PHILOSOPHICAL
TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Fisher KJ, Kryazhimskiy S, Lang GI. 2019 Detecting genetic interactions using parallel evolution in experimental populations. *Phil. Trans. R. Soc. B* **374**: 20180237.
http://dx.doi.org/10.1098/rstb.2018.0237

Accepted: 31 January 2019

One contribution of 16 to a theme issue 'Convergent evolution in the genomics era: new insights and directions'.

Subject Areas:
bioinformatics, evolution, genomics

Keywords:
genetic interactions, parallel evolution, experimental evolution, mutual information

Authors for correspondence:
Sergey Kryazhimskiy
e-mail: skryazhi@ucsd.edu
Gregory I. Lang
e-mail: glang@lehigh.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4466069.

THE ROYAL SOCIETY
PUBLISHING

Detecting genetic interactions using parallel evolution in experimental populations

Kaitlin J. Fisher[1], Sergey Kryazhimskiy[2] and Gregory I. Lang[1]

[1]Department of Biological Sciences, Lehigh University, Bethlehem, PA 18015, USA
[2]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA

SK, 0000-0001-9128-8705; GIL, 0000-0002-7931-0428

Eukaryotic genomes contain thousands of genes organized into complex and interconnected genetic interaction networks. Most of our understanding of how genetic variation affects these networks comes from quantitative-trait loci mapping and from the systematic analysis of double-deletion (or knockdown) mutants, primarily in the yeast *Saccharomyces cerevisiae*. Evolve and re-sequence experiments are an alternative approach for identifying novel functional variants and genetic interactions, particularly between non-loss-of-function mutations. These experiments leverage natural selection to obtain genotypes with functionally important variants and positive genetic interactions. However, no systematic methods for detecting genetic interactions in these data are yet available. Here, we introduce a computational method based on the idea that variants in genes that interact will co-occur in evolved genotypes more often than expected by chance. We apply this method to a previously published yeast experimental evolution dataset. We find that genetic targets of selection are distributed non-uniformly among evolved genotypes, indicating that genetic interactions had a significant effect on evolutionary trajectories. We identify individual gene pairs with a statistically significant genetic interaction score. The strongest interaction is between genes *TRK1* and *PHO84*, genes that have not been reported to interact in previous systematic studies. Our work demonstrates that leveraging parallelism in experimental evolution is useful for identifying genetic interactions that have escaped detection by other methods.

This article is part of the theme issue 'Convergent evolution in the genomics era: new insights and directions'.

1. Introduction

Determining the extent to which genetic variants interact to affect phenotypes is a central challenge in biology. Traditional methods such as quantitative-trait loci (QTL) mapping and double-deletion analysis have proven useful for identifying functional variants and genetic interactions in laboratory model systems such as the yeast *Saccharomyces cerevisiae*. However, both of these methods have limitations. QTL mapping provides a robust approach to identifying natural genetic variants that contribute to complex traits, but most studies are underpowered to detect genetic interactions. Large studies (with in the order of 10^3 segregants) have shown that QTL–QTL interactions contribute to a wide array of complex traits [1–4], but even the largest study to date did not have the statistical power to identify small-effect interactions [2]. In addition, genetic linkage makes it difficult in many cases to identify the causal variants underlying most QTLs.

Systematic phenotypic screens of double deletions/knockdowns in yeast and other organisms avoid these problems [5–8]. These types of studies have successfully identified a large number of genetic interactions, particularly within protein complexes [9]. By design, this approach is limited to detecting only strong pairwise interactions between loss-of-function variants. Most natural variation, however, is not loss of function [10,11], and thus, a comprehensive picture of genetic interactions will require tests of interactions between functional variants.

An alternative approach to identifying functionally important variants and interactions between them is to leverage the power of natural selection. When different populations of the same or different species face the same environmental challenge, natural selection often finds the same phenotypic [12–14] or even genetic [15–17] solution to this challenge. This phenomenon is referred to as convergent or parallel evolution. Thus, the observation of parallel genetic changes in multiple independent lineages can be used to identify variants that contribute to functionally important traits [18–20]. This approach has been successful in identifying key mutations in pathogen and tumour evolution [21–24]. The idea of convergence or parallelism has also been used to detect epistasis within genes [25–29] and more recently also between genes [30] in natural populations. In this type of analysis, pairs of variants are identified as genetically interacting if they co-occur in the same genotype more often than expected by chance. There are three challenges in using parallelism to detect functional variants and genetic interactions in natural populations. First, true functional parallelism is confounded by common ancestry. Second, because we rarely know what selection pressures drove the evolution of the functional variants, it is difficult to connect genotype with phenotype. Third, detecting epistasis requires many variants to accumulate and is therefore only feasible in either fast evolving populations or over very long time-scales.

Evolve and re-sequence experiments offer a complementary approach for detecting functional variants and genetic interactions. Like inferences from natural populations, this approach also relies on selection to find functional variants and genetic interactions between them. This approach, however, overcomes problems arising in studies of naturally evolving populations. Hundreds of replicate microbial populations can be propagated in identical conditions such that the selected phenotypes are either known or can be measured [31]. After hundreds or thousands of generations, entire populations or individual isolated clones are sequenced, and adaptive variants are identified by their parallel occurrence in replicate lines (e.g. [32–36]). Since replicate populations evolve independently, overabundance of parallel variants is a signal of positive selection, which is not confounded by common ancestry. Genetic interactions are known to contribute to adaptive evolution [37], and the data from evolve and re-sequence experiments must contain information about these genetic interactions. To the best of our knowledge, only one study so far has leveraged this type of data to detect epistasis and demonstrate how it affected evolutionary trajectories [36]. The challenge is that large datasets are required to detect overrepresented pairs of genes that contain interacting variants. However, unlike in QTL mapping approaches, the number of variants in experimentally evolved populations can be controlled to increase statistical power to reveal genetic interactions. At the same time, evolution in the laboratory, just like evolution in nature, assesses all types of variants, which in principle allows us to detect genetic interactions that may not be revealed in gene-deletion studies.

Here, we present an approach that leverages parallelism in experimental evolution to detect genetic interactions between genes that acquire mutations independently across populations. We detect genetic interactions between pairs of genes using mutual information [38–40]. This quantity captures the statistical dependence between the occurrences of mutations at two specific loci in the same genotype. We use this approach to analyse a recently published whole-genome dataset derived from experimentally evolved asexual populations of yeast. We find that the accumulated mutations are distributed between genotypes non-uniformly, indicating that genetic interactions have contributed to adaptive evolution in these laboratory populations. We identify specific pairs of genes that have acquired mutations in parallel more often than expected by chance, indicating putative genetic interactions. We experimentally verify that our top-hit pair, *TRK1* and *PHO84*, shows a positive genetic interaction when reconstructed in the ancestral background.

2. Material and methods

(a) Sequencing data re-analysis

Evolved mutations used for this analysis were obtained from 92 endpoint clones isolated from 42 populations of 4000 generation evolved autodiploids, previously reported in Fisher *et al.* [33]. Populations were grown in rich media in individual wells of unshaken 96-well plates at 30°C and diluted 1 : 1024 every 24 h. At approximately 60 generation intervals, populations were cryoarchived in 15% glycerol. We reanalysed the raw sequencing data to improve annotation quality. All raw data files were demultiplexed using a custom python script (barcodesplitter.py) from L. Parsons (Princeton University). Adapter sequences were trimmed using fastx_clipper (FASTX Toolkit). Reads were then aligned to a customized W303 genome using BWA v. 0.7.12 [41]. VCFtools was used to filter variants common to all samples and mating-type-specific polymorphisms (see [33]). Remaining polymorphisms were then annotated using a strain-background customized annotation file [42].

(b) Calculating mutual information

We used the evolved mutations generated by reprocessed sequence data to look for evidence of genetic interactions. To prevent false positives due to common ancestry, only one clone with the most mutations from each population was included in the analysis. We then excluded all intergenic and synonymous mutations. Lastly, to reduce the number of statistical tests, we looked for genetic interactions only among 'multi-hit' genes, i.e. those in which at least three mutations in independent populations were detected in the dataset. This was done to reduce noise by enriching for beneficial mutations. Nevertheless, we estimate, by simulation controlling for gene length, that 12% of genes receive three or more mutations by chance alone and are likely neutral. This reduced dataset includes 113 'multi-hit' genes from 46 independently evolved genotypes.

For all 6328 pairwise combinations of multi-hit genes, we calculated the mutual information statistic. To do so, we model an evolved genotype with a series of (possibly non-independent) Bernoulli random variables σ_i with $i = 1, 2, \ldots, K$, where $K = 113$, the total number of genes where mutations can possibly occur; σ_i takes value 1 if a mutation occurs (in the data) in gene i and it takes values 0 if it does not occur. We first estimate the marginal probability of a mutation occurring in gene i as

$$P(\sigma_i = 1) = C_{\mathrm{M}} \sum_{g=1}^{N} \tilde{M}_{gi}. \tag{2.1}$$

Here, $\tilde{M}_{gi} = M_{gi} + \varepsilon$ and $M_{gi} = 1$ if mutation in gene i is present in genotype g in the data. We regularize our estimates by adding a pseudocount $\varepsilon = 1/M$, where M is equal to the total number of mutations in the dataset [43]. Our results are robust with respect to the choice of ε (see below). The sum is taken over all $N = 46$ genotypes and $C_M = 1/N(1+\varepsilon)$ is the normalization constant. The probability of a mutation *not* occurring in gene i is then $P(\sigma_i = 0) = 1 - P(\sigma_i = 1)$. We also estimate the joint probability distribution $P(\sigma_i,\sigma_j)$ for each gene pair (i,j) as follows:

$$P(\sigma_i = 1,\, \sigma_j = 1) = C_J \sum_{g=1}^{N} \tilde{M}_{gi}\tilde{M}_{gj}, \tag{2.2}$$

$$P(\sigma_i = 0,\, \sigma_j = 1) = C_J \sum_{g=1}^{N} (1 + \varepsilon - \tilde{M}_{gi})\tilde{M}_{gj}, \tag{2.3}$$

$$P(\sigma_i = 1,\, \sigma_j = 0) = C_J \sum_{g=1}^{N} \tilde{M}_{gi}(1 + \varepsilon - \tilde{M}_{gj}) \tag{2.4}$$

$$\text{and} \quad P(\sigma_i = 0,\, \sigma_j = 0) = C_J \sum_{g=1}^{N} (1 + \varepsilon - \tilde{M}_{gi})(1 + \varepsilon - \tilde{M}_{gj}), \tag{2.5}$$

where $C_J = 1/N(1+\varepsilon)^2$. We use these estimates of joint probabilities to estimate the mutual information statistic MI_{ij} between random variables σ_i and σ_j as

$$MI_{ij} = \sum_{x,y\in\{0,1\}} P(\sigma_i = x,\, \sigma_j = y)\log_2 \frac{P(\sigma_i = x,\, \sigma_j = y)}{P(\sigma_i = x)P(\sigma_j = y)}. \tag{2.6}$$

The aggregate mutual information statistic MI_{tot} for the full dataset is then calculated as

$$MI_{\text{tot}} = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} MI_{ij}. \tag{2.7}$$

(c) Generating null datasets

To obtain the null distributions for the individual MI_{ij} statistics and for the aggregate MI_{tot} statistic, we generated 'null' datasets that are structurally identical to our real dataset, but in which the mutations are distributed randomly and independently across genotypes with the same marginal probabilities as in the real data. Specifically, in each 'null' dataset, we generated $N = 46$ genotypes by randomly and independently drawing each value M_{gi}, $g = 1, \ldots, N$, $i = 1, \ldots, K$ from the Bernoulli distribution with estimated marginal success probability $P(\sigma_i = 1)$ for each gene i. This method preserves the average numbers of mutations per gene and per clone.

To obtain the null distributions for each MI_{ij} and MI_{tot}, we generated 100 000 'null' datasets, and calculated all MI_{ij} and MI_{tot} statistics for each 'null' dataset as described above. We then estimated the p-value for all MI_{ij} and MI_{tot} and obtained nominally significant pairs of genes at different significance thresholds. Since the MI_{ij} statistics are not independent, we estimated the false discovery rate (FDR) and the p-values for the observed number of nominally significant pairs from our 'null' datasets [27].

(d) Strain construction

Evolved alleles of the most significant gene pair, *PHO84* and *TRK1*, were reconstructed into the ancestral background using CRISPR–Cas9 allele swaps. We first constructed plasmids starting from pML104 (Addgene 67638), which constitutively expresses Cas9 and a guide RNA (gRNA). We designed gRNAs to target one site in *PHO84* (5′-CCCGTAGAAAGCAACATCTAA-3′) and two sites in *TRK1* (5′-TTTTGGGTTCAAATCATCGAA-3′ and 5′-GGAGAACAACTCCTACTCGAC-3′). Plasmids were transformed into our ancestral background (yGIL1298: *MAT*α, *ade2-1*, *CAN1*, *his3-11*, *leu2-3*, *112*, *trp1-1*, *URA3*, *bar1*Δ::*ADE2*, *hml*$\alpha\Delta$::*LEU2*, *GPA1*::KanMX, *ura3*Δ::PFUS1-yEVenus) along with a 500 bp linear repair template (gBlock, IDT) encoding the appropriate evolved allele (*pho84*-A1071C, *trk1*-A733G and *trk1*-C1353G) as well as a synonymous PAM site change. Transformants were genotyped to confirm successful integration of each mutant allele. The *pho84*-A1071C mutant strain was backcrossed to yGIL432 (*MAT***a**, *GPA1*::NatMX, otherwise isogenic to yGIL1298) to move the *pho84*-A1071C allele to the *MAT***a** background. This *pho84*-A1071C *MAT***a** strain was crossed to yGIL1298 to generate heterozygous *pho84*-A1071C mutants, and to each of two *tkr1* mutants to generate heterozygous double mutants. Heterozygous single *trk1* mutants were created by crossing correct transformants to yGIL432. All *MAT***a**/α diploids were then converted to *MAT***a**/**a** to correspond with the autodiploid background in which the mutations arose by transforming diploids with pGIL088, which contains a galactose-inducible *HO* homing endonuclease, as reported in Fisher *et al.* [33].

(e) Fitness assay and interaction analysis

Fitness assays were performed as described previously [33]. Briefly, mutant cultures were mixed 1 : 1 with an autodiploid version of the ancestral strain (yGIL1064) labelled with ymCitrine at *URA3*. Cultures were propagated in a 96-well plate in an identical fashion to the evolution experiment for 50 generations. At 10-generation intervals, saturated cultures were sampled for flow cytometry. Analysis of flow cytometry data was performed with FlowJo 10.3. The selective coefficient was calculated as the slope of the best-fit line of the natural log of the ratio between query and reference strains against time.

Selection coefficients were measured for two technical replicates each of four biological replicates of *pho84*-A1071C and eight biological replicates of the remaining four query genotypes (*trk1*-A733G, *trk1*-C1353G, *pho84*-A1071C/*trk1*-A733G and *pho84*-A1071C/*trk1*-A1353G). One reconstructed clone had an abnormally high fitness, likely due to secondary mutations introduced during transformation, and was removed from the analysis. There was no significant difference in fitness between the two single *trk1* alleles ($t_{28} = 0.95$, $p = 0.35$) or between the two double mutants ($t_{28} = -1.087$, $p = 0.29$), so data for these genotypes were pooled. The expected additive fitness distribution of the double mutant was calculated by adding the mean selection coefficients and propagating the standard deviation of *trk1* and *pho84* single mutants. A one-tailed two-sample t-test was used to test for deviation from additive expectation.

(f) Network and clustering analysis

Hierarchical clustering and heatmap generation were done using the pheatmap R package [44]. Mutual information matrices were clustered by rows and columns using a Euclidean distance matrix. Subclusters shown were identified by trimming row and column dendrograms to five groups and identifying the four subclusters containing less than 20 genes. The significant pair network was generated via the R igraph package [45].

3. Results

(a) Identifying putative genetic interactions

We set out to look for genetic interactions between beneficial mutations that arose in a previously published yeast evolution experiment for which whole-genome sequencing data are publicly available [33]. In this experiment, 46 replicate autodiploid yeast populations evolved in the same laboratory environment for 4000 generations [33]. Using a custom bioinformatics pipeline (Methods), we identified 3835 unique new mutations that arose during evolution. We found 113 'multi-hit' genes, i.e. genes in which a non-synonymous or a nonsense mutation was discovered in at least three independent populations. Since we expect to find only 13.5 of such genes by chance if mutations were distributed randomly across the genome, multi-hit genes must be highly enriched for targets of selection.

We asked whether any pairs of multi-hit genes occurred in our data more or less often than expected by chance. Such over- or underrepresentation would indicate parallel evolution driven by genetic interactions. We calculated the aggregate mutual information statistic, MI_{tot} (Methods), which serves as an overall measure of mutational non-independence in our dataset, and found that $MI_{\text{tot}} = 87.7$ bits. We compared this value to the null distribution generated by randomly and independently distributing mutations among evolved genotypes 10^5 times (see Methods) and found that the observed value was significantly larger than expected by chance ($p < 10^{-3}$; figure 1; electronic supplementary material, figure S1). On average, the knowledge that a mutation in one gene is present in a given genotype provides a very small amount ($87.7/6328 = 0.014$ bits) of information about the presence of a mutation in any other specific gene. Nevertheless, the fact that mutated genes are distributed non-uniformly across evolved genotypes indicates that genetic networks subtly but significantly affected the mutational trajectories in our evolving populations.

Our estimates of mutual information depend on the value of the pseudocount parameter ε (see Methods). We re-ran our analysis (albeit with 10 simulations instead of 10^5) at varying values of ε between 0.0002 ($\varepsilon = 0.1/M$) and 0.004 ($\varepsilon = 2/M$) and found that our main result is robust with respect to the choice of ε (electronic supplementary material, figure S2).

Next, we compared the mutual information statistic MI_{ij} for each gene pair (i,j) in the dataset to its respective null distribution (see Methods; electronic supplementary material, figure S3). We identified a significant genetic interaction between two genes if the p-value for their MI_{ij} was less than 0.003. At this cut-off, we expect to observe 18.8 interacting gene pairs under our null model, but in fact we observe 33 (FDR of 0.57) and this excess is highly significant ($p <$ 0.005; electronic supplementary material, figure S4). Thirty-three significant gene pairs comprise 42 unique genes (electronic supplementary material, table S1).

Interactions between functional variants might be expected to exhibit allele specificity. We examined the identity of independently derived mutations in the top five most significant putative interactions (electronic supplementary material, table S2). Three of the nine genes in the top five pairs showed evidence of repeated loss of function as indicated by PROVEAN score (*IRA2*, *LTE1*, *WHI2*) [46]. Mutations in the remaining six genes show a mix of predicted effects. We examined the positions of mutations within each gene to look for patterns of site-specific variation. We found that the distribution of mutations across coding sequences was consistent with the uniform null hypothesis.

(b) Experimental verification of genetic interaction between mutations in *PHO84* and *TRK1*

Despite the high FDR, our epistasis analysis suggests that mutations in the top significant pair of genes, *PHO84* and *TRK1* (nominal $p < 10^{-5}$), exhibit a true genetic interaction. Mutations in these two genes co-occurred in the same genotype in our data three times and never exhibited a higher value of mutual information in any of the 100 000 simulations. When examining the complete dataset, including all clones descending from each population, we found that a mutation in the *PHO84* gene precedes a mutation in *TRK1* in at least one population and that all populations with a non-synonymous mutation in *PHO84* allele acquire a *TRK1* mutation (electronic supplementary material, figure S5).

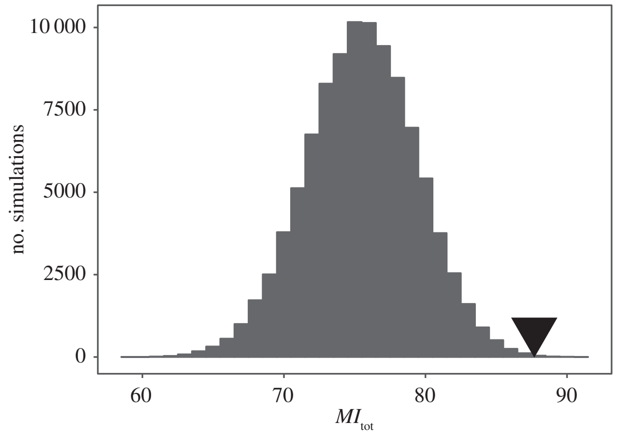

Figure 1. Histogram showing the null distribution of the aggregated MI_{tot} statistic based on 100 000 simulations (see Methods). Observed MI_{tot} is indicated by the black triangle.

To experimentally validate this positive genetic interaction, we reconstructed one allele of *pho84* and two alleles of *trk1* in the ancestral background both as single mutants and as double *pho84*/*trk1* mutants. All mutations were constructed as heterozygotes—the state in which they are maintained in the evolution experiment—and assayed for fitness. The mutant *pho84* and *trk1* alleles conferred small but measurable fitness benefits (0.009 ± 0.001 s.e. for *pho84* and 0.003 ± 0.001 s.e. for *trk1*). We found that the fitness of the *pho84*/*trk1* double mutant (0.015 ± 0.001 s.e.) was higher than the expectation based on the sum of fitnesses of single mutants (0.013 ± 0.001 s.e., figure 2), although the difference was only marginally significant ($t_{58} = 1.74$, $p = 0.043$).

(c) Structure of genetic interaction networks

We found that the set of putatively interacting genes is highly interconnected. The most significant 33 gene pairs consist of eight modules that contain at least three genes each and five isolated gene–gene interaction pairs (figure 3). The three largest modules encompass 42% of all candidate significant interactions. We performed hierarchical clustering on MI_{ij} and found that this matrix contains multiple small but tightly connected blocks (figure 4). On average, mutual information between any two genes within a block was eight times higher than between a random pair of genes (0.098 bits versus 0.012 bits). Notably, these blocks largely overlapped with the modules observed among putatively interacting pairs. This suggests that genetic interactions, rather than being exclusively strong pairwise interactions, are often dispersed among small networks of interacting genes.

4. Discussion

Even the simplest free-living microorganisms encode thousands of genes organized into complex and interconnected

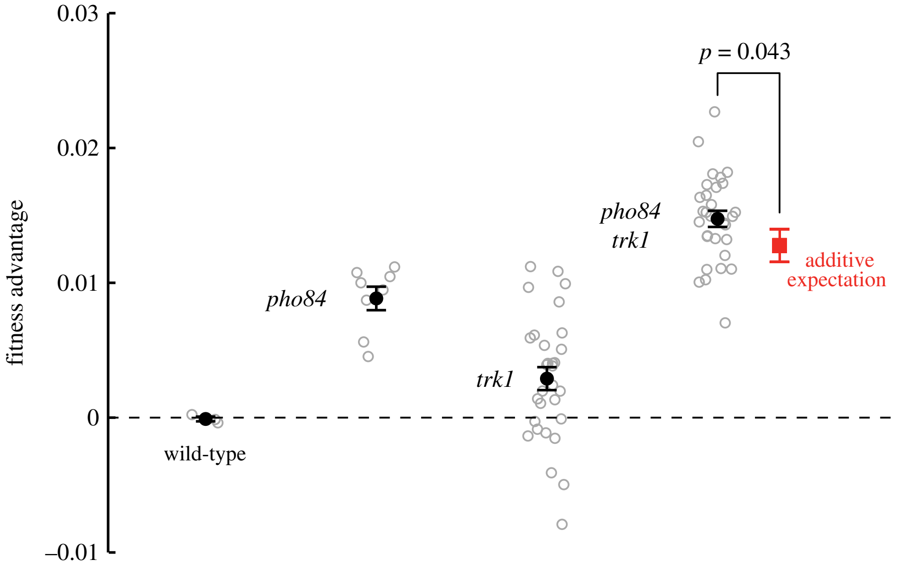

Figure 2. Fitness advantage of the single *TRK1* and *PHO84* mutations and of the double mutant. Replicate measurements are plotted as grey circles. Mean estimates are plotted as bold circles ± standard error. The red square indicates the additive expectation for the double mutant.

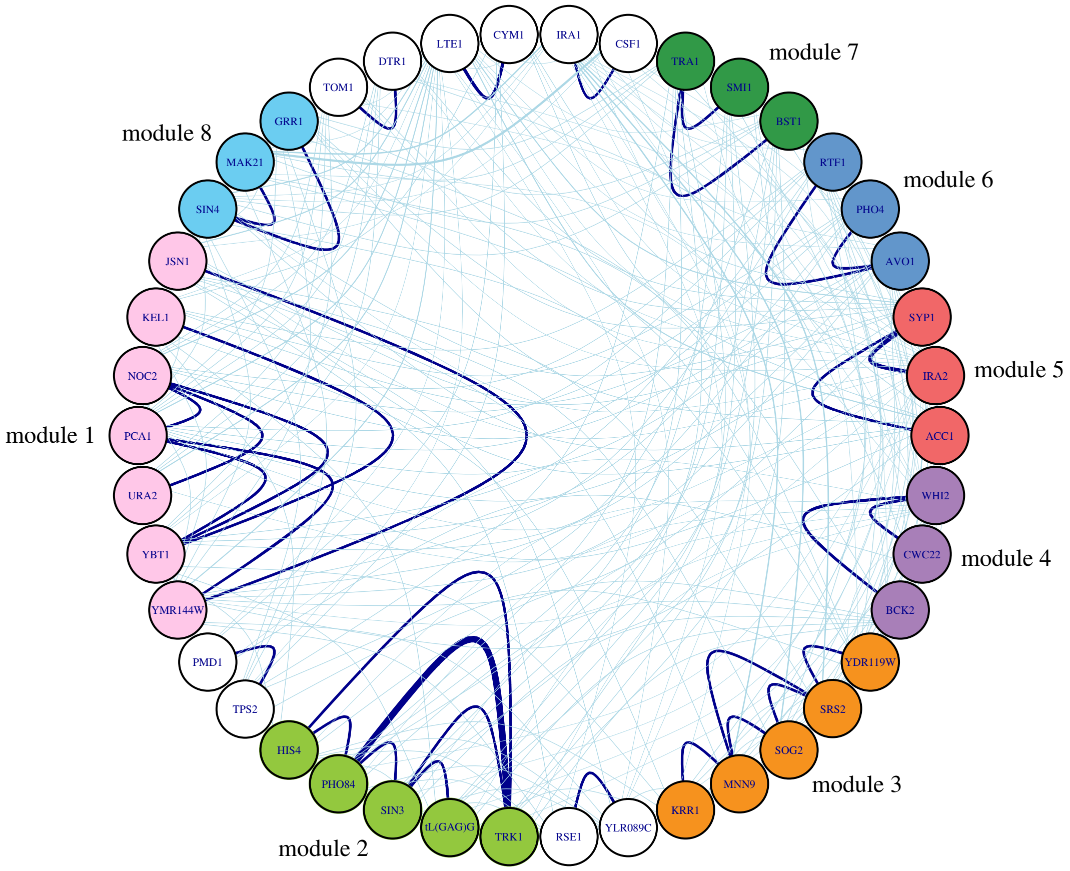

Figure 3. Network of all genes identified in significant gene pairs. Edges are scaled by MI_{ij} and connect all genes that co-occur in the same background at least once. Bolded lines represent significant pairwise MI_{ij}. Colours correspond to interconnected significant pairs. White circles indicate isolated gene pairs. Modules are labelled by size.

networks that collectively determine the organism's fitness. These genetic interaction networks constrain evolution such that populations evolving in identical conditions often find similar genetic solutions, both in nature and in the laboratory (e.g. [14,36,47]). Here, we developed a method, based on mutual information, that exploits genetic parallelism observed in microbial evolution experiments to infer genetic interactions between loci that acquired mutations in independent populations. With this method, we found that genetic interactions had an overall significant effect on mutational trajectories of evolved populations. We also identified 33 gene pairs (at FDR of 0.57) that exhibit the strongest genetic interactions in our dataset. We provide experimental support for one of these interactions, between genes *PHO84* and *TRK1*.

Our method for detecting genetic interactions complements existing approaches. Most of our understanding of

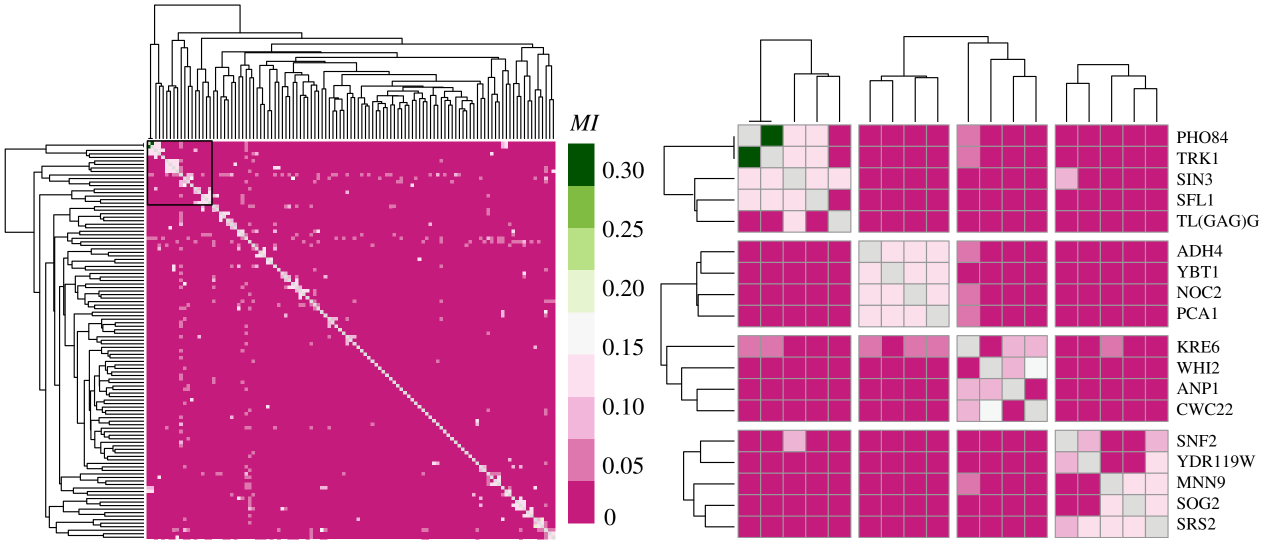

Figure 4. Hierarchical clustering of genes by pairwise mutual information captures the most significant pairs and networks among significant pairs. Subclusters shown were identified by trimming row and column dendrograms to five groups and identifying the four subclusters containing less than 20 genes.

genetic interactions comes from the systematic analysis of double-deletion/knockdown mutations [5,6,8,9,47–52]. By design, these approaches query only loss-of-function mutations, which represent less than 5% of natural variation in both yeast and humans [10,11]. By contrast, our approach can detect pairwise epistasis between all classes of beneficial variants, including gain-of-function mutations, mutations in essential genes and regulatory mutations that would be missed in gene-deletion studies. Indeed, out of the 33 most significant gene pairs, only one genetic interaction (between *SIN3* and *TRK1*) was known previously [6]. The most significant interaction discovered here is between genes *TRK1* and PHO84. *TRK1* encodes a high-affinity potassium transporter and *PHO84* encodes a high-affinity phosphate transporter. The biological cause of their interaction is unclear, although there is evidence of crosstalk between potassium and phosphate import [53,54].

While gene-deletion studies are particularly good at detecting strong negative pairwise interactions between deleterious mutations, such as synthetic lethality (reviewed in [9]), our method identifies primarily positive interactions between pairs of selectively accessible mutations. In theory, our approach could also capture negative interactions, but this would require observing an absence of certain mutational combinations more often than expected by chance. Such mutational incompatibilities have been observed in evolution experiments [36,47]; for example, mutations in an *HXT6/7* hexose transporter and its negative regulator, *MTH1*, in glucose-limited yeast chemostat populations are incompatible [47]. The absence of negative interactions in our list of significant pairs suggests that we are underpowered to detect them.

Several recent experimental evolution studies have found that adaptive mutations often exhibit a global (i.e. not specific to a particular gene pair) type of negative epistasis, which is referred to as 'diminishing returns epistasis' [55–58]. For example, we previously demonstrated that beneficial alleles in *gas1* and *ste12*, both approximately 3% fitness effect mutations, yield only a net approximately 5% benefit when combined [35]. If diminishing returns epistasis is indeed widespread, then pairwise interactions between specific mutations should be detected as deviations of the double-mutant fitness from the appropriate diminishing returns null model rather than from a naive additive model. Then, observing a double-mutant with higher-than-additive fitness, such as the *TRK1/PHO84* double mutant (figure 3), would be even more surprising compared to the diminishing returns null than to the additive null, and would provide even stronger evidence for a gene-specific positive genetic interaction.

Our approach has several important limitations. It suffers from a high rate of false discoveries (about 60%), at least for the dataset that we have analysed here. There are at least two reasons for such high FDR. First, we looked for signatures of epistasis among pairs of genes in which we observed three or more independent mutations. We assumed that all observed mutations in these 'multi-hit' genes are beneficial. However, this may not be the case. We estimate around 12% of the genes included in this analysis to have been mutated three or more times simply by chance. These mutations are distributed uniformly among genotypes and therefore decrease the signal-to-noise ratio in our data. One way to decrease FDR is to consider genes with an even higher degree of parallelism. Of course, this would come at a cost of potentially missing interesting genetic interactions among less frequently mutated genes.

Second, high FDR may in fact reflect a real biological phenomenon. Gene-deletion studies have shown that strong pairwise epistasis is relatively rare, around 4% if both positive and negative interactions are counted [6]. Thus, strong pairwise genetic interactions among beneficial mutations might also be rare. Weak epistasis might be more common, but it is also harder to detect. The highly significant value of the aggregate MI_{tot} statistic in our study suggests that genetic interactions jointly have affected the outcome of the evolutionary process at the genetic level. At the same time, the difficulty of reliably identifying individual interacting gene pairs suggests that genetic interactions, rather than being strong and concentrated in a small number of gene pairs, are weak and relatively dispersed. The power of our approach to detect weaker genetic interactions could be improved with more replicate populations. In our null model, co-occurrence of two mutations in the same genotype happens with probability in the order N^{-1}, where N is the number of independently evolved genotypes. For example, the p-value for two genes with three mutations each where all mutations co-occur in the same three genotypes scales is N^{-3}.

As mentioned above, our method is designed to detect pairwise genetic interactions. However, we observe that putative genetic interactions that we identify are clustered in groups that contain two to seven genes. It is tempting to conclude that such clustering is caused by real biological modules corresponding to physiologically distinct routes of adaptation. However, some degree of clustering is expected even if all of epistasis were pairwise and uniformly distributed among genes. The amount of such spurious clustering would depend on the strength and prevalence of epistasis and is hard to estimate. Increasing the number of replicate populations and reducing the duration of evolution experiments are likely to alleviate this problem.

Our approach does not eliminate the need for experimental validation of putative genetic interactions. However, current molecular techniques make genetic reconstructions feasible only for a relatively small number of mutations. Thus, our approach could serve as an initial filter for narrowing down the set of potentially interesting pairs of mutations for further experimental validation and investigation.

Our results demonstrate the feasibility of using experimental evolution and genetic parallelism to identify biologically interesting genetic interactions that might otherwise be difficult to uncover. In combination with other approaches, it will facilitate characterization of epistasis and, more broadly, help us understand the factors driving patterns of parallelism, diversification and genomic constraint in evolution.

Data accessibility. Electronic supplementary material, dataset 1 contains the reduced dataset used for pairwise epistasis detection. Mutual information and simulation code is archived on GitHub as a release: https://github.com/LangYeastEvoLab/Detect_Parallel_Epistasis; doi:10.5281/zenodo.1938607.

Authors' contributions. K.J.F. and G.I.L. conceived of the study. K.J.F., S.K. and G.I.L. conceptualized mutual information analysis and simulations. K.J.F. wrote the code. K.J.F., S.K. and G.I.L. analysed the data. K.J.F., S.K. and G.I.L. wrote the paper.

Competing interests. We have no competing interests.

Funding. This research was supported by grants to G.I.L. from the NIH (R01GM127420) and to S.K. from the BWF Career Award at Scientific Interface, the Alfred P. Sloan Foundation and the Hellman Foundation.

Acknowledgements. We thank D. Aggeli and S. Buskirk and members of the Lang Lab for comments on the manuscript.

References

1. Bloom JS, Ehrenreich IM, Loo WT, Lite TL, Kruglyak L. 2013 Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234–237. (doi:10.1038/nature11867)
2. Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. 2015 Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nat. Commun.* **6**, 8712. (doi:10.1038/ncomms9712)
3. Huang W *et al.* 2012 Epistasis dominates the genetic architecture of *Drosophila* quantitative traits. *Proc. Natl Acad. Sci. USA* **109**, 15 553–15 559. (doi:10.1073/pnas.1213423109)
4. Wilkening S *et al.* 2014 An evaluation of high-throughput approaches to QTL mapping in *Saccharomyces cerevisiae*. *Genetics* **196**, 853–865. (doi:10.1534/genetics.113.160291)
5. Babu M *et al.* 2014 Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in *Escherichia coli*. *PLoS Genet.* **10**, e1004120. (doi:10.1371/journal.pgen.1004120)
6. Costanzo M, *et al.* 2016 A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420. (doi:10.1126/science.aaf1420)
7. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG. 2006 Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet.* **38**, 896. (doi:10.1038/ng1844)
8. Tong AH *et al.* 2004 Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813. (doi:10.1126/science.1091317)
9. Baryshnikova A *et al.* 2010 Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* **7**, 1017–1024. (doi:10.1038/nmeth.1534)
10. Bergstrom A *et al.* 2014 A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31**, 872–888. (doi:10.1093/molbev/msu037)
11. Saleheen D *et al.* 2017 Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239. (doi:10.1038/nature22034)
12. Hagen DW, Gilbertson LG. 1972 Geographic variation and environmental selection in *Gasterosteus aculeatus* L. in the Pacific northwest, America. *Evolution* **26**, 32–51. (doi:10.1111/j.1558-5646.1972.tb00172.x)
13. O'Quin KE, Hofmann CM, Hofmann HA, Carleton KL. 2010 Parallel evolution of opsin gene expression in African cichlid fishes. *Mol. Biol. Evol.* **27**, 2839–2854. (doi:10.1093/molbev/msq171)
14. Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ. 2006 Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat. Genet.* **38**, 107–111. (doi:10.1038/ng1700)
15. Glazer AM, Cleves PA, Erickson PA, Lam AY, Miller CT. 2014 Parallel developmental genetic features underlie stickleback gill raker evolution. *EvoDevo* **5**, 19. (doi:10.1186/2041-9139-5-19)
16. McCracken K *et al.* 2009 Parallel evolution in the major haemoglobin genes of eight species of Andean waterfowl. *Mol. Ecol.* **18**, 3992–4005. (doi:10.1111/j.1365-294X.2009.04352.x)
17. Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012 Parallel molecular evolution in an herbivore community. *Science* **337**, 1634–1637. (doi:10.1126/science.1226630)
18. Rey C, Lanore V, Veber P, Guéguen L, Lartillot N, Sémon M, Boussau B. 2019 Detecting adaptive convergent amino acid evolution. *Phil. Trans. R. Soc. B* **374**, 20180234. (doi:10.1098/rstb.2018.234)
19. Witt KE, Huerta-Sánchez E. 2019 Convergent evolution in human and domesticate adaptation to high-altitude environments. *Phil. Trans. R. Soc. B* **374**, 20180235. (doi:10.1098/rstb.2018.0235)
20. Yang L, Ravikanthachari N, Mariño-Pérez R, Deshmukh R, Wu M, Rosenstein A, Kunte K, Song H, Andolfatto P. 2019 Predictability in the evolution of Orthopteran cardenolide insensitivity. *Phil. Trans. R. Soc. B* **374**, 20180246. (doi:10.1098/rstb.2018.0246)
21. Carroll MW *et al.* 2015 Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature* **524**, 97. (doi:10.1038/nature14594)
22. Gerlinger M *et al.* 2014 Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225. (doi:10.1038/ng.2891)
23. Kryazhimskiy S, Bazykin GA, Plotkin J, Dushoff J. 2008 Directionality in the evolution of influenza A haemagglutinin. *Proc. R. Soc. B* **275**, 2455–2464. (doi:10.1098/rspb.2008.0521)
24. Lieberman TD *et al.* 2011 Parallel bacterial evolution within multiple patients identifies candidate

pathogenicity genes. *Nat. Genet.* **43**, 1275–1280. (doi:10.1038/ng.997)

25. Codoñer FM, Fares MA. 2008 Why should we care about molecular coevolution? *Evol. Bioinform.* **4**, 117693430800400003. (doi:10.1177/117693430800400003)
26. Korber B, Farber RM, Wolpert DH, Lapedes AS. 1993 Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl Acad. Sci. USA* **90**, 7176–7180. (doi:10.1073/pnas.90.15.7176)
27. Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. 2011 Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* **7**, e1001301. (doi:10.1371/journal.pgen.1001301)
28. Lockless SW, Ranganathan R. 1999 Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299. (doi:10.1126/science.286.5438.295)
29. Shapiro B, Rambaut A, Pybus OG, Holmes EC. 2006 A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. *Mol. Biol. Evol.* **23**, 1724–1730. (doi:10.1093/molbev/msl037)
30. Neverov AD, Kryazhimskiy S, Plotkin JB, Bazykin GA. 2015 Coordinated evolution of influenza A surface proteins. *PLoS Genet.* **11**, e1005404. (doi:10.1371/journal.pgen.1005404)
31. Long A, Liti G, Luptak A, Tenaillon O. 2015 Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat. Rev. Genet.* **16**, 567–582. (doi:10.1038/nrg3937)
32. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009 Genome evolution and adaptation in a long-term experiment with *Escherichia coli. Nature* **461**, 1243–1247. (doi:10.1038/nature08480)
33. Fisher KJ, Buskirk SW, Vignogna RC, Marad DA, Lang GI. 2018 Adaptive genome duplication affects patterns of molecular evolution in *Saccharomyces cerevisiae*. *PLoS Genet.* **14**, e1007396. (doi:10.1371/journal.pgen.1007396)
34. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017 The dynamics of molecular evolution over 60 000 generations. *Nature* **551**, 45–50. (doi:10.1038/nature24287)
35. Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. 2013 Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**, 571–574. (doi:10.1038/nature12344)
36. Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012 The molecular diversity of adaptive convergence. *Science* **335**, 457–461. (doi:10.1126/science.1212986)
37. Phillips PC. 2008 Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867. (doi:10.1038/nrg2452)
38. Bindewald E, Shapiro BA. 2006 RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* **12**, 342–352. (doi:10.1261/rna.2164906)
39. Gloor GB, Martin LC, Wahl LM, Dunn SD. 2005 Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **44**, 7156–7165. (doi:10.1021/bi050293e)
40. Kim Y, Koyutürk M, Topkara U, Grama A, Subramaniam S. 2005 Inferring functional information from domain co-evolution. *Bioinformatics* **22**, 40–49. (doi:10.1093/bioinformatics/bti723)
41. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. (doi:10.1093/bioinformatics/btp324)
42. Matheson K, Parsons L, Gammie A. 2017 Whole-genome sequence and variant analysis of W303, a widely-used strain of *Saccharomyces cerevisiae. G3* **7**, 2219–2226. (doi:10.1534/g3.117.040022)
43. Schürmann T, Grassberger P. 1996 Entropy estimation of symbol sequences. *Chaos* **6**, 414–427. (doi:10.1063/1.166191)
44. Kolde R. 2012 Pheatmap: pretty heatmaps. R package version 61.
45. Csardi G, Nepusz T. 2006 The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695**, 1–9.
46. Choi Y, Chan AP. 2015 PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747. (doi:10.1093/bioinformatics/btv195)
47. Kvitek DJ, Sherlock G. 2011 Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet.* **7**, e1002056. (doi:10.1371/journal.pgen.1002056)
48. Szappanos B *et al.* 2011 An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* **43**, 656–662. (doi:10.1038/ng.846)
49. Van Opijnen T, Bodi KL, Camilli A. 2009 Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6**, 767–772. (doi:10.1038/nmeth.1377)
50. Breslow DK *et al.* 2008 A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat. Methods* **5**, 711–718. (doi:10.1038/nmeth.1234)
51. Costanzo M *et al.* 2010 The genetic landscape of a cell. *Science* **327**, 425–431. (doi:10.1126/science.1180823)
52. Jasnos L, Korona R. 2007 Epistatic buffering of fitness loss in yeast double deletion strains. *Nat. Genet.* **39**, 550–554. (doi:10.1038/ng1986)
53. Canadell D, González A, Casado C, Ariño J. 2015 Functional interactions between potassium and phosphate homeostasis in *Saccharomyces cerevisiae*. *Mol. Microbiol.* **95**, 555–572. (doi:10.1111/mmi.12886)
54. Rosenfeld L, Reddi AR, Leung E, Aranda K, Jensen LT, Culotta VC. 2010 The effect of phosphate accumulation on metal ion homeostasis in *Saccharomyces cerevisiae*. *J. Biol. Inorg. Chem.* **15**, 1051–1062. (doi:10.1007/s00775-010-0664-8)
55. Chou H-H, Chiu H-C, Delaney NF, Segrè D, Marx CJ. 2011 Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* **332**, 1190–1192. (doi:10.1126/science.1203799)
56. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011 Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332**, 1193–1196. (doi:10.1126/science.1203801)
57. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. 2014 Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* **344**, 1519–1522. (doi:10.1126/science.1250939)
58. Rojas Echenique JI, Kryazhimskiy S, Nguyen Ba AN, Desai MM. 2019 Modular epistasis and the compensatory evolution of gene deletion mutants. *PLoS Genet.* **15**, e1007958. (doi:10.1371/journal.pgen.1007958).