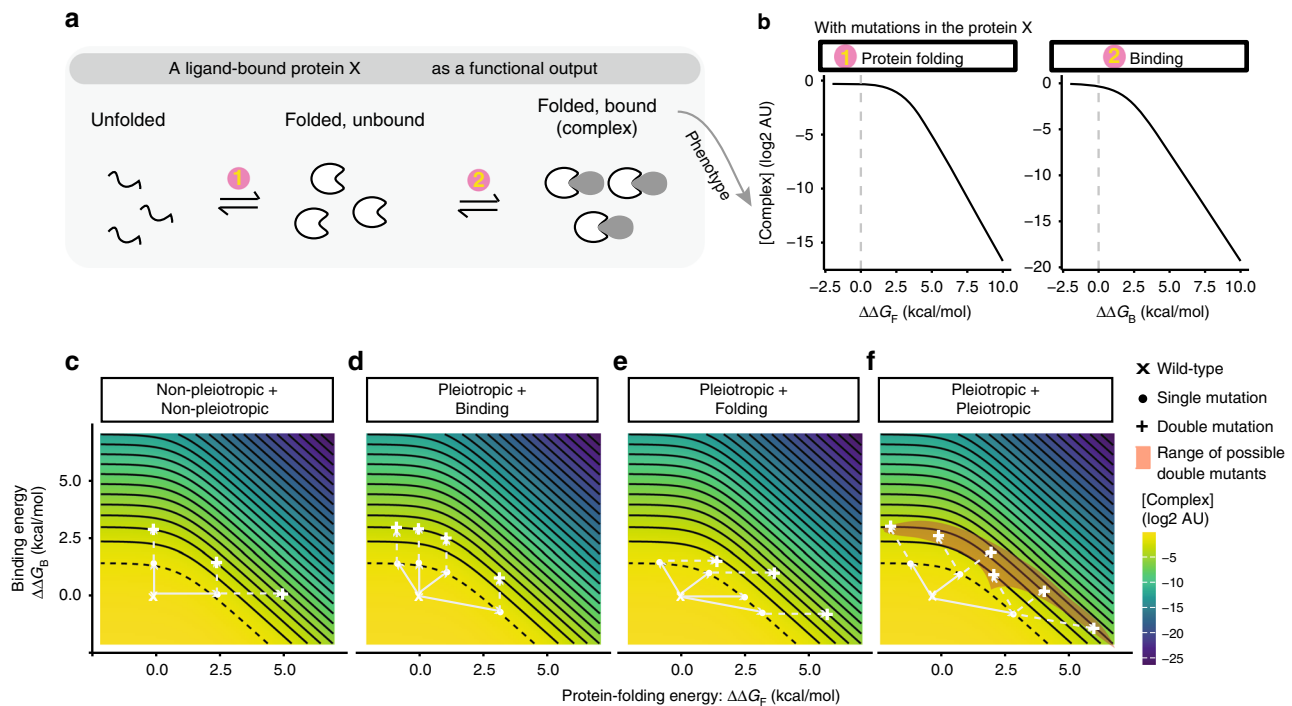


Fig. 6 Biophysical ambiguity in a protein-protein interaction system. **a** Statistical thermodynamic model of a protein binding to a ligand. The protein X exists in three states: unfolded, folded, and folded and bound to the ligand. The partitioning of these molecules depends on the Gibbs free-energy differences between states. **b** Mutations result in additive changes in the free energy of protein folding and binding, altering the concentration of the protein-ligand complex. **c–f** Free-energy-phenotype landscapes for mutations that affect the free energy of folding (*x*-axis) and/or binding energy (*y*-axis). Phenotypic isochores are drawn with an interval of 1 in log(2) scale. A continuous range of free-energy changes can underlie an observed phenotype (dashed isochore). Combining two mutations with the same effect can result in a range of double mutant phenotypes (red shaded areas in **f**). Example double mutant outcomes are shown when neither (**c**), one (**d**, **e**) or both (**f**) mutations are pleiotropic. See also Supplementary Fig. 5. Source data are provided as a Source data file.

stability and binding affinities of proteins—and many different changes in biophysical parameters can cause the same observed change in a trait. However, the phenotype of a double mutant depends on which of these biophysical properties is actually altered in each single mutant and so can take multiple values. Pleiotropic biophysical effects and non-monotonic input–output functions create further ambiguity when predicting how mutations of known effect combine to alter a phenotype.

The extent to which biophysical ambiguities will thwart the prediction of different phenotypes will depend on the number of parameters that can be affected by mutations, their biophysical pleiotropy, and monotonicity of input–output functions. The distributions of mutational effects on multiple biophysical parameters have been quantified for very few systems, but for both the lambda repressor and other proteins, mutations frequently affect both stability^{41,42} and binding to interaction partners^{41,43,44} with biophysical pleiotropy and non-monotonic functions also common^{31,35,45}. In other words, we expect biophysical ambiguity to confound phenotypic prediction in other systems including heteromeric complexes and beyond transcription factor-mediated repression.

To resolve ambiguities and accurately predict how mutations combine to alter phenotypes, additional information will always be required. Although ultimately it may be possible to predict from sequence how a particular mutation affects all the biophysical parameters of a protein, for the foreseeable future resolving ambiguities will require additional measurements to be made. High-throughput methods to quantify the effects of mutations on protein stability⁴², binding^{41,44,46} and activity⁴⁷ will help in this endeavour, particularly when used in combination to disentangle

biophysical effects. Moreover, quantifying how individual mutations interact with many other mutations in a system may allow the underlying causal changes in biophysical parameters to be inferred, at least when only two different parameters can be affected³⁵. Quantifying intermediate molecular phenotypes such as protein concentrations and additional higher-level phenotypes may also be useful (e.g., quantifying expression from P_R is sufficient to resolve the ambiguities resulting from the non-monotonicity of the P_{RM} dose–response curve), and experimentally quantifying the dose–response curves of individual mutations can also sometimes help to distinguish mutations with different biophysical effects⁴⁸.

However, the fundamental conclusion remains: even in this simple biological system (and in even simpler ones, Fig. 6 and Supplementary Fig. 5) it can be impossible to predict the combined effect of two mutations, even if we have perfect measurements of their individual effects on a trait. In such cases, additional information or measurements will always be required to accurately predict how genetic variants combine to alter phenotypes and cause disease.

Methods

Methods overview. Our model is based on Ackers' thermodynamic model of lambda repressor binding to its operator sites (O_{R1} , O_{R2} and O_{R3})¹⁷. Briefly, this model describes eight possible operator configuration states (c1–c8) in which the CI dimer can bind to the operators (Fig. 1b). Based on statistical thermodynamics, the downstream gene expression from promoters P_R and P_{RM} is determined by the probabilities of the ON and OFF *cis*-regulator configuration states¹⁷.

To examine CI coding mutants' effects on gene expression from P_R and P_{RM} promoters, we extended Ackers' model by including CI folding because many mutations destabilise proteins^{22–25}. Destabilising mutations will decrease the fraction

of the folded functional protein, and thus change gene expression from the downstream P_R or P_{RM} promoter. In other words, compared to Ackers' model, we have one more protein state—CI unfolded state $CI_{(U)}$ and the corresponding additional parameter—protein-folding energy ΔG_F (Supplementary Tables 1 and 2). The rest of our model is the same as Ackers' model. We consider the system as a single equilibrium, i.e. protein folding and dimerization are coupled reactions.

Below are the details of the model, which follow simple statistical thermodynamics.

CI configuration states. The total CI ($CI_{(Total)}$) molecule amount is the sum of all the CI molecules in the 10 different possible states as shown in Eq. (1). These different states include unfolded $CI_{(U)}$, folded monomer $CI_{(M)}$, free dimer CI_2 and seven operator-bound CI dimer states (Fig. 1b and Supplementary Table 1). The unit of molecule amount per cell is M in all the equations in our model.

$$CI_{(Total)} = CI_{(U)} + CI_{(M)} + 2 \cdot CI_2 + 2 \cdot OR_{(Total)} \sum_{i=2}^7 (k \cdot f_i) \quad (1)$$

Above, $OR_{(Total)}$ is the molecule amount of the operators, f_i is the relative probability that each of the seven *cis*-configuration states where CI is bound to operators occurs in relation to the not-bound state. i is the index for each *cis*-configuration state, and k is the number of CI dimers in the corresponding *cis*-configuration state (Supplementary Table 1). The amount of CI molecule for each operator-bound state is calculated based on the statistical thermodynamics but also multiplying the number of CI dimers (k) in each state and a factor 2 to account for two molecules for each dimer (Supplementary Table 1).

All the parameters in the model for wild-type CI are taken from literature (Supplementary Table 2).

Equilibrium between CI unfolded and folded monomer states. CI monomer folds in a simple folded $CI_{(M)}$ and unfolded $CI_{(U)}$ two-state fashion⁴⁹ that can be described as in the equation below:

$$\frac{CI_{(M)}}{CI_{(U)}} = \exp\left(\frac{-\Delta G_F}{RT}\right) \quad (2)$$

ΔG_F is the free-energy difference between the folded monomer and unfolded states of CI molecule. R is the gas constant ($R = 1.98 \times 10^{-3}$ kcal per M) and T is the absolute temperature for 37 °C (310.15 Kelvin).

Equilibrium between folded CI monomer and free dimer states.

$$\frac{CI_2}{CI_{(M)}^2} = \exp\left(\frac{-\Delta G_D}{RT}\right) \quad (3)$$

Equilibrium between free CI dimer and operator-bound states. We use Ackers' model to describe these relationships. Briefly, the likelihood of each configuration state (c1–c8 based on the *cis*-regulatory state) is a function of the binding energies and the free CI protein dimer concentration.

The probability that each of the eight *cis*-configuration states (f_i) occurs is:

$$f_i = \frac{\exp\left(\frac{-\Delta G_i}{RT}\right) CI_2^k}{\sum_{i=1}^8 \exp\left(\frac{-\Delta G_i}{RT}\right) CI_2^k} \quad (4)$$

Where ΔG_i is the total free energy of lambda repressor dimers in the respective *cis*-configuration state $i \in [1, 8]$ (Supplementary Table 1, where ΔG is free energy, with ΔG_T referring to the cooperation energy for two dimers binding to the adjacent operator sites); the exponent $k \in [0, 1, 2]$ is the total number of the lambda repressor dimers in the corresponding *cis*-configuration state i . As stated earlier, all the parameters are kept as originally described in Ackers' model (Supplementary Table 2).

CI distribution based on statistical thermodynamics. By combining Eqs. (1)–(4), we can describe the total expression level of $CI_{(Total)}$ as a function of CI free dimer concentration and Gibbs free energies:

$$CI_{(Total)} = \exp\left(\frac{\Delta G_U + \Delta G_F}{RT}\right) CI_2^{0.5} + 2CI_2 + \frac{2OR \left(\sum_{i=2}^7 \exp\left(\frac{-\Delta G_i}{RT}\right) CI_2 + 2 \times \sum_{i=5}^7 \exp\left(\frac{-\Delta G_i}{RT}\right) CI_2^2 + 3 \exp\left(\frac{-\Delta G_8}{RT}\right) CI_2^3 \right)}{\sum_{i=2}^7 \exp\left(\frac{-\Delta G_i}{RT}\right) CI_2 + \sum_{i=5}^7 \exp\left(\frac{-\Delta G_i}{RT}\right) CI_2^2 + \exp\left(\frac{-\Delta G_8}{RT}\right) CI_2^3} \quad (5)$$

Probability of P_R —ON. CI represses expression from the P_R promoter by binding to the operator sites that overlap with the RNA polymerase sigma factor binding site (Fig. 1b)¹⁷. Based on Ackers' model, two out of the eight *cis*-configuration states fail to repress gene expression from P_R —when CI is not bound to any operators (c1) and when CI only binds to the low-affinity OR_3 (c2) (Fig. 1b, Supplementary Table 1). Therefore, the probability of the P_R promoter to be active (P_{Pr}) is the sum of the probabilities of the two configuration states in which

promoter P_R is not repressed ($\sum_{i=\{1,2\}} f_i$), as shown in Eq. (6)¹⁷.

$$P_{Pr} = f_1 + f_2 = \frac{\exp\left(\frac{-\Delta G_1}{RT}\right) CI_2^0 + \exp\left(\frac{-\Delta G_2}{RT}\right) CI_2^1}{\sum_{i=1}^8 \exp\left(\frac{-\Delta G_i}{RT}\right) CI_2^k} \quad (6)$$

Probability of P_{RM} —ON. CI not only suppresses P_R promoter but also activates or suppresses the divergently transcribed P_{RM} promoter in response to changes in the CI concentration in the cell (Fig. 1c)^{10,50}. When CI is present and binds to OR_2 , it activates the P_{RM} promoter, while binding to OR_1 per se does not have any effects on P_{RM} activity^{10,16}. On the contrary, once CI binds to the low-affinity OR_3 , it blocks the access of RNA polymerase sigma factor, repressing expression from P_{RM} ⁵¹. Therefore, gene expression from P_{RM} is activated only when CI is bound to OR_2 and not bound to OR_3 (corresponding to the two *cis*-configuration states: c3 and c7) (Fig. 1b and Supplementary Table 1). Using Ackers' model and Eq. (4)¹⁷, we describe the probability that the P_{RM} promoter is activated as follows:

$$P_{Prm} = f_3 + f_7 = \frac{\exp\left(\frac{-\Delta G_3}{RT}\right) CI_2^1 + \exp\left(\frac{-\Delta G_7}{RT}\right) CI_2^2}{\sum_{i=1}^8 \exp\left(\frac{-\Delta G_i}{RT}\right) CI_2^k} \quad (7)$$

Calculating free dimer concentration. As seen from Eq. (5), we can easily calculate $CI_{(Total)}$ from CI_2 for a given set of free energies but not CI_2 from $CI_{(Total)}$. Therefore, we performed a parameter search for CI_2 values with each set of known biophysical parameters (ΔG values) that minimizes the absolute differences between the provided $CI_{(Total)}$ value and $CI_{(Total)}$ calculated based on Eq. (5). The Optimize⁵² function in R was used for the parameter search, with the tol parameter set to $1e-23$. We refer to this process using Eq. (8), where ΔG_s are all the Gibbs free energies of the system.

$$CI_2 = f(CI_{(Total)}, \Delta G_s) \quad (8)$$

Biophysical changes to phenotypes. The probabilities of the two promoters' ON-states as phenotypes can be calculated using a set of biophysical parameters (free energies) and $CI_{(Total)}$. We call this process a Forward Function (see Code availability). This function is composed of two steps: (1) parameter search for CI_2 for the given CI as described in the previous section (Calculating free dimer concentration) using Eq. (8); (2) calculating P_{PR} and P_{PRM} based on Eqs. (6) and (7).

Phenotypes to free energy for non-pleiotropic mutations. Mutations in the CI protein can affect protein-folding energy (ΔG_F), dimerization energy (ΔG_D), binding energy to the operator sites ($\Delta G_{OR1-OR3}$) and tetramerization energy (ΔG_T) at the biophysical level. We assume that mutations in CI that alter the free energy of DNA binding do so by the same magnitude for all three operators ($\Delta \Delta G_B = \Delta \Delta G_{OR1} = \Delta \Delta G_{OR2} = \Delta \Delta G_{OR3}$). To calculate only one biophysical change that can lead to the phenotype, we reversed the Forward Function described in the previous section. The Reverse Function for both P_{PR} and P_{PRM} is composed of two sub-functions. The first sub-function is the above-mentioned Forward Function, which calculates phenotypes from biophysical changes. This function is written in the form of $y = f(x)$, where y is the phenotype and x is a set of biophysical parameters including the total expression level of CI. The second sub-function is an Inverse Function that finds all roots for an equation in the form of $y - f(x) = 0$. A root-finding process is performed using the uniroot.all function in the R package rootSolve⁵³. Specifically, for each perturbation of biophysical parameter ($\Delta \Delta G$), we looked for all the roots within a range of -2 – 10 kcal per mol, and returned the $\Delta \Delta G$ values that produce the phenotypes while the other biophysical parameters are not perturbed.

Mutational effects are modelled at a fixed expression level of CI ($CI_{(Total)} = 8.4e - 7M$) that corresponds to $\sim 99\%$ repression of the P_R promoter and the CI concentration in a lysogen^{17,19}. To calculate changes in the biophysical parameters for single mutants with known effects on expression from P_R or P_{RM} , we first generated 136 evenly spaced phenotypes (with an interval of 0.1 in log(2) scale from -13.5 to 0). Then, for a given phenotype, we calculated corresponding changes in any of the four free-energy terms (biophysical parameters), each time allowing only one biophysical parameter to change using the Reverse Function explained in the previous paragraph.

Phenotypes to free energy for pleiotropic mutations. For any given phenotype, we systematically searched for combinations of biophysical changes that can produce the phenotype. Taking a pleiotropic mutation affecting both protein-folding energy (ΔG_F) and DNA-binding energy (ΔG_B) as an example, we first generated a fixed range of $\Delta \Delta G_F$ (-1 to 5 kcal per mol with an interval of 0.05 kcal per mol). Then, for each $\Delta \Delta G_F$, we calculated $\Delta \Delta G_B$ that produces the given phenotype using the Reverse Function as described for non-pleiotropic mutations. For mutations affecting three biophysical parameters (protein-folding energy ΔG_F , dimerization energy ΔG_D and DNA-binding energy ΔG_B), we first generated all

possible two-way combinations of $\Delta\Delta G_F$ and $\Delta\Delta G_D$, each from defined ranges of -1 to 5 kcal per mol with an interval of 0.05 kcal per mol. For each combination of $\Delta\Delta G_F$ and $\Delta\Delta G_D$ with the given phenotype, we calculated $\Delta\Delta G_B$, using the Reverse Function as described for non-pleiotropic mutations.

Double mutant phenotypes from single mutants' phenotypes. For each double mutant, we simply added the changes in the free energies of both single mutants to the corresponding wild-type free energy. Then, we used the updated parameters to calculate the downstream phenotypes based on the Forward Function explained in the section of Phenotypes to free energy for non-pleiotropic mutations. Double mutants' phenotypes are rounded to 2 decimal places in $\log(2)$ scale in order to avoid counting phenotypes with very similar values as different phenotypes.

Thermodynamic model of simple protein interactions. We considered the protein of interest (that is mutated) to be in three different configuration states: (1) unfolded, (2) folded, and (3) folded and bound (or dimer) (Fig. 6a and Supplementary Fig. 5a). The steady-state equilibrium is in the same format as shown for CI protein in Eqs. (2) and (3). When protein binds to a substrate instead of to itself, it follows Eq. (9).

$$\frac{[\text{Complex}]}{[\text{ProteinX}] \cdot [\text{Ligand}]} = \exp\left(\frac{-\Delta G_B}{RT}\right). \quad (9)$$

Above, [complex] is the concentration of the bound Protein X to its ligand (or substrate molecule). The parameters we used in the model for Figs. 6 and S5 are $\Delta G_{F, WT} = -1$ kcal per mol; $\Delta G_{B (or D), WT} = -2$ kcal per mol. [Protein X]: [Ligand] = 1:1.

3D visualisation of CI bound to O_R1-3 . The 3D structure of CI bound to O_R1-3 was generated based on PDB structure 3bdn, using YASARA software (v 19.7.20).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data supporting this work are provided within the paper, the supplementary information and the source data file. Source data are provided with this paper.

Code availability

Scripts are publicly available from https://github.com/lehner-lab/Biophysical_Ambiguity. Source data are provided with this paper.

Received: 22 April 2020; Accepted: 4 September 2020;

Published online: 01 October 2020

References

- Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
- Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* **14**, 168–178 (2013).
- Starita, L. M. & Fields, S. Deep mutational scanning: A highly parallel method to measure the effects of mutation on protein function. *Cold Spring Harb. Protoc.* **2015**, 711–714 (2015).
- Shendure, J. & Akey, J. M. The origins, determinants, and consequences of human mutations. *Science* **349**, 1478–1483 (2015).
- Jelier, R., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Predicting phenotypic variation in yeast from individual genome sequences. *Nat. Genet.* **43**, 1270–1274 (2011).
- Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
- Domingo, J., Baeza-Centurion, P. & Lehner, B. The causes and consequences of genetic interactions (epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 083118–014857 (2019).
- Costanzo, M. et al. Global genetic networks and the genotype-to-phenotype relationship. *Cell* **177**, 85–100 (2019).
- Hartl, D. L., Dykhuizen, D. E. & Dean, A. M. Limits of adaptation: The evolution of selective neutrality. *Genetics* **111**, 655–674 (1985).
- Ptashne, M. *A Genetic Switch: Phage Lambda Revisited* (Cold Spring Harbor Laboratory Press, 2004).
- Sauer, R. T., Jordan, S. R. & Pabo, C. O. λ Repressor: a model system for understanding protein–DNA interactions and protein stability. *Adv. Protein Chem.* **40**, 1–61 (1990).
- Hecht, M. H., Nelson, H. C. & Sauer, R. T. Mutations in lambda repressor's amino-terminal domain: implications for protein stability and DNA binding. *Proc. Natl Acad. Sci. USA* **80**, 2676–2680 (1983).
- Sepúlveda, L., Xu, H., Zhang, J. & Wang, M. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science* **351**, 1218–1222 (2016).
- Golding, I. Decision making in living cells: lessons from a simple system. *Annu. Rev. Biophys.* **40**, 63–80 (2011).
- Ptashne, M. et al. How the lambda repressor and cro work. *Cell* **19**, 1–11 (1980).
- Meyer, B. J. & Ptashne, M. Gene regulation at the right operator (OR) of bacteriophage λ . III. λ Repressor directly activates gene transcription. *J. Mol. Biol.* **139**, 195–205 (1980).
- Ackers, G. K., Johnson, A. D. & Shea, M. A. Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl Acad. Sci. USA* **79**, 1129–1133 (1982).
- Shea, M. A. & Ackers, G. K. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J. Mol. Biol.* **181**, 211–230 (1985).
- Li, X., Lalic, J., Baeza-Centurion, P., Dhar, R. & Lehner, B. Changes in gene expression predictably shift and switch genetic interactions. *Nat. Commun.* **10**, 3886 (2019).
- Lagator, M., Paixao, T., Barton, N., Bollback, J. P. & Guet, C. C. On the mechanistic nature of epistasis in a canonical *cis*-regulatory element. *Elife* **6**, e25192 (2017).
- Bray, D. Protein molecules as computational elements in living cells. *Nature* **376**, 307–312 (1995).
- Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
- Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.* **44**, 575–588 (2019).
- Casadio, R., Vassura, M., Tiwari, S., Fariselli, P. & Luigi Martelli, P. Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum. Mutat.* **32**, 1161–1170 (2011).
- Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
- Gimble, F. S. & Sauer, R. T. λ Repressor mutants that are better substrates for RecA-mediated cleavage. *J. Mol. Biol.* **206**, 29–39 (1989).
- Nelson, H. C. & Sauer, R. T. Lambda repressor mutations that increase the affinity and specificity of operator binding. *Cell* **42**, 549–558 (1985).
- Nelson, H. C. M., Hecht, M. H. & Sauer, R. T. Mutations defining the operator-binding sites of bacteriophage repressor. *Cold Spring Harb. Symp. Quant. Biol.* **47**, 441–449 (1983).
- Stayrook, S., Jaru-Ampornpan, P., Ni, J., Hochschild, A. & Lewis, M. Crystal structure of the λ repressor and a model for pairwise cooperative operator binding. *Nature* **452**, 1022–1025 (2008).
- Beckett, D. et al. Isolation of λ repressor mutants with defects in cooperative operator binding. *Biochemistry* **32**, 9073–9079 (1993).
- Nelson, H. C. M. & Sauer, R. T. Interaction of mutant λ repressors with operator and non-operator DNA. *J. Mol. Biol.* **192**, 27–38 (1986).
- Hecht, M. H., Sturtevant, J. M. & Sauer, R. T. Effect of single amino acid replacements on the thermal stability of the NH2-terminal domain of phage lambda repressor. *Proc. Natl Acad. Sci. USA* **81**, 5685–5689 (1984).
- Hecht, M. H., Hehir, K. M., Nelson, H. C. M., Sturtevant, J. M. & Sauer, R. T. Increasing and decreasing protein stability: Effects of revertant substitutions on the thermal denaturation of phage λ repressor. *J. Cell. Biochem.* **29**, 217–224 (1985).
- Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nat. Rev. Genet.* **11**, 572–582 (2010).
- Otwinowski, J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).
- Wodak, S. J. et al. Allosteric in its many disguises: from theory to applications. *Structure* **27**, 566–578 (2019).
- Horowitz, A., Fleisher, R. C. & Mondal, T. Double-mutant cycles: new directions and applications. *Curr. Opin. Struct. Biol.* **58**, 10–17 (2019).
- Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* **369**, 1318–1332 (2007).
- Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis. *Proc. Natl Acad. Sci. USA* **115**, E7550–E7558 (2018).
- Gjuvsland, A. B., Wang, Y., Plahte, E. & Omholt, S. W. Monotonicity is a key feature of genotype-phenotype maps. *Front. Genet.* **4**, 216 (2013).
- Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).

42. Matreyek, K. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
43. Woodsmith, J. et al. Protein interaction perturbation profiling at amino-acid resolution. *Nat. Methods* **14**, 1213–1221 (2017).
44. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *Elife* **7**, e32472 (2018).
45. Keren, L. et al. Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell* **166**, 1282–1294 (2016).
46. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).
47. Mighell, T. L., Evans-Dutson, S. & O’Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
48. Chure, G. et al. Predictive shifts in free energy couple mutations to their phenotypic consequences. *Proc. Natl Acad. Sci. USA* **116**, 18275–18284 (2019).
49. Huang, G. S. & Oas, T. G. Structure and stability of monomeric λ repressor: NMR evidence for two-state folding. *Biochemistry* **34**, 3884–3892 (1995).
50. Reichardt, L. & Kaiser, A. D. Control of lambda repressor synthesis. *Proc. Natl Acad. Sci. USA* **68**, 2185–2189 (1971).
51. Maurer, R., Meyer, B. J. & Ptashne, M. Gene regulation at the right operator (OR) of bacteriophage λ . I. OR3 and autogenous negative control by repressor. *J. Mol. Biol.* **139**, 147–161 (1980).
52. Brent, R. P. in *Algorithms for Minimization Without Derivatives* 61–80, <https://doi.org/10.1109/TAC.1974.1100629> (1973).
53. Soetaert, K. & Herman, P. M. J. *A Practical Guide to Ecological Modelling: Using R as a Simulation Platform* (Springer, 2008).

Acknowledgements

This work was supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2017-89488-P and SEV-2012-0208), the Bettencourt Schueller Foundation, Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), and the CERCA Program/Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership and the Centro de Excelencia Severo Ochoa.

Author contributions

X.L. performed all analyses and made the figures; X.L. and B.L. conceived the study, designed the analyses and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-18694-0>.

Correspondence and requests for materials should be addressed to B.L.

Peer review information *Nature Communications* thanks Elena Kuzmin and other, anonymous, reviewers for their contributions to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020