March 18, 2020

Decoupling gene knockout effects from gene functions by evolutionary analyses

Li Liu[#], Mengdi Liu[#], Di Zhang, Shanjun Deng, Piaopiao Chen, Jing Yang, Yunhan Xie & Xionglei He

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

[#] These authors contributed equally to this work.

[*] Correspondence should be addressed to X.H. (hexiongl@mail.sysu.edu.cn).

Abstract

Genic functions have long been confounded by pleiotropic mutational effects. To understand such genetic effects, we examine HAP4, a well-studied transcription factor in *Saccharomyces cerevisiae* that functions by forming a tetramer with HAP2, HAP3, and HAP5. Deletion of HAP4 results in highly pleiotropic gene expression responses, some of which are clustered in related cellular processes (clustered effects) while most are distributed randomly across diverse cellular processes (distributed effects). Strikingly, the distributed effects that account for much of HAP4 pleiotropy tend to be non-heritable in a population, suggesting they have little evolutionary consequences. Indeed, these effects are poorly conserved in closely related yeasts. We further show substantial overlaps of clustered effects, but not distributed effects, among the four genes encoding the HAP2/3/4/5 tetramer. This pattern holds for other biochemically characterized yeast protein complexes or metabolic pathways. Examination of a set of cell morphological traits of the deletion lines yields consistent results. Hence, only some gene deletion effects support related biochemical understandings with the rest being pleiotropic and evolutionarily decoupled from the gene's normal functions.

Introduction

Mutation analysis has long been used to understand the functions of a gene(*1*). It appears now clear that a gene can often affect various seemingly unrelated traits(*2*), a phenomenon termed pleiotropy(*3*). For instance, a large-scale gene knockdown assay

in the nematode worm *Caenorhabditis elegans* finds on average a gene affects ~10% of 44 assessed traits(*4*). Attempts to understand such pleiotropic mutational effects are mainly from mechanistic perspectives(*5, 6*), by considering the focal gene's multiple molecular functions or multiple cellular processes associated with a single molecular function(*7*). The resulting pictures are, however, often complex, confusing our understandings in how a gene functions.

Since biological systems are all evolutionary products with history, mechanistic perspectives alone may bias the efforts for delineating a biological phenomenon(*8, 9*). This is exemplified by the debates on the ENCODE project in which up to 80% of the human genome was claimed to be functional despite that only 10% appears to be under selection(*10-12*). A simple example explains how the confusion arose. Suppose there is a transcription factor (TF) that recognizes a DNA motif, say, ATCGATC. The human genome with ~3×10^9 base pairs in length contains over one hundred thousand ATCGATC motif sequences, some of which are evolutionarily selected for certain biological processes while the rest exist as *ad hoc* entities due to the equilibrium of random mutations in such a long genome(*11*). From a purely mechanistic perspective originally adopted by the ENCODE consortium(*10*), the myriad ATCGATC sequences are all called functional so long as they are bound by the TF. However, the claim of *ad hoc* entities as functional would only confuse our understandings in how the system is organized to function. Such confusions forced the ENCODE consortium to eventually abandon their evolution-free view on biochemical functionality(*13*). Notably, the same problem actually also applies to the genetic effects defined in reverse

77 genetics analysis. The common practice is to knockout or knockdown a gene and find
78 the traits significantly altered(*14*), which represents a purely mechanistic framework.
79 In the above TF versus ATCGATC motif example, when the TF is deleted, the
80 expression of those genes with the motif at promoter region could all be affected. The
81 resulting pleiotropic effects, which are either *ad hoc* or evolutionarily selected
82 according to the nature of the focal motifs, together would lead to a very complex
83 picture on the functionality of the TF.

84 The necessity of adopting an evolutionary view in reverse genetic analysis lies also
85 in the effect size of the knockout or knockdown mutations experimentally introduced,
86 which is much larger than that of typical segregating alleles in natural populations(*15*).
87 Hence, while the normal functions of a gene is necessarily built by natural selection,
88 responses to such experimental inactivation of the functions may not be shaped by
89 evolution(*16*). Then, how can the "non-evolutionary" responses be in line with the
90 evolutionarily selected gene functions? With this question in mind we here examine
91 the evolutionary nature of a set of gene deletion effects. We show widespread
92 decoupling of gene deletion effects from the gene's normal functions, calling for an
93 evolutionary framework for reverse genetic analysis.

94

95 **Results**

96 We started with a known yeast *Saccharomyces cerevisiae* gene HAP4(*17*). It is a
97 non-essential transcription factor that has been subjected to extensive studies since its
98 discovery 30 years ago(*18*). We deleted the open reading frame of HAP4 in *S.*

cerevisiae strain BY4741, and checked the expression trait of the other yeast genes by sequencing the transcriptome of the strains that grow in the rich medium YPD at 30 °C (Table S1). We found 195 responsive genes each with a significant expression change under a stringent statistical cutoff (Table S2). Gene ontology (GO) analysis of the 195 genes revealed one third of them (65/195) clustered more than expected by chance in dozens of GO terms. These GO terms are related with each other and reflect well the functional annotations of HAP4 as a regulator of mitochondria activities(*19*) (Fig. 1A). The remaining two thirds (130/195) distribute rather randomly across diverse biological processes, underscoring the strong pleiotropy of HAP4. The two sets of genes are all functionally characterized with clear GO annotations (Table S2), and have comparable fitness importance ($P = 0.1$, Mann-Whitney U-test; Fig. S1). Notably, the 65 clustered deletion effects and 130 distributed deletion effects are supported by similar P-values and fold changes ($P = 0.20$ and 0.46, respectively, Mann-Whitney U-test; Fig. 1B).

Because evolution happens in a population rather than in an individual, it is important to test the population-level heritability of the deletion effects. We crossed two *S. cerevisiae* strains to obtain a population of yeast segregants. Specifically, a wild-type strain BY4742(*MATalpha*), which is identical to BY4741 except at the mating locus, was crossed with the HAP4 deletion line of DBVPG1373(*MATa*) (Fig. 1C). This way, the comparison between wild-type and null alleles at the HAP4 locus would match the comparison conducted in the isogenic BY4741 background. We dissected six tetrads of the hybrid and obtained 12 HAP4 wild-type and 12 HAP4 null segregants. For each of the 195 deletion effects we computed its heritability (h^2_{HAP4})

121 in the segregant population (Methods). The h^2_{HAP4} measures the fraction of variance
122 of an expression trait that is attributed to the HAP4 locus. We noted that the
123 heritability analysis resembles a forward genetic assay with a candidate genetic locus.
124 An h^2_{HAP4} close to zero suggests HAP4 is not a QTL (Quantitative Trait Locus) of the
125 expression trait. We found the 65 clustered effects in general have much higher h^2_{HAP4}
126 than the 130 distributed effects ($P = 7.8\times10^{-4}$, Mann-Whitney U-test; Fig. 1D).
127 Approximately 26.2% (17/65) of the clustered effects have a statistically significant
128 h^2_{HAP4}, while the number is 0.8% (1/130) for the distributed effects ($P = 2.1\times10^{-8}$,
129 Fisher's exact test; Fig. 1E). It is worth pointing out that mutational effects sensitive
130 to genetic backgrounds have been documented in a wide range of organisms(*20-23*).
131 Because non-heritable phenomena in biology last only one generation, with little
132 evolutionary consequences, the deletion effects with low population-level heritability
133 should be evolutionarily unconserved. We obtained, under the same environment and
134 the same statistical cutoff as in BY4741, the deletion effects of HAP4 in *Saccharomyces*
135 *paradoxus* strain N17, a closely related yeast species diverged from *S. cerevisiae* ~10
136 million years ago(*24*) (Fig. 1F). Only 5.4% (7/130) of the distributed effects found in
137 *S. cerevisiae* were also observed in the HAP4 deletion line of *S. paradoxus* N17, while
138 the number was 27.7% (18/65) for the clustered effects ($P = 3.1\times10^{-5}$, Chi-square test).
139 The difference was robust as evidenced by plotting the expression responses of the 195
140 genes in N17(*Δhap4*) (Fig. 1G). Notably, although the statistical signals were not
141 directly comparable between the conservation analysis and heritability analysis, there
142 were 14 overlaps between the 25 conserved effects defined here and the 18 effects with

143 significant h^2_{HAP4}. Hence, the heritability analysis based on a rather arbitrary and
144 small population appeared to represent well the situation in nature. We also examined
145 intra-species conservation of the HAP4 deletion effects by looking at two other *S.*
146 *cerevisiae* strains DBVPG1373 and GIL104, the former of which is 0.35% diverged
147 from the strain BY4741 and the latter 0.07% diverged at the genomic level(*25*).
148 Similar1y, the 130 distributed effects were largely unreproducible in strains
149 DBVPG1373(*Δhap4*) and GIL104(*Δhap4*), while the 65 clustered effects were much
150 more conserved ($P = 0.5\times10^{-4}$ for DBVPG1373(*Δhap4*) and $P = 2.6\times10^{-5}$ for
151 GIL104(*Δhap4*), Chi-square test; Fig. S2). As expected, both types of the deletion
152 effects are more reproducible in more related yeasts (Fig. 1H). We confirmed the
153 results cannot be explained by different detectability of expression changes between the
154 two gene sets by excluding those genes lowly expressed in wild-type BY4741 (Fig. S3).
155 These data, together with the heritability analysis, suggest the clustered effects of HAP4
156 tend to be evolutionarily selected; on the contrary, the distributed genetic effects appear
157 largely non-evolutionary, likely representing *ad hoc* responses to the gene deletion(*16*).
158 HAP4 functions by forming a tetramer with HAP2, HAP3 and HAP5, which is a
159 result of evolution(*18*). We hypothesized clustered effects should support this
160 biochemistry understanding better than distributed effects, because the latter is non-
161 evolutionary. To test the hypothesis, we deleted the other three genes that encode the
162 tetramer, respectively, in *S. cerevisiae* BY4741, and measured the expression profiles
163 of the deletion lines. We defined clustered effects and distributed effects for each of
164 the lines using the same method as in the HAP4 deletion line. We obtained 43, 150

and 50 clustered effects, and 61, 306 and 111 distributed effects for the deletions of HAP2, HAP3 and HAP5, respectively (Fig. 2A-C; Table S3). Consistent with the hypothesis, we found 20 overlapped clustered effects across the four gene deletion lines, 14.5 times higher than that of the distributed effects ($P < 0.001$, simulation test, Fig. 2D). Notably, the 20 overlapped clustered effects are not the strongest in BY4741(*Δhap4*) (Fig. S4). To avoid the potential bias that expression responses to the tetramer may have been considered in the GO annotations of the responsive genes, we excluded all expression-related evidences for GO annotations to define new clustered and distributed effects. We obtained essentially the same result (Fig. S5). Because there are publicly available microarray data for HAP2, HAP3, HAP4 and HAP5 deletion lines(*26*), we also repeated the analysis using the public expression data and observed a similar pattern (Fig. S6).

In addition to considering the protein complex formed by HAP4, we could also consider protein-DNA interactions since HAP4 is a transcription factor. Data from a chromatin immuno-precipitation (CHIP) assay of the promoters bound by HAP4 show that, among the 195 responsive genes observed in BY4741(*Δhap4*), 13 are direct targets of HAP4 (Fig. 2E)(*27, 28*). Interestingly, there is 24-fold enrichment of direct targets in the clustered effects relative to the distributed effects ($P = 8.6\times10^{-6}$, Fisher's exact test); among the 20 overlapped clustered effects 50% (10/20) are direct targets of HAP4, while the genomic background is 0.64% (33/5146) ($P = 4.6\times10^{-18}$, Hypergeometric test) (Fig. 2F). Hence, the CHIP data well support the distinction of the two effect types.

Collectively, these results are consistent with a previous model(*16*) (Box 1), in

187 which the null phenotype of a gene can be ascribed to either the loss of the gene's native
188 functions, or the gain of spurious functions that arise from passive adjustments of the
189 cellular system after the perturbation. The key difference of the two function types is
190 their evolutionary nature: native functions are historical, selected, and evolutionary,
191 while spurious functions are ahistorical, *ad hoc*, and non-evolutionary (*29-31*).
192 Accordingly, the distributed effects examined here likely represent spurious functions
193 created by the HAP4 deletion, and the clustered effects could be in a large part ascribed
194 to the native functions of HAP4. It is intriguing how the two effect types defined by
195 GO could fit the two function types described in the model. We reasoned that
196 evolutionarily optimized native functions are likely to regulate specific pathways or
197 processes; losing them would thus cause coordinating changes of the related genes(*9*),
198 which are detected by GO analysis. In contrast, spurious functions may affect the
199 transcriptome in a rather random way, resulting in distributed changes across diverse
200 cellular processes, most of which cannot be covered by overrepresented GO terms.
201 This may explain why GO clustering here could echo evolutionary effectiveness.

202 Regardless of the underlying logic, clustered genetic effects seem to be well
203 matched with related biochemical understandings. This would help address a long-
204 standing challenge to molecular biology - the gap between genetic analysis and
205 biochemistry analysis(*14, 32*); specifically, genes with intimate biochemical
206 interactions do not have common genetic effects and genes with common genetic
207 effects do not show intimate biochemical interactions(*33, 34*). To test the generality
208 of the finding that was based on the HAP2/3/4/5 tetramer, we examined other

biochemically characterized protein complexes by using publicly available expression data. To avoid bias we considered a single dataset comprising microarray-based expression profiles of over one thousand yeast gene deletion lines(*26*). There are 54 protein complexes annotated by a previous study suitable for our analysis(*35*) . In 24 cases the overlaps of clustered effects are significantly more than what would be expected from distributed effects at a 99% confidence level, and the enrichments range from 2.7-fold to over 100-fold with a median 5.3-fold (Fig. 3A and Table S4). The overlapped clustered effects of each protein complex represent specific functions (Fig. 3B-C and Fig. S7). For example, the ten overlapped clustered effects of the elongator holoenzyme complex are tens to hundreds times overrepresented in a few transcription-related GO terms as well as proteasome-related GO terms (Fig. 3B), the former of which echo well the annotated functions of the complex while the latter appear to suggest new understandings(*36*). As another example, the genes encoding the protein kinase CK2 complex have 21 overlapped clustered effects that appear to affect specifically the metabolism of various amino acids (Fig. 3C), a functional insight not been well recognized(*37*).

We also checked genes on the same KEGG pathways. There are 41 pathways that are related to metabolism, genetic information processing, cellular processes, and so on, suitable for our analysis (Table S5). The rate of overlaps of clustered effects is significant higher than that of distributed effects in nine cases, and the enrichments range from 5.6-fold to over 100-fold with a median 46.9-fold (Fig. 3D and Table S5). Consistently, the overlapped clustered effects of each pathway represent distinct

functions (Fig. 3E-F and Fig. S7). For the many cases in which clustered effects show no more overlaps than distributed effects, the involved genes may execute additional functions irrelevant to the focal complex or pathway. Notably, in none of the cases distributed effects represent related biochemical understandings better than clustered effects, highlighting the cryptic nature of them. Taken together, focusing on clustered effects appears to be a readily operational approach to narrowing the gap between genetic analysis and biochemical data.

The above analyses considered gene expression traits. We next examined the yeast cell morphological traits that are based on the microscopic images of cells stained by fluorescent dyes(*38*). With the help of a computer software as many as 405 quantitative traits can be obtained from cell wall and nuclear stained cell images(*39*). These traits are typically about area, distance, and angle calculated based on dozens of coordinate points, lines and angles that describe the shape of mother cell and bud, and the shape and localization of the nuclei in mother cell and bud (Fig. 4A). This large set of yeast traits had served as a valuable resource for studying genotype-phenotype relationships(*9, 40, 41*). Deletion of HAP4 in *S. cerevisiae* significantly altered 78 morphological traits, among which 24 are also significantly affected in *S. paradoxus* by HAP4 deletion (Table S6). To test if the evolutionarily conserved effects of HAP4 are shared with HAP2, HAP3 and HAP5 more than the non-conserved effects, we also measured the morphological traits affected by each of the other three genes, respectively, in *S. cerevisiae*. We found that 58.3% (14/24) of the conserved effects are shared with all the other three genes, which is significantly higher than the number (18/54 = 33.3%)

253 for the non-conserved effects of HAP4 (P = 0.035, Fisher's exact test; Fig. 4B). The
254 estimations are not explained by correlated traits (Fig. S8), and the difference remains
255 largely unchanged when only traits with small measuring noise are considered (Fig. S9).
256 Hence, the cell morphology data also support the role of evolution in separating genetic
257 effects.
258
259 **Discussion**
260 Thanks to the mature framework of measuring the selective constraints on DNA
261 sequence(*42*), the evolution-free functionality of DNA elements defined in ENCODE
262 was challenged immediately after its emergence(*11, 12*). Notably, the gene-trait
263 interactions defined in reverse genetic analyses are also based on an evolution-free
264 framework. However, this century-old problem has been largely ignored, despite
265 exceptions(*43, 44*), primarily due to the lack of a readily available measure of the
266 underlying evolutionary constraint. In this study we performed, for the first time to
267 the best of our knowledge, a rigorous test of the evolutionary nature of a set of gene
268 deletion effects by examining their within-population heritability and intra-/inter-
269 species conservation. We found only some of them subject to effective selection, with
270 the rest likely being *ad hoc* and non-evolutionary. That being said, we cautioned some
271 effects might be under very weak selection that was beyond the detection power of our
272 analyses. This concern would be alleviated by a reasonable assumption that effects
273 under very weak selection are not distinct from those under no selection in the
274 functional properties examined in the study. Similar to the *ad hoc* "functional" DNA

275 elements defined in ENCODE(*10*), the *ad hoc* genetic effects are presumably explained
276 by mutation equilibrium or spurious functions arising from the gene deletion(*16*).
277 Importantly, since such *ad hoc* effects have not yet been shaped by evolution, they are
278 unlikely to be compatible with the roles the focal gene has long played in evolution(*9*).
279 This may explain in great part the origin of gene pleiotropy.

280 Conceptually speaking, our evolutionary view on genetic effects is an extension of
281 the evolutionary view on the biochemical activities of DNA elements(*11, 12*). Hence,
282 pros and cons that have widely discussed in the debates on the ENCODE project apply
283 similarly to this study. For example, because detecting selection involves multiple
284 lineages, one cannot rule out the possible that an entity under no detectable selection is
285 actually subject to lineage-specific selection(*11*). However, since the lineages
286 examined are often closely related, lineage-specific entities selected in a short time
287 window should be rare compared to those acquired during the long time period
288 predating the split of the lineages. Operationally speaking, the evolutionary view on
289 the functionality of DNA elements relies on DNA sequence comparison, which is
290 straightforward and now mature. However, an evolutionary separation of genetic
291 effects requires rather complex experimental designs; also, there is no available
292 framework for modeling the turnover rate of gene-trait interactions under no selection.
293 Hence, we could, as in this study, only perform enrichment analysis for a group of
294 genetic effects. Nevertheless, the current limitation in operationality does not
295 challenge the validity of the concept. A surprising finding of this study is GO
296 clustering can serve as a useful and readily operational proxy for selection when

297 expression traits are examined. The underlying rationale, namely, functional
298 coordination built by selection, may help us design more efficient strategies for
299 delineating the evolutionary nature of genetic effects in the future.

300 In summary, by examining the evolutionary nature of a set of gene deletion effects
301 we revealed widespread decoupling of gene deletion effects from gene functions. This
302 calls for an expanded framework for reverse genetic analysis (Fig. 4C). Specifically,
303 the conventional framework relies solely on statistical tests to separate the mutant
304 versus wild-type differences into significant and insignificant effects. In the expanded
305 framework significant effects are further separated into evolutionarily selected and
306 evolutionarily *ad hoc* ones. Only the former would support the biochemical
307 understandings with the latter being pleiotropic and decoupled from the gene functions.

308

309 **Materials and Methods**

310 **Yeast gene deletions**

311 Three *S. cerevisiae* (SC) strains BY4741 (*MATa, his3, leu2, met15, ura3*), GIL104
312 (*MATa, URA3, leu2, trp1, CAN1, ade2, his3, bar1Δ::ADE2*; derived from the W303)
313 and DBVPG1373 (*MATa, ura3*), and one *S. paradoxus* (SP) strain N17 (*MATa, ura3*)
314 were included in the study. Unless otherwise stated, the *S. cerevisiae* strains were
315 cultured in the rich medium YPD (1%Yeast extract, 1%Peptone, 2% Dextrose) at 30°C,
316 and *S. paradoxus* N17 was cultured in YPD at 25°C. The wild type URA3 in GIL104
317 was first replaced by a LEU2 cassette. HAP4 was replaced by a URA3 cassette in
318 each of the four strains. HAP2, HAP3, and HAP5 were also replaced, respectively,
319 by a URA3 cassette in BY4741. Notably, for all gene replacements the whole open
320 reading frame from the starting codon to the stop codon of a focal gene was replaced.
321 As described in our previous study(*22*), the standard LiAc transformation method(*45*)
322 was used to transform DNA into the yeast cells and gene replacements were achieved
323 by homologous recombination. The transformation protocol was slightly modified for
324 *S. paradoxus*(*46*); specifically, heat shock was performed for seven minutes at 37°C.
325 Synthetic medium deprived of uracil or leucine was used to select the clones with
326 successful replacement for the target gene. All gene replacements were verified using
327 polymerase chain reaction (PCR). For each gene deletion line, 3-5 independent clones
328 were obtained for further examination, which effectively controlled the potential effects

of secondary mutations introduced during the gene replacement.

Because haploid yeast cells tend to flocculate, which is not suitable for cell morphology characterization, diploid yeasts are required in the analysis of morphological traits. We first obtained haploid gene deletion strains (SC-BY4741 or SP-N17 background; *MATa*), which were then crossed with the corresponding *MATalpha* wild-type strain, respectively. The diploid heterozygous gene deletion strains were sporulated by following the method of a previous study(*47*). Specifically, the cells were incubated in YEP (1% yeast extract,1% Bacto peptone,0.05% NaCl) containing 2% potassium acetate) for five hours at 30 °C to start the sporulation process. The culture was centrifuged (2,000g, for 2 min), the cell pellet was washed three times by sterile water, and re-suspended in sporulation media (10g/l potassium acetate and 50mg/l zinc acetate) for five days at 25°C with shaking. The products were incubated with 200U/ml lyticase (Sigma #L4025) for 30 min at 30°C followed by 15 min at 50°C. The products were washed by sterile water and then plated on synthetic medium deprived of uracil for two days at 30°C for SC strains, or 25°C for SP strains. The genotypes of the colonies were determined by PCR. For each gene the haploid deletion strains of both mating types were obtained. A pair of *MATa* and *MATalpha* strains with the same gene deletion were crossed to obtain a diploid homozygous gene deletion strain. For each gene deletion line three independent clones were obtained and examined.

Obtain a population of segregants

A wild-type strain BY4742 (*MATalpha*), which is identical to BY4741 except at the mating locus, was crossed with a HAP4 deletion stain of DBVPG1373 (*MATa*). Two biological replications were carried out. The diploid heterozygous deletion strains were sporulated for 3-5 days in sporulation medium on a shaking table at 25°C. Tetrads were obtained and incubated with 200U/ml lyticase for 3-5 min at 30°C, and then streaked onto a YPD plate for tetrad dissection using the MSM400 dissection microscope (Singer Instrument Company Ltd). Spores were grown on YPD plates at 30°C for two days, and the genotypes of the colonies were determined by PCR. We selected only those tetrads that produce four segregants with genotypes *MATa*+HAP4, *MATa*+*Δhap*4, *MATalpha*+HAP4, and *MATalpha*+*Δhap4*, respectively. A total of 24 segregants from six such tetrads were obtained for the heritability analysis.

RNA sequencing and data analysis

For each strain a single colony on agar plate was picked and cultured in YPD liquid overnight at 30 °C with shaking. Approximately 200 μl saturated culture was added into 10ml fresh YPD, which resulted in an optical density OD600 ~ 0.1 (UNICO UV/VIS Spectrophotometer), and cells of 3ml culture at OD600 = 0.5-0.65 were harvested. Total RNA was extracted by QIAGEN RNeasy Plus mini kit (Cat No.74136). The mRNA sequencing was performed using the paired-end module on a HiSeq platform at Genwiz by following the standard procedure. To ensure the high quality of expression analysis, we sequenced the mRNA of 3-6 independent clones for each wild-type or gene deletion line.

RNA-seq reads were mapped to reference yeast genomes using STAR (Version 2.6.0c)(*48*). For BY4741 and GIL104, we used the genome of *S. cerevisiae* strain S288C as the reference (version R64-2-1_20150113; http://www.yeastgenome.org). The reference genomes of SC-DBVPG1373 and SP-N17 were downloaded from SGRP (https://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html). For a typical clone there were about 6.5 million paired-end reads mapped to the coding sequences. Gene expression levels were determined by Featurecounts (version 1.6.2)(*49*) with default settings and RPKM (reads per kilobase per million) of each gene were calculated by R package edgeR. The wild-type versus mutant differential expression analysis was performed by DESeq2(*50*) with default parameters, and genes with an adjusted P-value smaller than 0.05 and a fold change (FC) greater than 1.5 were defined as significantly changed genes. In the conservation, heritability or overlapping analysis, an effect is called conserved, heritable or overlapped only when it shows the same direction in the various conditions examined. Genes of the uracil biosynthesis pathway (YBL039C, YEL021W, YJL130C, YJR103W, YKL024C, YKL216W, YLR420W, YML106W, YMR271C) were excluded from further analysis. The expression level measured by RPKM of HAP2, HAP3, HAP4 and HAP5 in wild-type BY4741 is 84.4, 64, 194 and 90.5, respectively, and all becomes zero in the corresponding deletion lines. The fitness of yeast gene deletion lines was produced by a previous study(*51*). Table S1 contains details of the RNA-seq expression information of all yeast lines examined in this study.

GO analysis

The GO analyses of the responsive genes derived from our RNA-seq data were conducted in the SGD website using GO Term Finder (Version 0.86; https://www.yeastgenome.org/goTermFinder), by excluding computational analysis evidences and other less reliable evidences: IBA, IC, IEA, IKR, IRD, ISA, ISM, ISO, ISS, NAS, ND, TAS. In a strict analysis which required the exclusion of all expression-related evidences, only three GO evidence codes IDA, HDA and IPI that represent direct experimental assays were considered. For the GO analyses of public microarray data the R package clusterProfiler(*52*) was used with default settings. The cutoffs used to define an enriched GO term include an adjusted P-value smaller than 0.01 and a fold enrichment greater than 2. To improve specificity only GO terms containing less than 200 genes were considered. The fold enrichment was calculated as (number of changed genes in the GO term / number of all changed genes) / (number of genes in the GO term / number of genes in all GO terms of the class). The GO semantic similarity scores were calculated by R package GOSemSim(*53*).

Heritability analysis

Following a previous study(*54*), for each of the 195 genes the expression is expressed as $y = \mu 1_N + u + e$, where y is a vector of the expression level ($\log_2$RPKM) in the 24 segregants, μ is the mean expression level in the 24 segregants, 1_N is a vector of N ones, u is a vector of random additive genetic effects from the HAP4 locus, and e is a vector of residuals. The variance structure of an expression trait is written as $V = A\sigma_u^2 +$

$I\sigma_e^2$, where A is relatedness matrix based exclusively on the HAP4 locus (1 for wild-type allele and 0 for null allele), I is identity matrix, σ_u^2 is additive genetic variance explained by HAP4 locus, and σ_e^2 is error variance. Then, the value of h^2_{HAP4} is equal to $\sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. R package rrBLUP was used to estimate the variance components.

To test the statistical significant of an h^2_{HAP4}, 24 segregants were divided into two groups: 12 with the wild-type allele of HAP4 and 12 with the null allele of HAP4. We compared the expression levels of the focal gene between the two segregant groups using DESeq2. The obtained 195 raw P-values were adjusted for multiple testing using the Benjamini-Hochberg controlling procedure. An adjusted P-value smaller than 0.05 was considered significant.

Analyze protein complexes and metabolic pathways

The public microarray data of ~1,400 yeast gene deletion lines were obtained from a previous study(*26*), and P-values and fold changes (FC) provided in the data were directly used. Specifically, $P < 0.05$ and absolute $FC > 1.2$ were used to define genes with significant expression changes; if the number of significantly changed genes was over 1,500, a more stringent cutoff $P < 0.01$ was used. To avoid the effects of genes with ubiquitous expression responses we excluded from further analyses the top 10% genes that each show significant changes in at least 12% of the gene deletion lines. GO analyses were performed by R package clusterProfiler to define clustered and distributed effects for each deletion line, with the results summarized in Table S7. To examine the overlapped clustered effects between genes of the same protein complex or pathway, we only considered the deletion lines with at least 20 clustered effects, resulting in a set of 422 deletion lines suitable for further analyses.

Information of 518 protein complexes was obtained from a previous study (*35*). The KEGG pathways of the yeast *S. cerevisiae* were downloaded from KEGG website (https://www.genome.jp/kegg-bin/get_htext?sce00001). There are 54 complexes and 41 pathways each with at least two member genes found in the above defined mutant set.

For each protein complex or pathway, the overlaps of clustered effects and the overlaps of distributed effects were compared in number. The numbers of clustered effects and distributed effects of the involved genes were normalized to make the overlaps between the two effect types comparable. To estimate the confidence interval of a comparison we used random samplings. If clustered effects are less than distributed effects in all genes, which is true in most of the cases examined, we sampled (without replacements) a random subset of distributed effects to ensure the two effect types of a gene equal in number. If clustered effects are more than distributed effects in all genes, we sampled (without replacements) a random subset of distributed effects to ensure the two effect types of a gene equal in number. If the above consistent patterns do not exist, we sampled consistently from one side (either clustered effects or distributed effects) but with replacements for the gene with an insufficient number of effects on this side. For each complex or pathway 1,000 such random samplings were carried out to derive the 99% confidence interval, and an observed difference is called

significant if it is not within the interval. Table S4 and Table S5 have details about the protein complexes and KEGG pathways examined, respectively.

Analyze cell morphological traits

Diploid yeast cells were examined by following the protocol of previous studies with slight modifications(*38, 39*). In brief, a single yeast colony was picked and cultured in YPD liquid overnight with shaking to the saturation phase. Then, 1.5 μl culture was transferred to 100μl fresh YPD in a 96-well plate and grew for 3-4 hours at 30°C for SC strains or 25°C for SP strains. Cells were fixed with 3.7% formaldehyde solution. Cell wall was stained by FITC-ConA (fluorescein isothiocyanate-conjugated, concanavalin A, Sigma-Aldrich C7642). Cell nucleus was stained by hochest-mix (Thermo Fisher, Hoechst 33342 Solution) instead of DAPI to enhance the specificity. We did not stain actin because the dye Rhodamine phalloidin was not stable enough to support the following high-throughput automated image capturing which takes about 10 hours for scanning 96 wells of a plate. The stained cells were plated into a microplate (Greiner 781091) with ~5.0×10^4 cells per well and images were captured by IN Cell Analyzer 2200 (GE Healthcare) using the 60× objective lens.

Five SC lines (all diploid with BY4741 background: wild-type, *Δhap2*, *Δhap3*, *Δhap4* and *Δhap5*) and two SP lines (all diploid with N17 background: wild-type and *Δhap4*) were examined. Because the trait measuring is quite sensitive to batch effect, for each line we conducted 18-24 replicates of staining and image capturing. The images were analyzed by CalMorph(*38, 39*) with default settings, and only 405 rather than 501 traits were extracted in this study because actin is not stained. At least 1,000 cells were captured and analyzed (with at least 100 informative cells for each cell-cycle stage) for a high-quality replicate. In the end, there were 13~23 high-quality replicates for each of the lines included in further analysis. Trait values were compared between replicates of a gene deletion line and replicates of the corresponding wild-type line using T-test, and the resulting 405 P-values were adjusted for multiple testing using the Benjamini-Hochberg controlling procedure. Because of the many replicates included in the comparison, many traits showed a statistically significant but biologically negligible difference between wild-type and mutant lines. Hence, a trait is called affected by a gene only when the adjusted $P < 0.05$ and the difference between wild-type and mutant is large than 5%. Table S8 has complete information regarding the morphological trait analysis.

Acknowledgments

We are grateful to financial support from NSFC (grant #31630042 and #91731302 to X. H.), technical support from Z. Zhou and X. Chen, and helpful discussions with C. Wu, J. Yang, J. Zhang, W. Qian, and Y. Zhang.

References

1. H. J. Muller, Further studies on the nature and causes of gene mutations. *Proceedings of the 6th International Congress of Genetics*, 213-255 (1932).
2. Z. Wang, B. Y. Liao, J. Zhang, Genomic patterns of pleiotropy and the evolution of complexity.

Proc Natl Acad Sci U S A **107**, 18034-18039 (2010).

3. E. J. E. Caspari, PLEIOTROPIC GENE ACTION. **6**, 1-18 (1952).
4. B. Sonnichsen *et al.*, Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature* **434**, 462-469 (2005).
5. A. B. Paaby, M. V. Rockman, The many faces of pleiotropy. *Trends Genet* **29**, 66-73 (2013).
6. F. W. Stearns, One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767-773 (2010).
7. X. L. He, J. Z. Zhang, Toward a molecular understanding of pleiotropy. *Genetics* **173**, 1885-1891 (2006).
8. F. J. Ayala, "Nothing in biology makes sense except in the light of evolution": Theodosius Dobzhansky: 1900-1975. *J Hered* **68**, 3-10 (1977).
9. H. Chen, C. I. Wu, X. He, The Genotype-Phenotype Relationships in the Light of Natural Selection. *Mol Biol Evol* **35**, 525-542 (2018).
10. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
11. D. Graur *et al.*, On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biol Evol* **5**, 578-590 (2013).
12. W. F. Doolittle, Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A* **110**, 5294-5300 (2013).
13. M. Kellis *et al.*, Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131-6138 (2014).
14. A. J. F. Griffiths, *An Introduction to genetic analysis*. (W.H. Freeman, New York, ed. 5th, 1993), pp. xi, 840 p.
15. D. Alzoubi, A. A. Desouki, M. J. Lercher, Alleles of a gene differ in pleiotropy, often mediated through currency metabolite production, in E. coli and yeast metabolic simulations. *Sci Rep* **8**, 17252 (2018).
16. X. He, The Biology Complicated by Genetic Analysis. *Mol Biol Evol* **33**, 2177-2181 (2016).
17. S. L. Forsburg, L. Guarente, Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev* **3**, 1166-1178 (1989).
18. M. Bolotin-Fukuhara, Thirty years of the HAP2/3/4/5 complex. *Biochim Biophys Acta Gene Regul Mech* **1860**, 543-559 (2017).
19. S. Buschlen *et al.*, The S. Cerevisiae HAP complex, a key regulator of mitochondrial function, coordinates nuclear and mitochondrial gene expression. *Comparative and functional genomics* **4**, 37-46 (2003).
20. D. W. Threadgill *et al.*, Targeted disruption of mouse EGF receptor: effect of genetic background on mutant phenotype. *Science* **269**, 230-234 (1995).
21. V. Vu *et al.*, Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. *Cell* **162**, 391-402 (2015).
22. P. Chen, D. Wang, H. Chen, Z. Zhou, X. He, The nonessentiality of essential genes in yeast provides therapeutic insights into a human disease. *Genome Res* **26**, 1355-1362 (2016).
23. A. Evangelou *et al.*, Unpredictable Effects of the Genetic Background of Transgenic Lines in Physiological Quantitative Traits. *G3 (Bethesda)* **9**, 3877-3890 (2019).
24. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. S. Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254 (2003).
25. G. Liti *et al.*, Population genomics of domestic and wild yeasts. *Nature* **458**, 337-341 (2009).

26. P. Kemmeren *et al.*, Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**, 740-752 (2014).
27. C. T. Harbison *et al.*, Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).
28. K. D. MacIsaac *et al.*, An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 113 (2006).
29. K. Neander, Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of Science* **58**, 168-184 (1991).
30. R. G. Millikan, In Defense of Proper Functions. *Philosophy of Science* **56**, 288-302 (1989).
31. R. Amundson, G. V. Lauder, Function without purpose. *Biology and Philosophy* **9**, 443-469 (1994).
32. S. L. Wong, L. V. Zhang, F. P. Roth, Discovering functional relationships: biochemistry versus genetics. *Trends Genet* **21**, 424-427 (2005).
33. X. He, W. Qian, Z. Wang, Y. Li, J. Zhang, Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat Genet* **42**, 272-276 (2010).
34. M. Costanzo *et al.*, A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, (2016).
35. J. J. Benschop *et al.*, A consensus of core protein complex compositions for Saccharomyces cerevisiae. *Mol Cell* **38**, 916-928 (2010).
36. J. Q. Svejstrup, Elongator complex: how many roles does it play? *Current Opinion in Cell Biology* **19**, 331-336 (2007).
37. D. W. Litchfield, Protein kinase CK2: structure, regulation and role in cellular decisions of life and death. *The Biochemical journal* **369**, 1-15 (2003).
38. Y. Ohya *et al.*, High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci U S A* **102**, 19015-19020 (2005).
39. H. Okada, S. Ohnuki, Y. Ohya, Quantification of cell, actin, and nuclear DNA morphology with high-throughput microscopy and CalMorph. *Cold Spring Harb Protoc* **2015**, 408-412 (2015).
40. W. C. Ho, J. Zhang, The genotype-phenotype map of yeast complex traits: basic parameters and the role of natural selection. *Molecular biology and evolution* **31**, 1568-1580 (2014).
41. S. Nogami, Y. Ohya, G. Yvert, Genetic complexity and quantitative trait loci mapping of yeast morphological traits. *Plos Genet* **3**, e31 (2007).
42. D. Graur, W.-H. Li, *Fundamentals of molecular evolution*. (Sinauer Associates, Sunderland, Mass., ed. 2nd, 2000), pp. xiv, 481 p.
43. C. H. Chandler, S. Chari, I. Dworkin, Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends Genet* **29**, 358-366 (2013).
44. C. H. Chandler *et al.*, How well do you know your mutation? Complex effects of genetic background on expressivity, complementation, and ordering of allelic effects. *PLoS Genet* **13**, e1007075 (2017).
45. R. D. Gietz, R. H. J. N. p. Schiestl, High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31-34 (2007).
46. D. R. Scannell *et al.*, The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3-Genes Genom Genet* **1**, 11-25 (2011).

47. A. H. Enyenihi, W. S. Saunders, Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae. *Genetics* **163**, 47-54 (2003).
48. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
49. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
50. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
51. W. Qian, D. Ma, C. Xiao, Z. Wang, J. Zhang, The Genomic Landscape and Evolutionary Resolution of Antagonistic Pleiotropy in Yeast. *Cell reports* **2**, 1399-1410 (2012).
52. G. C. Yu, L. G. Wang, Y. Y. Han, Q. Y. He, clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omics-a Journal of Integrative Biology* **16**, 284-287 (2012).
53. G. C. Yu *et al.*, GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976-978 (2010).
54. S. H. Lee, N. R. Wray, M. E. Goddard, P. M. Visscher, Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).

Figure legends

Fig. 1. Tests of evolutionary effectiveness of HAP4 deletion effects.

(A): The 195 responsive genes are enriched in dozens of GO terms that are related and also reflect well the functional annotations of HAP4. The heatmap shows the pairwise similarity of the enriched GO terms, with three subclasses each corresponding to certain biological processes that are summarized at the right.

(B): The *P*-value (adjusted for multiple testing) and fold change (FC) of the 65 clustered effects and 130 distributed effects. Each dot represents a responsive gene (i.e., an effect).

(C): Obtain a population of segregants with different genetic backgrounds to test the heritability of HAP4 deletion effects.

(D): The 65 clustered effects have greater h^2_{HAP4} than the 130 distributed effects ($P = 7.8\times10^{-4}$, Mann-Whitney U-test). Each dot represents an effect, and *P*-value measures the statistical significance of h^2_{HAP4}, with the vertical dashed line showing adjusted $P = 0.05$.

(E): The proportions of deletion effects that are significantly heritable with adjusted $P < 0.05$. Error bars represent SE.

(F): A dendrogram showing the phylogeny the four yeast strains examined in this study.

(G): Conservation analysis of the HAP4 deletion effects. The 195 responsive genes defined in BY4741(*Δhap4*) are examined with respect to their expression responses in *S. paradoxus* N17 (*Δhap4*). The horizontal dashed line shows adjusted $P = 0.05$ and vertical dashed lines show $\log_2FC = \pm0.58$. (cyan: clustered effects; red: distributed effects; cycle: down-regulated in BY4741(*Δhap4*); triangle: up-regulated in BY4741(*Δhap4*))

(H): The rate of conservation in the three related yeasts for the 65 clustered effects and 130 distributed effects defined in BY4741(*Δhap4*), respectively. Error bars represent SE.

Fig. 2. Clustered effects of the four genes encoding the HAP2/3/4/5 tetramer overlap a lot more than their distributed effects do.

(A, B, C): The P-value (adjusted for multiple testing) and fold change (FC) of the clustered effects and distributed effects defined in the three BY4741 strains *Δhap2*, *Δhap3*, and *Δhap5*, respectively. Each dot represents a responsive gene (i.e., an effect), and the total number of responsive genes is shown at the bottom next to the effect type.

(D): There are 20 overlapped clustered effects for the four genes encoding the HAP2/3/4/5 tetramer, which is significantly higher than expectation. The expectation is estimated by random sampling of the distributed effects of the four genes to calculate overlaps, and 1,000 such simulations were conducted.

(E): Among the 195 responsive genes found in BY4741(*Δhap4*) 13 are direct target of HAP4 according to a chromatin immune-precipitation assay.

(F): The proportion of direct target of HAP4 in different gene sets. Error bars represent SE.

Fig. 3. Clustered effects support related biochemistry understandings much

better than distributed effects in a variety of protein complexes and KEGG pathways.

(A): The clustered effects of genes encoding a protein complex in general overlap more than their distributed effects. Each circle represents a complex, and the filled ones are significant at a 99% confidence level estimated by random sampling. A total of 54 protein complexes are included here, with 24 cases showing at least twice more overlapped clustered effects than overlapped distributed effects (below the line y = 0.5x). The numbers of effects have been normalized such that in each case the overlaps of clustered effects and the overlaps of distributed effects can be directly compared.

(B): The representative GO terms of the overlapped clustered effects of the elongator holoenzyme complex. Only four genes encoding the complex, which are highlighted in orange, have suitable expression data for the analysis. There are 10 overlapped clustered effects of the four genes, which are over 300 times more than expected. The expectation is estimated by random sampling of the distributed effects of the focal genes to calculate overlaps. The orange circle each represents an overlapped clustered effect, and the blue circles represent the enriched GO terms of the overlapped clustered effects with the number inside showing the fold enrichment in the given term.

(C): The representative GO terms of the overlapped clustered effects of the protein kinase CK2 complex. There are 21 overlapped clustered effects, 83.3 times more than expected.

(D): The clustered effects of genes in the same KEGG pathway also tend to overlap more than their distributed effects. Each circle represents a pathway, and the filled ones are significant at a 99% confidence level. A total of 41 pathways are included, with nine showing at least five times more overlapped clustered effects than overlapped distributed effects (below the line y = 0.2x). The number of effects have been normalized such that in each case the overlaps of clustered effects and the overlaps of distributed effects can be directly compared.

(E): The representative GO terms of the overlapped clustered effects of the four genes in the metabolic pathway sce00260. There are six overlapped clustered effects, over 100 times more than expected.

(F): The representative GO terms of the overlapped clustered effects of the four genes in the genetic information processing pathway sce03010. There are 10 overlapped clustered effects, 21.4 times more than expected.

Fig. 4. Examination of cell morphological traits also supports the role of evolution in separating genetic effects.

(A): The yeast cell morphology outlined by coordinate points, lines and angles (only some are shown) based on which a total of 405 quantitative traits can be derived by a computer software.

(B): The traits affected by HAP4 in both *S. cerevisiae* and *S. paradoxus* (i.e., conserved effects) are more likely to overlap with those affected by HAP2, HAP3 and HAP5 than the traits affected by HAP4 only in *S. cerevisiae* (non-conserved effects) (P = 0.035, one-tailed Fisher's exact test). A total of 78 morphological traits significantly affected

by HAP4 deletion in *S. cerevisiae* are examined, among which 24 are conserved effects and 54 non-conserved effects. Overlaps refer to traits significantly affected by all four gene deletions in *S. cerevisiae.* Error bars represent SE.

(C): Proposition of an expanded framework for reverse genetic analysis. Statistically significant genetic effects defined in conventional framework are further separated into evolutionarily selected and *ad hoc* ones, with the former supporting related biochemistry understandings and the latter being pleiotropic and decoupled from the gene's normal functions.

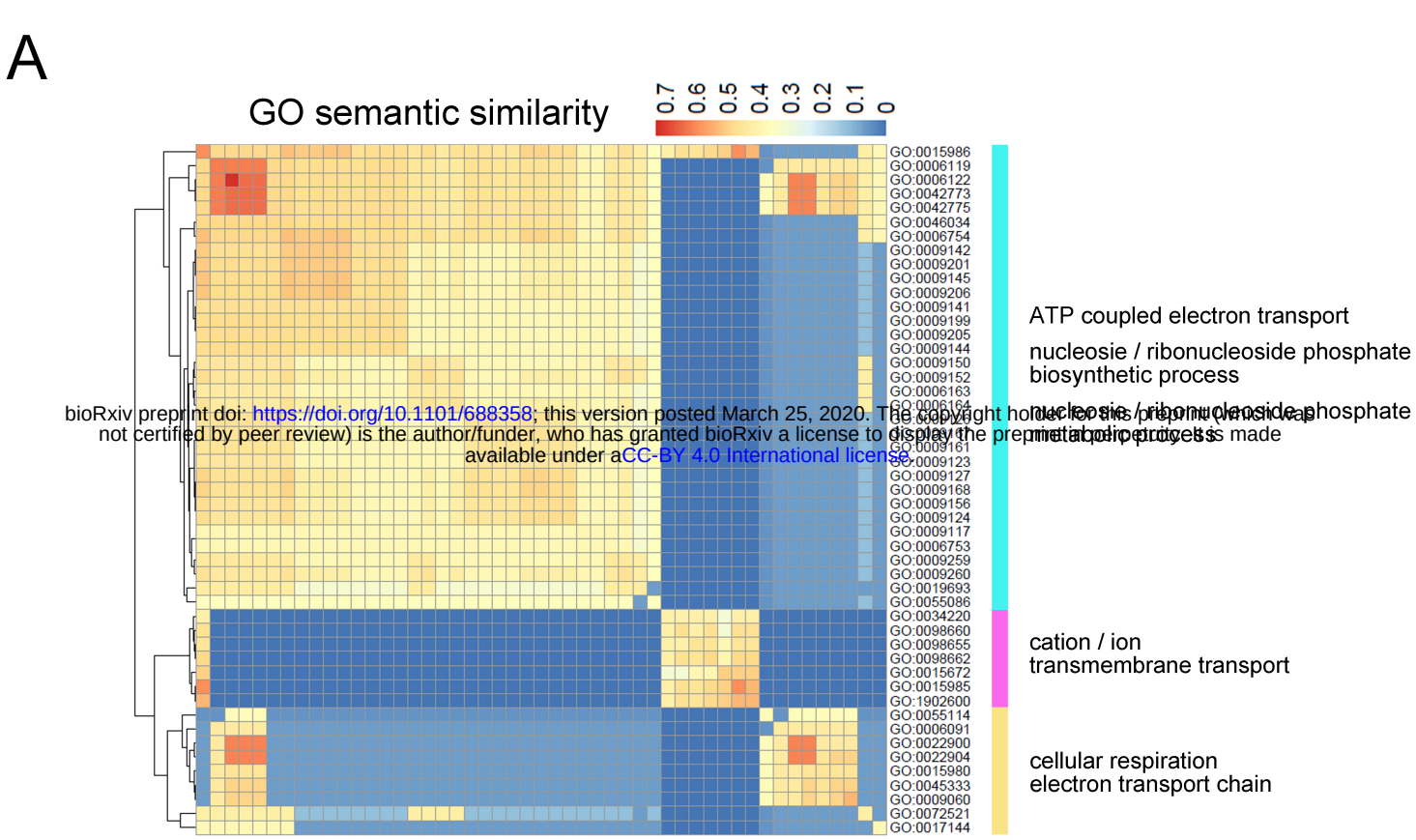


Fig. 1

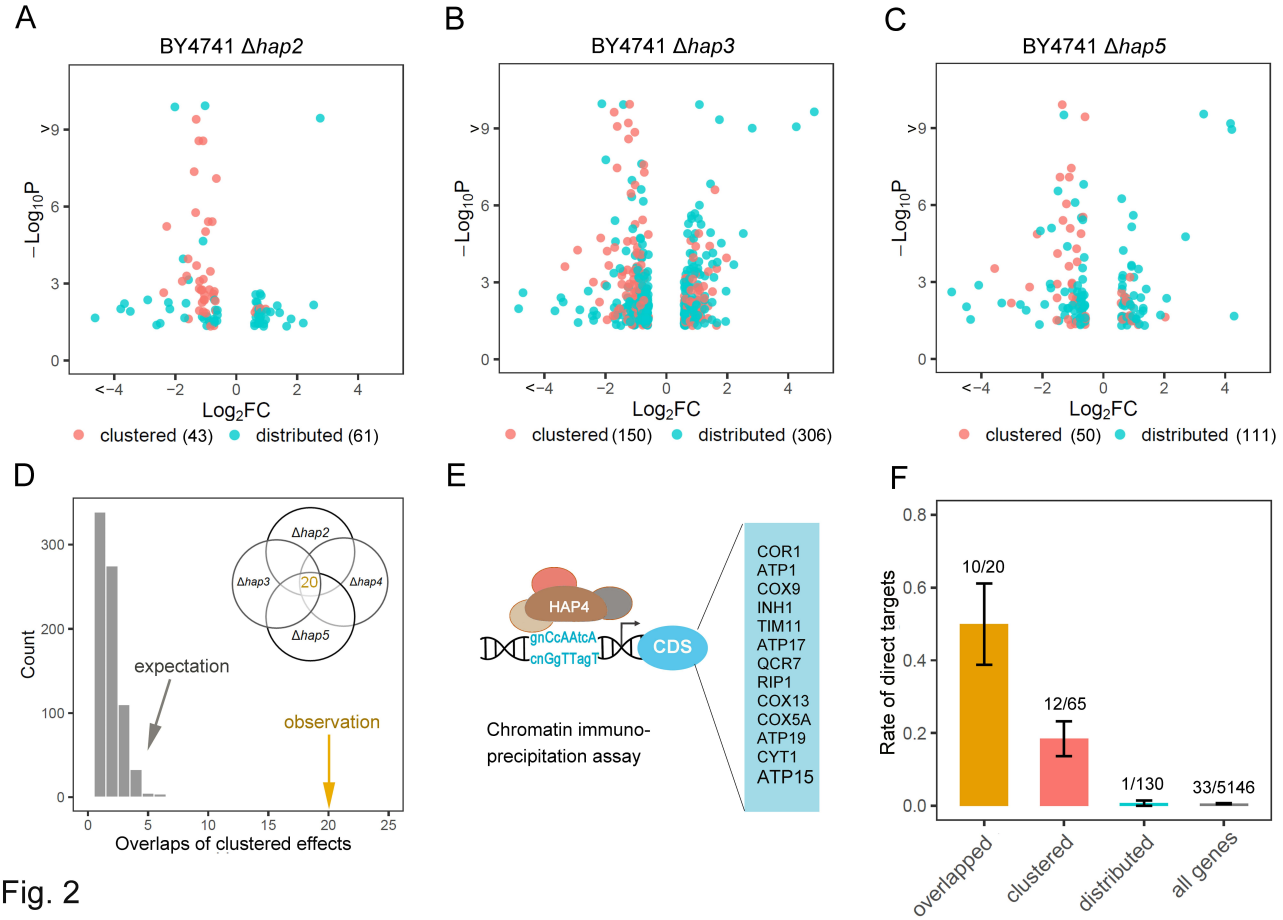
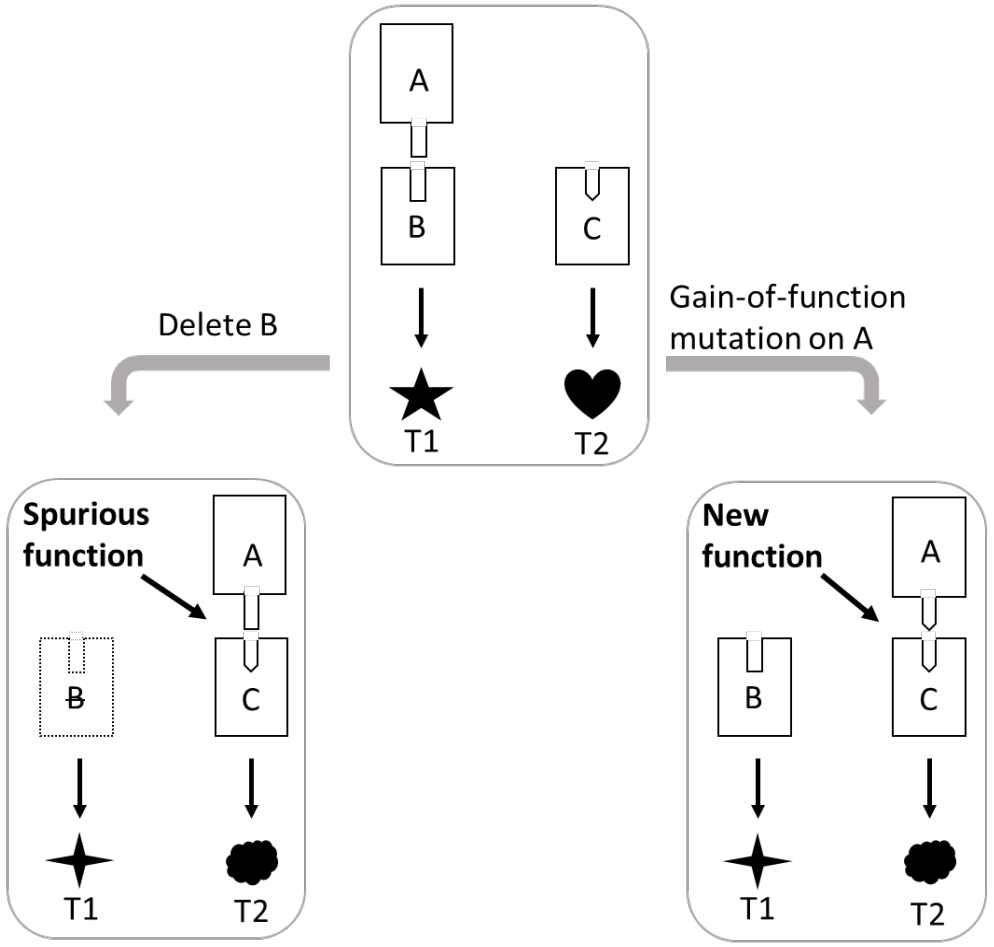


Fig. 2

Box 1



Nearly 90 years ago Hermann J. Muller coined the terms amorph, hypermorphy, hypomorph, antimorph, and neomorph to classify mutations based on their loss- or gain-of-function nature (see ref. 1). The basic idea of the classification has been fundamental to genetic analysis. In particular, **amorph** refers to null mutation on a gene, and the resulting phenotype is believed to represent the gene's native function. In contrast, **neomorph** refers to gain-of-function mutation on a gene, and the resulting phenotype does not represent the gene's native function. These concepts, although intuitively valid, lack rigorous tests. The diagram on the left illustrates how confusion could arise.

Suppose there is a living system with three genes (A, B and C) and two traits (T1 and T2). The function of proteins A and B is to form a dimer to regulate T1, and the function of protein C is to regulate T2. The understand the system we may apply genetic analysis. Deletion of B will break the A-B dimer, altering T1. This phenotype change represents the **native function** of B. However, when B is absent, A may find C to form a new, although less intimate, dimer A-C, which would alter T2. This is plausible since proteins with a structurally similar domain are prevalent in a eukaryotic genome. The change of T2 does not represent the native function of B; instead, it is explained by the non-native A-C dimer, a **spurious function** that arises from the deletion of B.

Notably, the spurious function arising from the deletion of B is by nature same as the **new function** caused by a gain-of-function mutation on A. It is well accepted that phenotype changes resulting from gain-of-function mutations do not represent native functions. From an evolutionary perspective only the A-B dimer is **selected**; the A-C dimer in both mutated systems is ***ad hoc***.

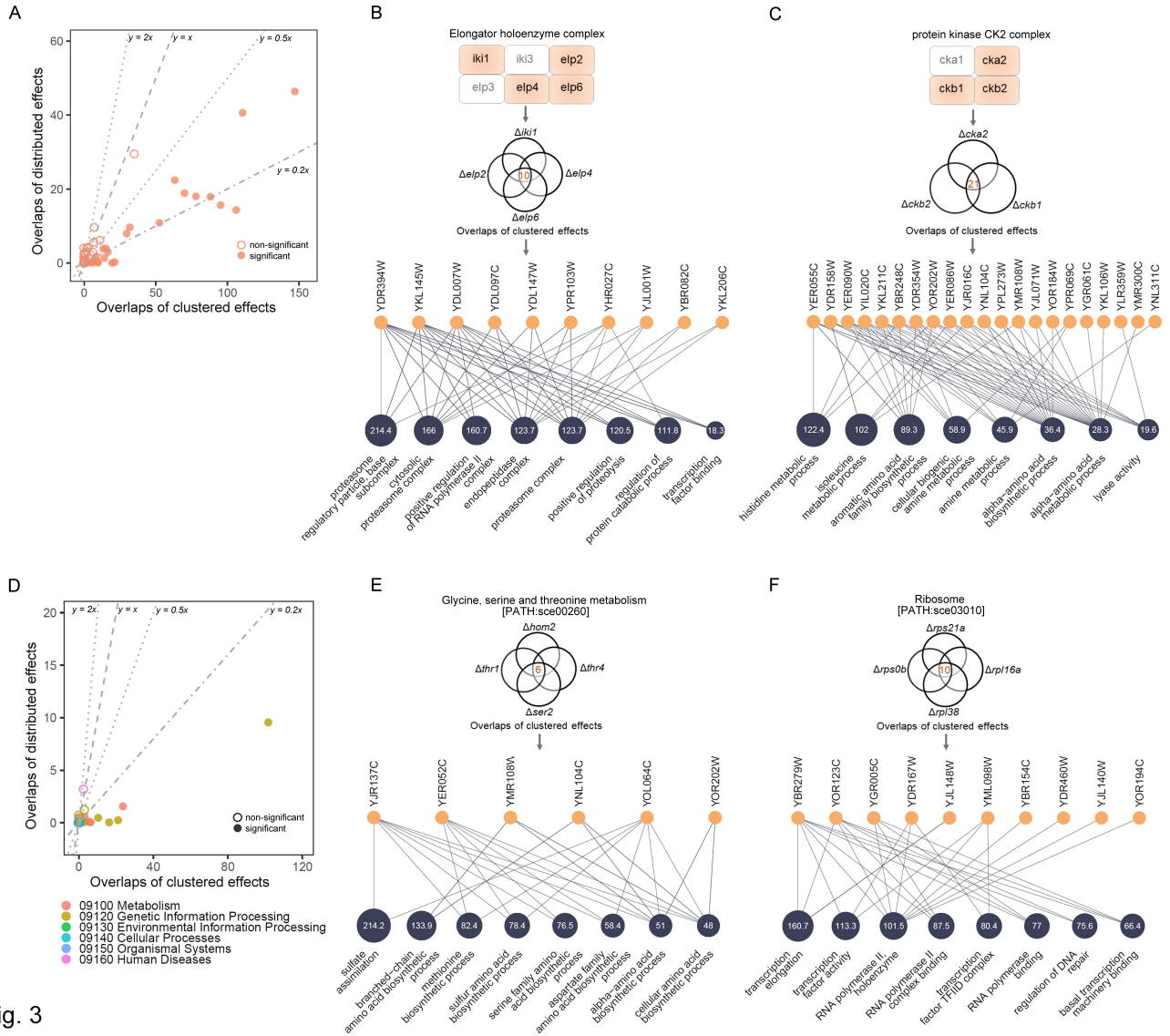
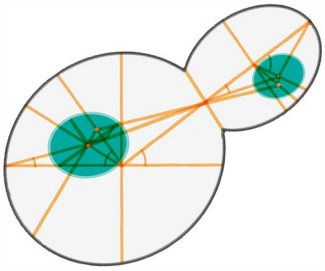
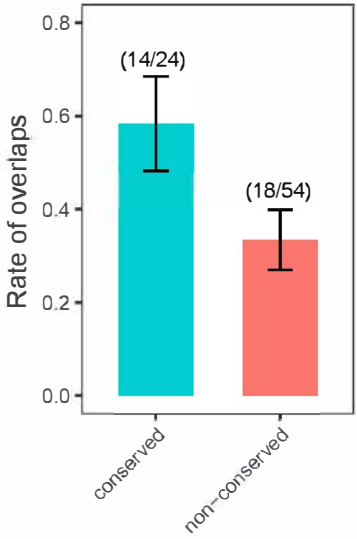

A
Overlaps of distributed effects
Overlaps of clustered effects
y = 2x
y = x
y = 0.5x
y = 0.2x
non-significant
significant
B
Elongator holoenzyme complex
iki1
iki3
elp2
elp3
elp4
elp6
Δiki1
Δelp2
Δelp4
Δelp6
11
10
Overlaps of clustered effects
YDR394W
YKL145W
YDL007W
YDL097C
YDL147W
YPR103W
YHR027C
YJL001W
YBR082C
YKL206C
214.4
166
160.7
123.7
123.7
120.5
111.8
18.3
proteasome regulatory particle base subcomplex
cytosolic proteasome complex
positive regulation of RNA polymerase II complex
endopeptidase complex
proteasome complex
positive regulation of proteolysis
regulation of protein catabolic process
transcription factor binding
C
protein kinase CK2 complex
cka1
cka2
ckb1
ckb2
Δcka2
Δckb2
Δckb1
21
Overlaps of clustered effects
YER055C
YDR158W
YER090W
YIL020C
YKL211C
YBR248C
YDR354W
YOR202W
YER086W
YJR016C
YNL104C
YPL273W
YMR108W
YJL071W
YOR184W
YPR069C
YGR061C
YKL106W
YLR359W
YMR300C
YNL311C
122.4
102
89.3
58.9
45.9
36.4
28.3
19.6
histidine metabolic process
isoleucine metabolic process
aromatic amino acid family biosynthetic process
cellular biogenic amine metabolic process
amine metabolic process
alpha-amino acid biosynthetic process
alpha-amino acid metabolic process
lyase activity
D
Overlaps of distributed effects
Overlaps of clustered effects
y = 2x
y = x
y = 0.5x
y = 0.2x
non-significant
significant
09100 Metabolism
09120 Genetic Information Processing
09130 Environmental Information Processing
09140 Cellular Processes
09150 Organismal Systems
09160 Human Diseases
E
Glycine, serine and threonine metabolism
[PATH:sce00260]
Δhom2
Δthr1
Δthr4
Δser2
6
Overlaps of clustered effects
YJR137C
YER052C
YMR108W
YNL104C
YOL064C
YOR202W
214.2
133.9
82.4
78.4
76.5
58.4
51
48
sulfate assimilation
branched-chain amino acid biosynthetic process
methionine biosynthetic process
sulfur amino acid biosynthetic process
serine family amino acid biosynthetic process
aspartate family amino acid biosynthetic process
alpha-amino acid biosynthetic process
cellular amino acid biosynthetic process
F
Ribosome
[PATH:sce03010]
Δrps21a
Δrps0b
Δrpl16a
Δrpl38
10
Overlaps of clustered effects
YBR279W
YOR123C
YGR005C
YDR167W
YJL148W
YML098W
YBR154C
YDR460W
YJL140W
YOR194C
160.7
113.3
101.5
87.5
80.4
77
75.6
66.4
transcription elongation
transcription factor activity
RNA polymerase II holoenzyme
RNA polymerase II complex binding
transcription factor TFIID complex
RNA polymerase binding
regulation of DNA repair
basal transcription machinery binding

Fig. 3

A



B


0.8
0.6
0.4
0.2
0.0
Rate of overlaps
(14/24)
(18/54)
conserved
non-conserved

C

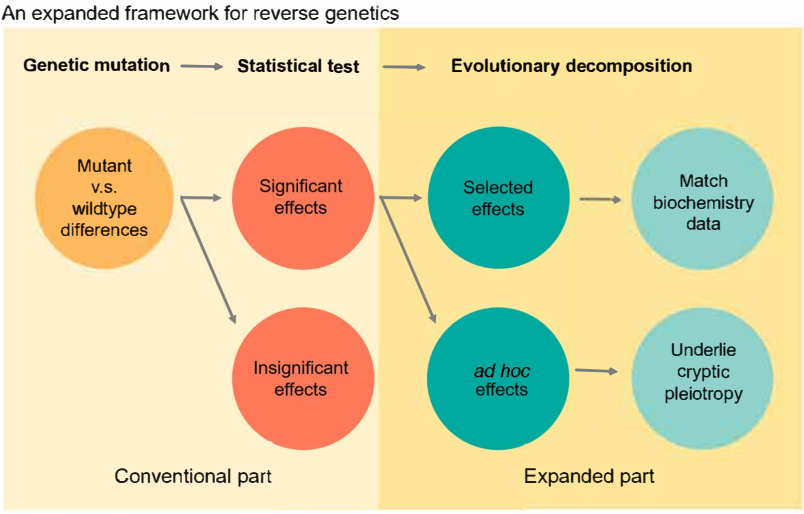

An expanded framework for reverse genetics
Genetic mutation
Statistical test
Evolutionary decomposition
Mutant v.s. wildtype differences
Significant effects
Insignificant effects
Selected effects
ad hoc effects
Match biochemistry data
Underlie cryptic pleiotropy
Conventional part
Expanded part

Fig. 4

Supporting Information of

"Decoupling gene knockout effects from gene functions by evolutionary analyses"

Li Liu[#], Mengdi Liu[#], Di Zhang, Shanjun Deng, Piaopiao Chen, Jing Yang, Yunhan Xie & Xionglei He*

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

This file contains:

Legends of Tables S1 to S8

Figs. S1 to S9

Legends of supplementary tables

Table S1: RNA-seq-based gene expression levels for each HAP4 deletion and wild-type strain.

Table S2: RNA-seq-based expression changes (P-value and FC) of the 195 responsive genes in BY4741(*Δhap4*) and other deletion lines, with the values of h^2_{HAP4} also included.

Table S3: RNA-seq-based gene expression changes after deleting HAP2, HAP3, HAP4 and HAP5, respectively, in BY4741.

Table S4: Summary of the analyses of protein complexes in this study.

Table S5: Summary of the analyses of KEGG pathways in this study.

Table S6: The affected morphological traits in a variety of gene deletion lines.

Table S7: Summary of the clustered effects and distributed effects defined in each mutant that has public microarray data.

Table S8: Summary of the trait information of each diploid gene deletion or wild-type yeast strain, with the 405 trait values, the number of examined cells, and the number of replications included.

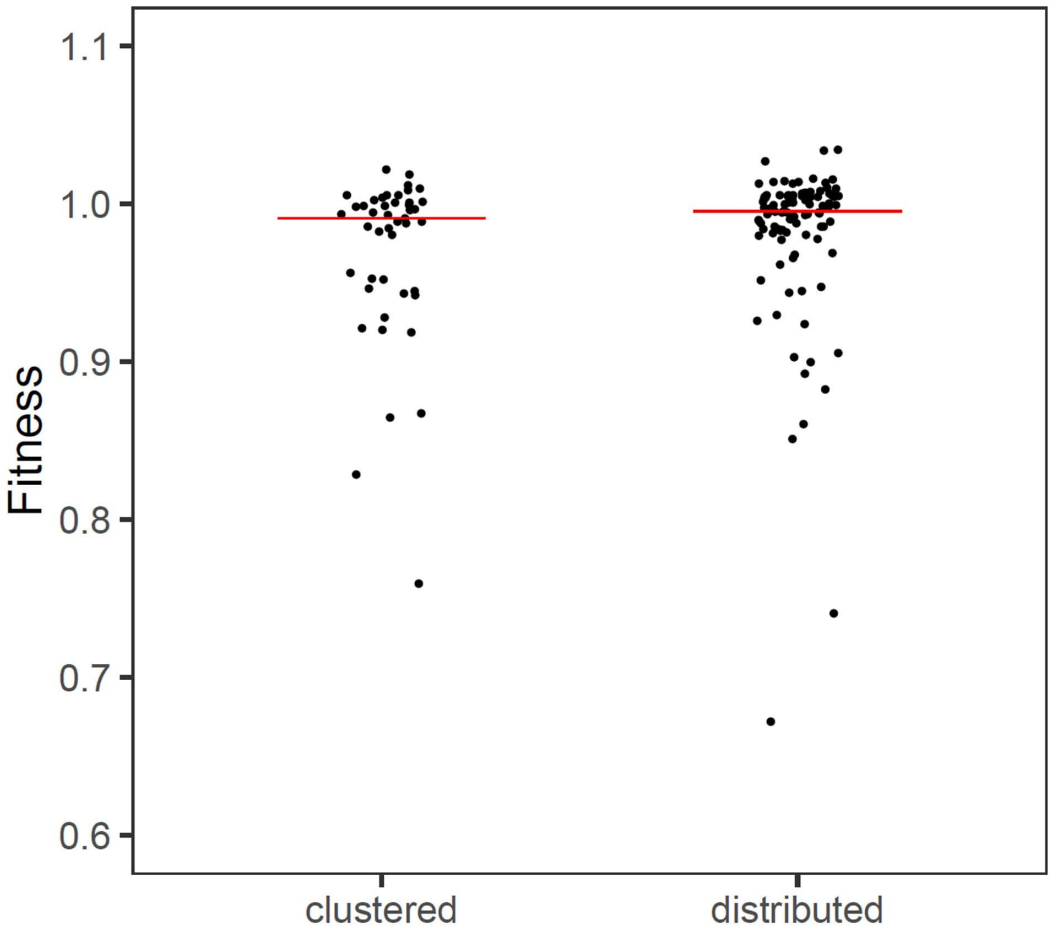


Fig. S1. Fitness importance is comparable between the clustered responsive genes and distributed responsive genes defined in BY4741(Δhap4) (*P* = 0.10, Mann-Whitney U-test). Fitness importance of a gene is measured by the relative growth rate of the gene deletion line to wild-type. The horizontal line shows the median.

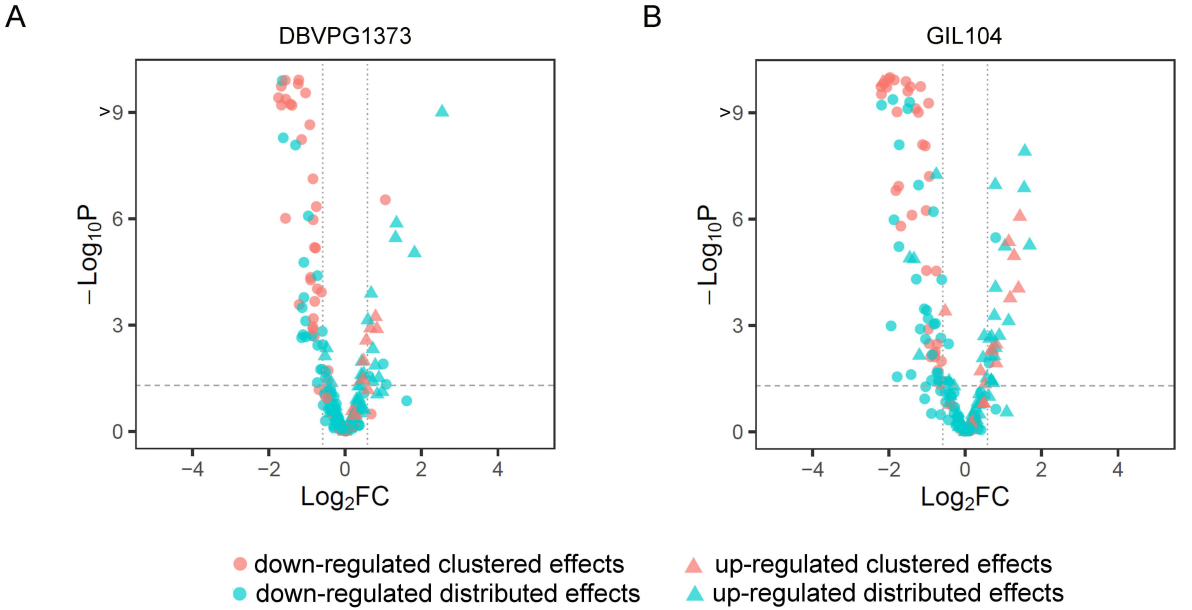


Fig. S2. The intra-species conservation analysis of the HAP4 deletion effects. The 195 responsive genes defined in BY4741(Δhap4) are examined with respect to their expression responses in *S. cerevisiae* DBVPG1373(Δhap4) and GIL104(Δhap4), respectively. The horizontal dashed line shows adjusted P = 0.05 and vertical dashed lines show log2FC = ±0.58. (cyan: clustered effects; red: distributed effects; cycle: down-regulated in BY4741(Δhap4); triangle: up-regulated in BY4741(Δhap4)).

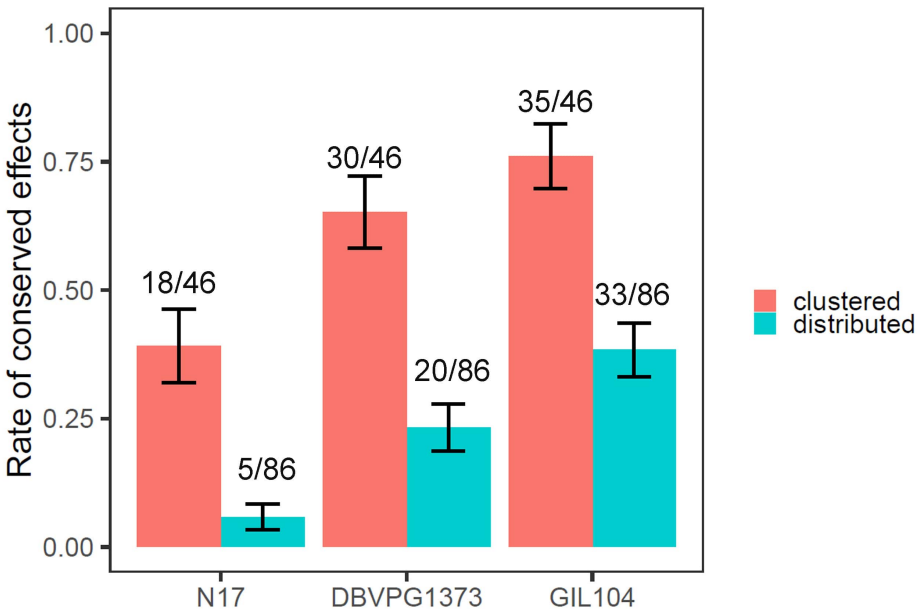


Fig. S3. Conservation analysis of the HAP4 deletion effects by considering only genes with a strong expression level in wild-type BY4741. This analysis is to address the concern that lowly expressed genes in wild-type tend not to have detectable down-regulation due to technical bias. Hence, for the 195 responsive genes defined in BY4741(Δhap4) only those with log2RPKM > 5 in wild-type BY4741 are considered here, leaving 46 clustered effects and 86 distributed effects. Error bars represent SE.

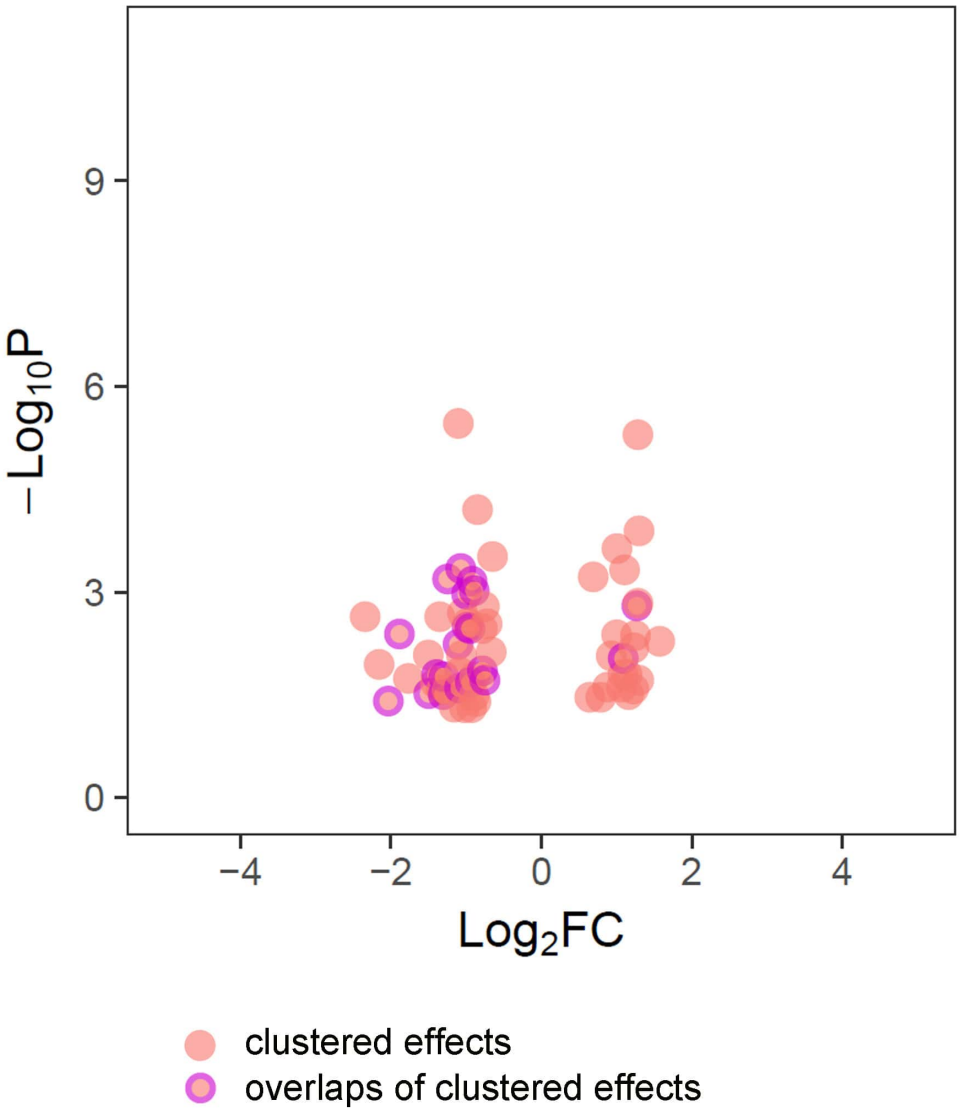
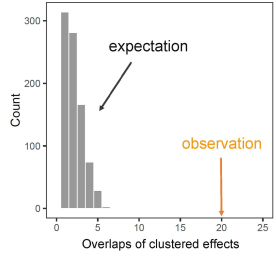


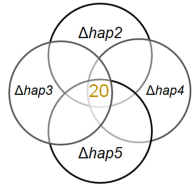
Fig. S4. The 20 overlapped clustered effects are comparable to the rest 45 (65-20) clustered effects defined in BY4741(Δhap4) with regard to their effect size. The differences are not statistically significant for both the P-values and fold changes observed in BY4741(Δhap4) (P = 0.75 and 0.51, respectively, Mann-Whitney U-test)

A

Deleted gene	Number of clustered effects	Number of distributed effects
HAP2	46	58
HAP3	140	316
HAP4	70	125
HAP5	53	108

B





C

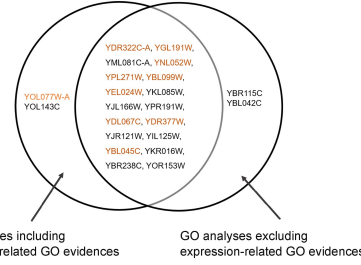
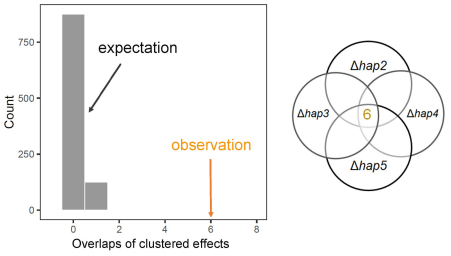


Fig. S5. The strong overlaps of clustered effects of the HAP2/3/4/5 complex genes are not biased by expression-related GO evidences for annotating the deletion effects. (A) Only the evidences IDA, HDA and IPI are used to re-define the clustered and distributed deletion effects of the four genes, respectively. (B) There are 20 overlapped clustered effects for the four genes encoding the HAP2/3/4/5 tetramer, which is significantly higher than expectation. The expectation is estimated by random sampling of the distributed effects of the four genes to calculate overlaps, and 1,000 such simulations were conducted. (C) The overlapped clustered effects are largely the same before and after excluding expression-related GO evidences. Genes that are the direct target of HAP4 are highlighted in yellow.

A

Deleted gene	Number of clustered effects	Number of distributed effects
HAP2	130	173
HAP3	33	41
HAP4	81	119
HAP5	12	39

B



C

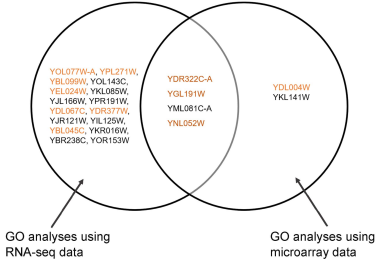


Fig. S6. The enrichment of overlapped clustered effects in HAP2/3/4/5 complex is reproduced by using public microarray data of the four gene deletion lines. (A) Microarray-based expression data are used to re-define the clustered and distributed deletion effects of the four genes, respectively. (B) There are only six overlapped clustered effects for the four genes encoding the HAP2/3/4/5 tetramer, which is significantly higher than expectation. The reduced number is primarily due to the small number (12) of clustered effects observed in HAP5 deletion. The expectation is estimated by random sampling of the distributed effects of the four genes to calculate overlaps, and 1,000 such simulations were conducted. (C) Comparison of the six overlapped clustered effects defined using microarray data with the 20 overlapped clustered effects defined using RNA-seq data. Genes that are the direct target of HAP4 are highlighted in yellow.

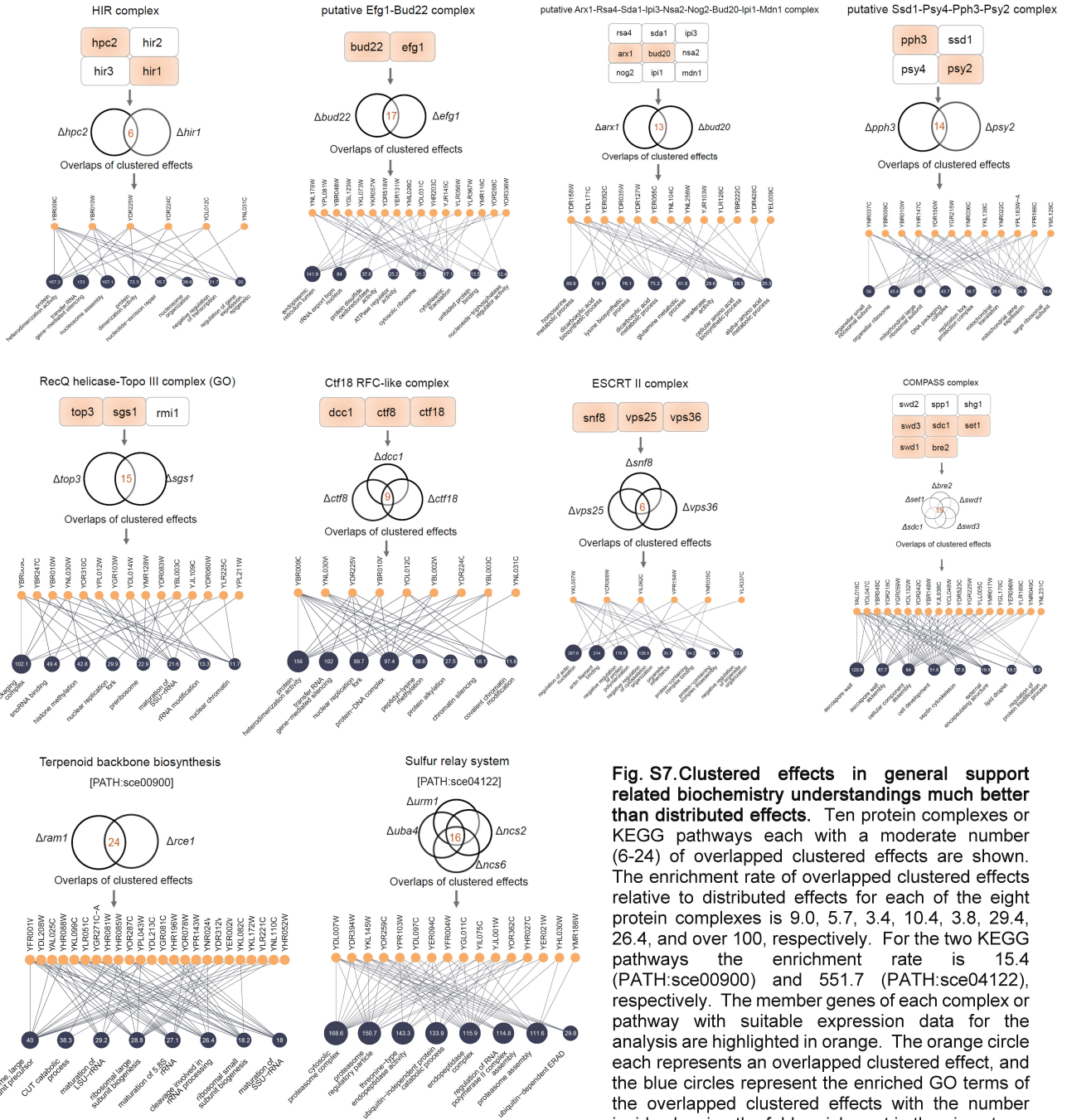


Fig. S7. Clustered effects in general support related biochemistry understandings much better than distributed effects. Ten protein complexes or KEGG pathways each with a moderate number (6-24) of overlapped clustered effects are shown. The enrichment rate of overlapped clustered effects relative to distributed effects for each of the eight protein complexes is 9.0, 5.7, 3.4, 10.4, 3.8, 29.4, 26.4, and over 100, respectively. For the two KEGG pathways the enrichment rate is 15.4 (PATH:sce00900) and 551.7 (PATH:sce04122), respectively. The member genes of each complex or pathway with suitable expression data for the analysis are highlighted in orange. The orange circle each represents an overlapped clustered effect, and the blue circles represent the enriched GO terms of the overlapped clustered effects with the number inside showing the fold enrichment in the given term.

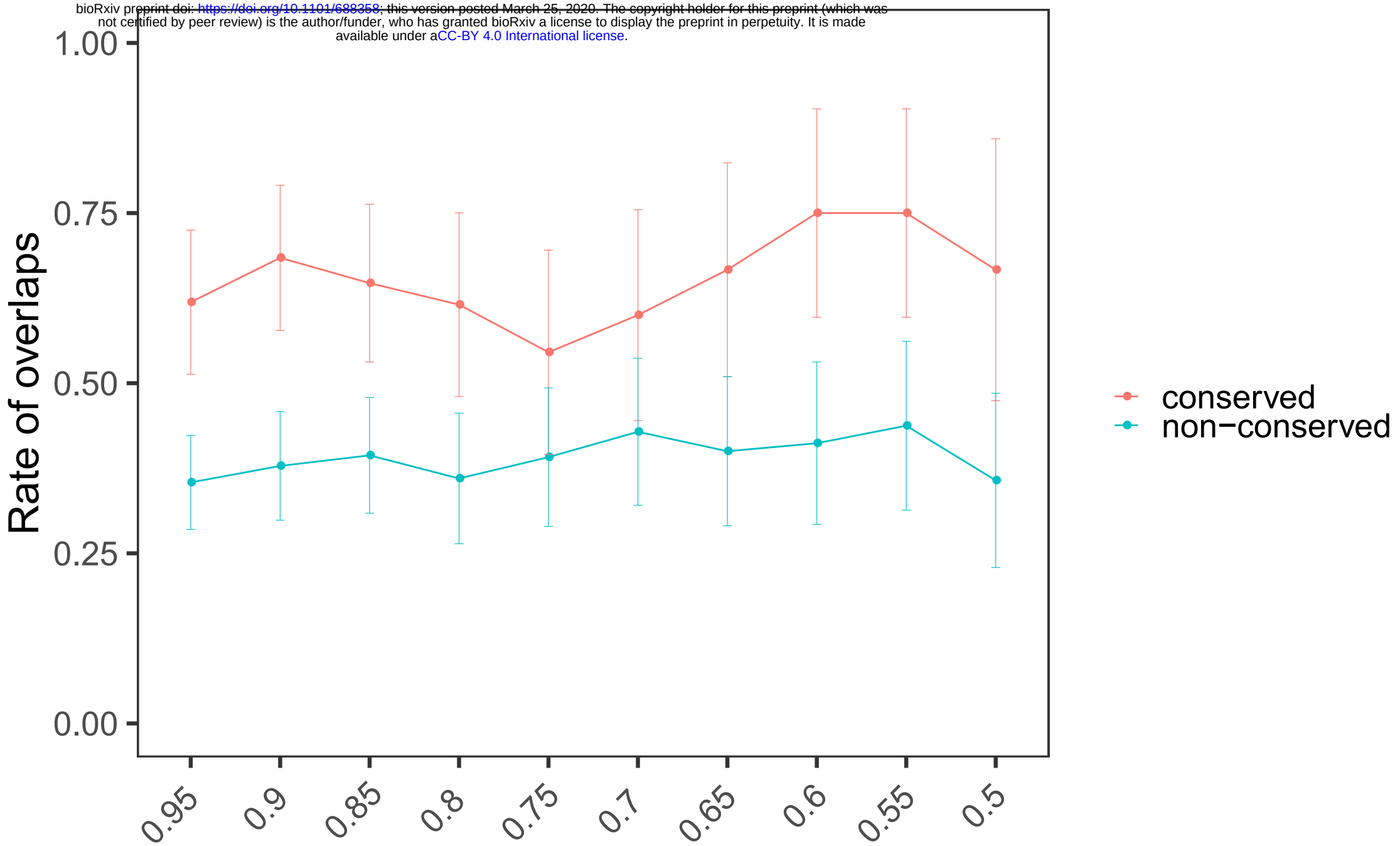


Fig. S8. The estimated rate of overlaps cannot be explained by correlated traits. The Pearson's R of all trait pairs is calculated using the trait values generated in ref. 4 for 4,718 yeast mutants. We then removed traits one by one from those with the highest absolute R until no two traits have R^2 greater than a threshold, which is set to be 0.95, 0.9, 0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55, and 0.5, respectively. The number of remaining traits are 346, 312, 277, 247, 223, 204, 185, 161, 153, and 127, respectively. Error bars represent SE.

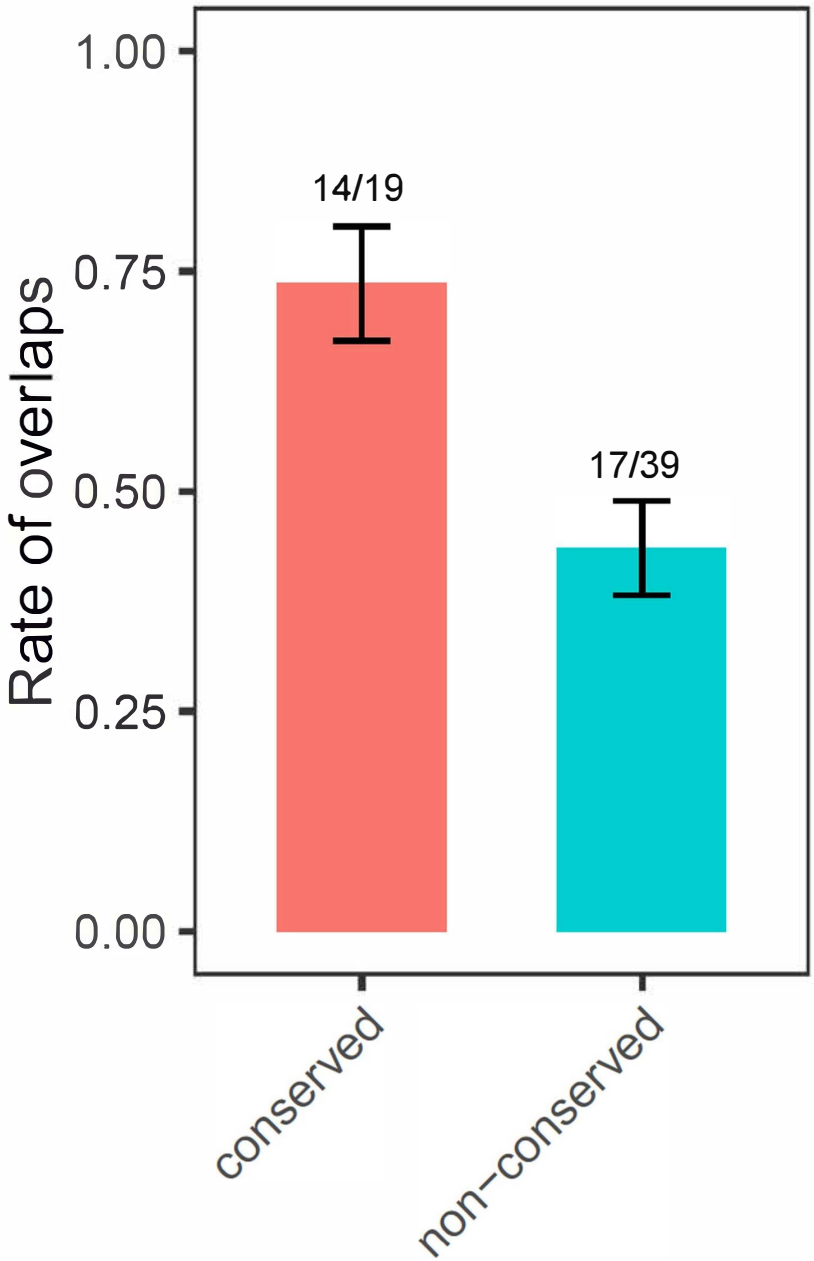


Fig. S9. The comparison in Fig. 4B is robust against trait measuring noise. To address the potential technical bias that traits with large measuring noise tend to be both non-conserved and non-overlapping, only traits with measuring CV < 0.1 across the replicates in wild-type BY4741 are considered. This results in 58 traits that are significantly affected by HAP4 deletion in S. cerevisiae, among which 19 are conserved effects and 39 non-conserved effects. The rate of overlaps in the conserved set remains significantly higher than the non-conserved set ($P = 0.029$, one-tailed Fisher's exact test). Overlaps refer to traits significantly affected by all four gene deletions in *S. cerevisiae*.