October 18th, 2019

Evolutionary decomposition of genetic effects

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

The view of evolution-free biochemical functionalities defined in the Encyclopedia of DNA Elements (ENCODE) project is challenged immediately after its emergence. Ironically, genetic effects defined in reverse genetics are also evolution-free, but this century-old fact has received little attention because of no readily available measure of the underlying evolutionary constraint. In this study we conducted systematic tests for selection on a set of pleiotropic genetic effects by assessing their within-population heritability and inter/intra-species conservation. We found only some of the effects are subject to effective selection, and the rest represent evolutionarily *ad hoc* gene-trait interactions. Interestingly, the selected portion supports well related biochemistry understandings while the *ad hoc* portion accounts for cryptic pleiotropy. We argue that, given the ENCODE consortium already abandoned the evolution-free view on biochemical functionalities, it is time for geneticists to develop an evolution-included framework for genetic effects.

Introduction

Mutation analysis has long been used to understand the normal function 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 were primarily from mechanistic perspectives[5,6], by considering the focal gene's multiple

55 molecular functions or multiple cellular processes associated with a single molecular
56 function[7]. However, since extant species are all evolutionary products with history,
57 mechanistic perspectives alone are not sufficient to delineate a biological phenomenon[8].
58 This is exemplified by the debates on the ENCODE project in which up to 80% of the
59 human genome was claimed to be functional despite the fact that only 10% is under
60 selection[9-16]. A simple example explains how the confusions arise. Suppose there is
61 a transcription factor (TF) that recognize a DNA motif, say, ATCGATCG. The human
62 genome with $3x10^9$ base pairs in length contains over one hundred thousand
63 ATCGATCG motif sequences, some of which are evolutionarily selected for certain
64 biological processes while the rest exist as *ad hoc* entities due to the equilibrium of
65 random mutations in such a long genome[12]. From an evolution-free perspective
66 originally adopted by the ENCODE consortium[9], the myriad ATCGATCG sequences
67 are all called functional so long as they are bound by the TF. However, the claim of
68 *ad hoc* entities as functional would only confuse our understandings in how the system
69 is organized to function. The resulting confusion forced the ENCODE consortium to
70 eventually abandon their evolution-free view on biochemical functionality[16].

71 Notably, the same problem actually also applies to genetic causalities defined in
72 reverse genetics. The common practice in reverse genetics is to knockout or
73 knockdown a gene and find the traits significantly altered[17]. The revealed gene-trait
74 interactions are evolution-free, representing solely mechanistic understandings. In the
75 above TF versus ATCGATCG motif example, when the TF is deleted, the expression of
76 those genes with the motif at promoter region could all be affected. The resulting

77 pleiotropic genetic effects, which are either *ad hoc* or evolutionarily selected according
78 to the nature of the focal motifs, together would lead to a very complex picture on the
79 functionality of the TF. Since even the ENCODE consortium, a community that is
80 heavily driven by mechanism but not history, is now willing to adopt an evolution-
81 included view on biochemical functionalities, isn't it the time for geneticists to abandon
82 the evolution-free view on genetic effects? Toward this goal we here tested the
83 evolutionary constraint on a set of genetic effects, revealing novel insights.
84

85 **Results**

86 We started with a known yeast *Saccharomyces cerevisiae* gene HAP4[18]. It is a
87 non-essential transcription factor that has been subjected to extensive studies since its
88 discovery 30 years ago[19]. We deleted the open reading frame of HAP4 in *S. cerevisiae*
89 strain BY4741, and checked the expression trait of the other yeast genes by sequencing
90 the transcriptome of the strains that grow in the rich medium YPD at 30 °C (Table S1).
91 We found 195 responsive genes each with a significant expression change under a
92 stringent statistical cutoff (Table S2). Gene ontology (GO) analysis of the 195 genes
93 revealed one third of them (65/195) clustered more than expected by chance in dozens
94 of GO terms. These GO terms are related with each other and reflect well the
95 functional annotations of HAP4 as a regulator of mitochondria activities[20] (Fig. 1a).
96 The remaining two thirds (130/195) distribute rather randomly across diverse biological
97 processes, underscoring the strong pleiotropy of HAP4. The two sets of genes are all
98 functionally characterized (Table S2), with comparable fitness importance (P = 0.1,

99 Mann-Whitney U-test; Fig. S1). Notably, the 65 clustered deletion effects and 130
100 distributed deletion effects are supported by similar P-values and fold changes (P =
101 0.20 and 0.46, respectively, Mann-Whitney U-test; Fig. 1b).

102 Genetic effects observed in an isogenic background should be heritable in
103 population if they are selected in evolution. We crossed two *S. cerevisiae* strains to
104 obtain a population of yeast segregants. Specifically, a wild-type strain
105 BY4742(MATapha), which is identical to BY4741 except at the mating locus, was
106 crossed with the HAP4 deletion line of DBVPG1373(MATa) (Fig. 1c). This way, the
107 comparison between wild-type and null alleles at the HAP4 locus would match the
108 comparison conducted in the isogenic BY4741 background. We dissected six tetrads
109 of the hybrid and obtained 12 HAP4 wild-type and 12 HAP4 null segregants. For each
110 of the 195 deletion effects we defined a heritability score HS = Var(wt, null) / Var(total),
111 where Var(wt, null) is the variance of the focal gene's expression level between the two
112 segregant groups, and Var(total) is the total variance across the segregants. This
113 design resembles a forward genetic assay with a candidate genetic locus. The HS is
114 equivalent to the narrow-sense heritability of an expression trait with HAP4 as the
115 candidate quantitative trait locus (i.e., h^2_{QTL}). An HS close to zero suggests HAP4 not
116 be a QTL of the expression trait. We found the 65 clustered effects in general have
117 much higher HS than the 130 distributed effects ($P = 1.1\times10^{-3}$, Mann-Whitney U-test;
118 Fig. 1d). Approximately 26.2% (17/65) of the clustered effects have a statistically
119 significant HS, while the number is 0.8% (1/130) for the distributed effects ($P = 2.1\times10^{-}$
120 ^{8}, Fisher's exact test; Fig. 1e).

121 We further tested the evolutionary conservation of the genetic effects. We
122 obtained, under the same environment and the same statistical cutoff as in BY4741, the
123 deletion effects of HAP4 in *Saccharomyces paradoxus* strain N17, a closely related
124 yeast species diverged from *S. cerevisiae* ~10 million years ago[21] (Fig. 1f). Only 5.4%
125 (7/130) of the distributed effects found in *S. cerevisiae* were also observed in the HAP4
126 deletion line of *S. paradoxus* N17, while the number was 27.7% (18/65) for the
127 clustered effects ($P = 3.1\times10^{-5}$, Chi-square test). The difference is robust as evidenced
128 by plotting the expression responses of the 195 genes in N17(Δhap4) (Fig. 1g). We
129 also examined intra-species conservation of the HAP4 deletion effects by looking at
130 two other *S. cerevisiae* strains DBVPG1373 and GIL104, the former of which is 0.35%
131 diverged from the strain BY4741 and the latter 0.07% diverged at the genomic level[22].
132 Similar1y, the 130 distributed effects were largely unreproducible in strains
133 DBVPG1373(Δhap4) and GIL104(Δhap4), while the 65 clustered effects were much
134 more conserved ($P = 0.5\times10^{-4}$ for DBVPG1373(Δhap4) and $P = 2.6\times10^{-5}$ for
135 GIL104(Δhap4), Chi-square test; Fig. S2). As expected, both types of the deletion
136 effects are more reproducible in more related yeasts (Fig. 1h). We confirmed the
137 results cannot be explained by different detectability of expression changes between the
138 two gene sets by excluding those genes lowly expressed in wild-type BY4741 (Fig. S3).
139 These data, together with the heritability analysis, suggest the clustered effects of HAP4
140 tend to be selected in evolution; on the contrary, the distributed genetic effects appear
141 non-evolutionary, likely representing *ad hoc* responses to the gene deletion[23].
142 HAP4 functions by forming a tetramer with HAP2, HAP3 and HAP5 as a result of

143 evolution[19]. We hypothesized clustered genetic effects should support this
144 biochemistry understanding better than distributed genetic effects, because the latter is
145 non-evolutionary. To test the hypothesis, we deleted the other three genes that encode
146 the tetramer, respectively, in *S. cerevisiae* BY4741, and measured the expression
147 profiles of the deletion lines. We defined clustered effects and distributed effects for
148 each of the lines using the same method as in the HAP4 deletion line. We obtained
149 43, 150 and 50 clustered effects, and 61, 306 and 111 distributed effects for the deletions
150 of HAP2, HAP3 and HAP5, respectively (Fig. 2a-c; Table S3). Consistent with the
151 hypothesis, we found 20 overlapped clustered effects across the four gene deletion lines,
152 14.5 times higher than that of the distributed effects ($P < 0.001$, simulation test, Fig.
153 2d). Notably, the 20 overlapped clustered effects are not the strongest in
154 BY4741(Δhap4) (Fig. S4). To avoid the potential bias that expression responses to
155 the tetramer may have been considered in the GO annotations of the responsive genes,
156 we excluded all expression-related evidences for GO annotations to define new
157 clustered and distributed effects. We obtained essentially the same result (Fig. S5).
158 Because there are publicly available microarray data for HAP2, HAP3, HAP4 and
159 HPA5 deletion lines[24], we also repeated the analysis using the public expression data
160 and observed a similar pattern (Fig. S6).

161 In addition to considering the protein complex formed by HAP4, we could also
162 consider protein-DNA interactions since HAP4 is a transcription factor. Data from a
163 chromatin immuno-precipitation (CHIP) assay of the promoters bound by HAP4 show
164 that, among the 195 responsive genes observed in BY4741(Δhap4), 13 are direct targets

165 of HAP4 (Fig. 2e)[25,26]. Interestingly, there is 24-fold enrichment of direct targets in
166 the clustered effects relative to the distributed effects ($P = 8.6\times10^{-6}$, Fisher's exact test);
167 among the 20 overlapped clustered effects 50% (10/20) are direct targets of HAP4,
168 while the genomic background is 0.64% (33/5146) ($P = 4.6\times10^{-18}$, Hypergeometric test)
169 (Fig. 2f). Hence, the CHIP data well support the distinction of the two effect types.

170 Collectively, these results are consistent with a recently proposed model[23] (Box 1),
171 in which the null phenotype of a gene can be ascribed to either the loss of the gene's
172 native function, or the gain of spurious function that arises from passive adjustments
173 within the cellular system after the genetic perturbation. The key difference of the two
174 function types is their evolutionary effectiveness: native functions are historical and
175 evolutionarily selected, while spurious functions are ahistorical and *ad hoc*[27-29].
176 Accordingly, the distributed effects examined here likely represent spurious functions
177 created by the HAP4 deletion, and the clustered effects could be in a large part ascribed
178 to the native functions of HAP4. It is intriguing how the two effect types defined by
179 GO could fit the two function types described in the model. We reason that
180 evolutionarily optimized native functions are likely to regulate specific pathways or
181 processes; losing them would thus cause coordinating changes of the related genes[30],
182 which are detected by GO analysis. In contrast, spurious functions may affect the
183 transcriptome in a rather random way, resulting in distributed changes across diverse
184 cellular processes, most of which cannot be covered by overrepresented GO terms.
185 This may explain why GO clustering here could echo evolutionary effectiveness.

186 Regardless of the underlying logic, clustered genetic effects seem to be well

187 matched with related biochemistry understandings. This would help resolve a long-
188 standing challenge to molecular biology - the gap between genetic analysis and
189 biochemistry analysis[17,31]; specifically, genes with intimate biochemical interactions do
190 not have common genetic effects and genes with common genetic effects do not show
191 intimate biochemical interactions[32,33]. To test the generality of the finding that was
192 based on the HAP2/3/4/5 tetramer, we examined other biochemically characterized
193 protein complexes by using publicly available expression data. To avoid bias we
194 considered a single dataset comprising microarray-based expression profiles of over
195 one thousand yeast gene deletion lines[24]. There are 54 protein complexes annotated
196 by a previous study suitable for our analysis[34] . In 24 cases the overlaps of clustered
197 effects are significantly more than what would be expected from distributed effects at
198 a 99% confidence level, and the enrichments range from 2.7-fold to over 100-fold with
199 a median 5.3-fold (Fig. 3a and Table S4). The overlapped clustered effects of each
200 protein complex represent specific functions (Fig. 3b-c and Fig. S7). For example, the
201 ten overlapped clustered effects of the elongator holoenzyme complex are tens to
202 hundreds times overrepresented in a few transcription-related GO terms as well as
203 proteasome-related GO terms (Fig. 3b), the former of which echo well the annotated
204 functions of the complex while the latter appear to suggest new understandings[35]. As
205 another example, the genes encoding the protein kinase CK2 complex have 21
206 overlapped clustered effects that appear to affect specifically the metabolism of various
207 amino acids (Fig. 3c), a functional insight not been well recognized[36].

208 We also checked genes on the same KEGG pathways. There are 41 pathways that

209 are related to metabolism, genetic information processing, cellular processes, and so on,
210 suitable for our analysis (Table S5). The rate of overlaps of clustered effects is
211 significant higher than that of distributed effects in nine cases, and the enrichments
212 range from 5.6-fold to over 100-fold with a median 46.9-fold (Fig. 3d and Table S5).
213 Consistently, the overlapped clustered effects of each pathway represent distinct
214 functions (Fig. 3e-f and Fig. S7). For the many cases in which clustered effects show
215 no more overlaps than distributed effects, the involved genes may execute additional
216 functions irrelevant to the focal complex or pathway. Notably, in none of the cases
217 distributed effects represent related biochemistry understandings better than clustered
218 effects, highlighting the cryptic nature of them. Taken together, focusing on clustered
219 effects appears to be a readily operational approach to narrowing the gap between
220 genetic analysis and biochemistry data.

221 The above analyses considered gene expression traits. We next examined the
222 yeast cell morphological traits that are based on the microscopic images of cells stained
223 by fluorescent dyes[37]. With the help of a computer software as many as 405
224 quantitative traits can be obtained from cell wall and nuclear stained cell images[38].
225 These traits are typically about area, distance, and angle calculated based on dozens of
226 coordinate points, lines and angles that describe the shape of mother cell and bud, and
227 the shape and localization of the nuclei in mother cell and bud (Fig. 4a). This large
228 set of yeast traits had served as a valuable resource for studying genotype-phenotype
229 relationships[30,39,40]. Deletion of HAP4 in *S. cerevisiae* significantly altered 78
230 morphological traits, among which 24 are also significantly affected in *S. paradoxus* by

231 HAP4 deletion (Table S6). To test if the evolutionarily conserved effects of HAP4 are
232 shared with HAP2, HAP3 and HAP5 more than the non-conserved effects, we also
233 measured the morphological traits affected by each of the other three genes, respectively,
234 in *S. cerevisiae*. We found that 58.3% (14/24) of the conserved effects are shared with
235 all the other three genes, which is significantly higher than the number (18/54 = 33.3%)
236 for the non-conserved effects of HAP4 (P = 0.035, Fisher's exact test; Fig. 4b). The
237 estimations are not explained by correlated traits (Fig. S8), and the difference remains
238 largely unchanged when only traits with small measuring noise are considered (Fig. S9).
239 Hence, the cell morphology data also support the role of evolution in separating genetic
240 effects.
241

242 **Discussion**

243 Thanks to the mature framework of measuring the selective constraints on DNA
244 sequence[41], the evolution-free functionality of DNA elements defined in ENCODE was
245 challenged immediately after its emergence[12,13]. Notably, the genetic-trait interactions
246 defined in reverse genetics are also based on an evolution-free framework. However,
247 this century-old problem has ever attracted little attention[1], primarily due to the lack of
248 a readily available measure of the underlying selective force. In this study we, for the
249 first time to the best of our knowledge, performed a rigorous test of the evolutionary
250 effectiveness of a set of genetic effects by examining their evolutionary conservation
251 as well as within-population heritability. We found only some of them subject to
252 effective selection, and the rest represent evolutionarily *ad hoc* gene-trait interactions.

253 Similar to the *ad hoc* "functional" DNA elements defined in ENCODE[9], the *ad hoc*
254 genetic interactions are presumably explained by mutation equilibrium or spurious
255 functions arising from the focal genetic perturbation[23]. Such *ad hoc* interactions,
256 serving as the raw materials of evolution, would be selected for or against in the future.
257 Importantly, since they have not yet been shaped by selection, they tend not to be
258 compatible/coordinated with the pre-existing functions[30]. Hence, including them
259 would likely blur the role the focal gene has long played in the past. This helps
260 understand, at least partly, the origin of pleiotropic effects.

261 Conceptually speaking, our evolution-included view on genetic effects is a natural
262 extension of the evolution-included view on the functionality of DNA elements[12,13].
263 Hence, pros and cons that have widely discussed in the debates on the ENCODE project
264 apply similarly to this study. For example, because detecting selection involves
265 multiple lineages, one cannot rule out the possible that an entity under no detectable
266 selection is actually subject to lineage-specific selection[12]. However, since the
267 lineages examined are often closely related, lineage-specific entities selected in a short
268 time window should be rare compared to those acquired during the long time period
269 predating the split of the lineages. Operationally speaking, the evolution-included
270 view on the functionality of DNA elements relies on DNA sequence comparison, which
271 is straightforward and now mature. However, an evolutionary separation of genetic
272 effects requires rather complex experimental designs; also, there is no available
273 framework for modeling the turnover rate of gene-trait interactions under no selection.
274 Hence, we could, as in this study, only perform enrichment analysis for a group of gene-

trait interactions. Nevertheless, the current limitation in operationality does not challenge the validity of the concept. A particularly interesting finding of this study is GO clustering can serve as a useful and readily operational proxy for selection when expression traits are examined. The underlying rationale, namely, functional coordination built only by selection, may help design a more efficient strategy for delineating the evolutionary nature of genetic effects.

In summary, by re-assessing genetic causality in the light of evolution this study suggests an expanded framework for reverse genetics (Fig. 4c). Specifically, the conventional framework is evolution-free, relying solely on statistical tests to separate the mutant versus wild-type differences into significant and insignificant effects. In the expanded framework evolutionary effectiveness is included to further separate significant effects into evolutionarily selected and *ad hoc* ones, the former of which would support related biochemistry understandings while the latter account for cryptic pleiotropy.

Materials and Methods

Yeast gene deletions

Three *S. cerevisiae* (SC) strains BY4741 (*MATa, his3, leu2, met15, ura3*), GIL104 (*MATa, URA3, leu2, trp1, CAN1, ade2, his3, bar1Δ::ADE2*; derived from the W303) and DBVPG1373 (*MATa, ura3*), and one S. paradoxus (SP) strain N17 (*MATa, ura3*) were included in the study. Unless otherwise stated, the *S. cerevisiae* strains were cultured in the rich medium YPD (1%Yeast extract, 1%Peptone, 2% Dextrose) at 30°C, and *S. paradoxus* N17 was cultured in YPD at 25°C. The wild type URA3 in GIL104 was first replaced by a LEU2 cassette. HAP4 was replaced by a URA3 cassette in each of the four strains. HAP2, HAP3, and HAP5 were also replaced, respectively, by a URA3 cassette in BY4741. Notably, for all gene replacements the whole open reading frame from the starting codon to the stop codon of a focal gene was replaced. As described in our previous study[42], the standard LiAc transformation method[43] was used to transform DNA into the yeast cells and gene replacements were achieved by

homologous recombination. The transformation protocol was slightly modified for *S. paradoxus*[44]; specifically, heat shock was performed for seven minutes at 37°C. Synthetic medium deprived of uracil or leucine was used to select the clones with successful replacement for the target gene. All gene replacements were verified using polymerase chain reaction (PCR). For each gene deletion line, 3-5 independent clones 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[45]. 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 obt
ained. 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 (MATapha), 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+ Δhap4, 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 200uL saturated culture was added into 10mL fresh YPD, which resulted in an optical density OD600 ~ 0.1 (UNICO

348 UV/VIS Spectrophotometer), and cells of 3mL culture at OD600 = 0.5-0.65 were
349 harvested. Total RNA was extracted by QIAGEN RNeasy Plus mini kit (Cat
350 No.74136). The mRNA sequencing was performed using the paired-end module on a
351 HiSeq platform at Genwiz by following the standard procedure. To ensure the high
352 quality of expression analysis, we sequenced the mRNA of 3-6 independent clones for
353 each wild-type or gene deletion line.
354 RNA-seq reads were mapped to reference yeast genomes using STAR (Version
355 2.6.0c)[46]. For BY4741 and GIL104, we used the genome of *S. cerevisiae* strain S288C
356 as the reference (version R64-2-1_20150113; http://www.yeastgenome.org). The
357 reference genomes of SC-DBVPG1373 and SP-N17 were downloaded from SGRP
358 (https://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html). For a
359 typical clone there were about 6.5 million paired-end reads mapped to the coding
360 sequences. Gene expression levels were determined by Featurecounts (version
361 1.6.2)[47] with default settings and RPKM (reads per kilobase per million) of each gene
362 were calculated by R package edgeR. The wild-type versus mutant differential
363 expression analysis was performed by DESeq2[48] with default parameters, and genes
364 with an adjusted P-value smaller than 0.05 and a fold change (FC) greater than 1.5 were
365 defined as significantly changed genes. In the conservation, heritability or
366 overlapping analysis, an effect is called conserved, heritable or overlapped only when
367 it shows the same direction in the various conditions examined. Genes of the uracil
368 biosynthesis pathway (YBL039C, YEL021W, YJL130C, YJR103W, YKL024C,
369 YKL216W, YLR420W, YML106W, YMR271C) were excluded from further analysis.
370 The expression level measured by RPKM of HAP2, HAP3, HAP4 and HAP5 in wild-
371 type BY4741 is 84.4, 64, 194 and 90.5, respectively, and all becomes zero in the
372 corresponding deletion lines. The fitness of yeast gene deletion lines was produced
373 by a previous study[49]. Table S1 contains details of the RNA-seq expression
374 information of all yeast lines examined in this study.
375
376 **GO analysis**
377 The GO analyses of the responsive genes derived from our RNA-seq data were
378 conducted in the SGD website using GO Term Finder (Version 0.86;
379 https://www.yeastgenome.org/goTermFinder), by excluding computational analysis
380 evidences and other less reliable evidences: IBA, IC, IEA, IKR, IRD, ISA, ISM, ISO,
381 ISS, NAS, ND, TAS. In a strict analysis which required the exclusion of all
382 expression-related evidences, only three GO evidence codes IDA, HDA and IPI that
383 represent direct experimental assays were considered. For the GO analyses of public
384 microarray data the R package clusterProfiler[50] was used with default settings. The
385 cutoffs used to define an enriched GO term include an adjusted P-value smaller than
386 0.01 and a fold enrichment greater than 2. To improve specificity only GO terms
387 containing less than 200 genes were considered. The fold enrichment was calculated
388 as (number of changed genes in the GO term / number of all changed genes) / (number
389 of genes in the GO term / number of genes in all GO terms of the class). The GO
390 semantic similarity scores were calculated by R package GOSemSim[51].
391

392 **Heritability analysis**
393 To estimate the heritability of the 195 HAP4 deletion effects, a population of 24
394 segregants were divided into two groups: 12 with the wild-type allele of HAP4 and 12
395 with the null allele of HAP4. The heritability score HS was the variation between the
396 two groups divided by the total variation of the population. To test the statistical
397 significant of an HS, we compared the expression levels of the focal gene between the
398 two segregant groups using DESeq2. The obtained 195 raw P-values were adjusted
399 for multiple testing using the Benjamini-Hochberg controlling procedure. An
400 adjusted P-value smaller than 0.05 was considered significant.
401
402 **Analyze protein complexes and metabolic pathways**
403 The public microarray data of ~1,400 yeast gene deletion lines were obtained from a
404 previous study[24], and P-values and fold changes (FC) provided in the data were directly
405 used. Specifically, $P < 0.05$ and absolute $FC > 1.2$ were used to define genes with
406 significant expression changes; if the number of significantly changed genes was over
407 1,500, a more stringent cutoff $P < 0.01$ was used. To avoid the effects of genes with
408 ubiquitous expression responses we excluded from further analyses the top 10% genes
409 that each show significant changes in at least 12% of the gene deletion lines. GO
410 analyses were performed by R package clusterProfiler to define clustered and
411 distributed effects for each deletion line, with the results summarized in Table S7. To
412 examine the overlapped clustered effects between genes of the same protein complex
413 or pathway, we only considered the deletion lines with at least 20 clustered effects,
414 resulting in a set of 422 deletion lines suitable for further analyses.
415 Information of 518 protein complexes was obtained from a previous study [34]. The
416 KEGG pathways of the yeast *S. cerevisiae* were downloaded from KEGG website
417 (https://www.genome.jp/kegg-bin/get_htext?sce00001). There are 54 complexes and
418 41 pathways each with at least two member genes found in the above defined mutant
419 set.
420 For each protein complex or pathway, the overlaps of clustered effects and the
421 overlaps of distributed effects were compared in number. The numbers of clustered
422 effects and distributed effects of the involved genes were normalized to make the
423 overlaps between the two effect types comparable. To estimate the confidence
424 interval of a comparison we used random samplings. If clustered effects are less than
425 distributed effects in all genes, which is true in most of the cases examined, we sampled
426 (without replacements) a random subset of distributed effects to ensure the two effect
427 types of a gene equal in number. If clustered effects are more than distributed effects
428 in all genes, we sampled (without replacements) a random subset of distributed effects
429 to ensure the two effect types of a gene equal in number. If the above consistent
430 patterns do not exist, we sampled consistently from one side (either clustered effects or
431 distributed effects) but with replacements for the gene with an insufficient number of
432 effects on this side. For each complex or pathway 1,000 such random samplings were
433 carried out to derive the 99% confidence interval, and an observed difference is called
434 significant if it is not within the interval. Table S4 and Table S5 have details about the
435 protein complexes and KEGG pathways examined, respectively.

436

437 **Analyze cell morphological traits**

438 Diploid yeast cells were examined by following the protocol of previous studies with
439 slight modifications[37,38]. In brief, a single yeast colony was picked and cultured in
440 YPD liquid overnight with shaking to the saturation phase. Then, 1.5uL culture was
441 transferred to 100uL fresh YPD in a 96-well plate and grew for 3-4 hours at 30°C for
442 SC strains or 25°C for SP strains. Cells were fixed with 3.7% formaldehyde solution.
443 Cell wall was stained by FITC-ConA (fluorescein isothiocyanate-conjugated,
444 concanavalin A, Sigma-Aldrich C7642). Cell nucleus was stained by hochest-mix
445 (Thermo Fisher, Hoechst 33342 Solution) instead of DAPI to enhance the specificity.
446 We did not stain actin because the dye Rhodamine phalloidin was not stable enough to
447 support the following high-throughput automated image capturing which takes about
448 10 hours for scanning 96 wells of a plate. The stained cells were plated into a
449 microplate (Greiner 781091) with ~5.0×10^4 cells per well and images were captured by
450 IN Cell Analyzer 2200 (GE Healthcare) using the 60× objective lens.

451 Five SC lines (all diploid with BY4741 background: wild-type, Δhap2, Δhap3,
452 Δhap4 and Δhap5) and two SP lines (all diploid with N17 background: wild-type and
453 Δhap4) were examined. Because the trait measuring is quite sensitive to batch effect,
454 for each line we conducted 18-24 replicates of staining and image capturing. The
455 images were analyzed by CalMorph[37,38] with default settings, and only 405 rather than
456 501 traits were extracted in this study because actin is not stained. At least 1,000 cells
457 were captured and analyzed (with at least 100 informative cells for each cell-cycle stage)
458 for a high-quality replicate. In the end, there were 13~23 high-quality replicates for
459 each of the lines included in further analysis. Trait values were compared between
460 replicates of a gene deletion line and replicates of the corresponding wild-type line
461 using T-test, and the resulting 405 P-values were adjusted for multiple testing using the
462 Benjamini-Hochberg controlling procedure. Because of the many replicates included
463 in the comparison, many traits showed a statistically significant but biologically
464 negligible difference between wild-type and mutant lines. Hence, a trait is called
465 affected by a gene only when the adjusted $P < 0.05$ and the difference between wild-
466 type and mutant is large than 5%. Table S8 has complete information regarding the
467 morphological trait analysis.

468

469 **Acknowledgments**

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

473

474 **References**

475 1 Muller, H. J. Further studies on the nature and causes of gene mutations. *Proceedings of the*
476 *6th International Congress of Genetics*, 213-255 (1932).
477 2 Wang, Z., Liao, B. Y. & Zhang, J. Genomic patterns of pleiotropy and the evolution of complexity.
478 *Proc Natl Acad Sci U S A* **107**, 18034-18039, doi:10.1073/pnas.1004666107 (2010).
479 3 Caspari, E. J. E. PLEIOTROPIC GENE ACTION. **6**, 1-18 (1952).

480 4 Sonnichsen, B. *et al.* Full-genome RNAi profiling of early embryogenesis in Caenorhabditis
481 elegans. *Nature* **434**, 462-469, doi:10.1038/nature03353 (2005).
482 5 Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet* **29**, 66-73,
483 doi:10.1016/j.tig.2012.10.010 (2013).
484 6 Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767-773,
485 doi:10.1534/genetics.110.122549 (2010).
486 7 He, X. L. & Zhang, J. Z. Toward a molecular understanding of pleiotropy. *Genetics* **173**, 1885-
487 1891, doi:10.1534/genetics.106.060269 (2006).
488 8 Ayala, F. J. "Nothing in biology makes sense except in the light of evolution": Theodosius
489 Dobzhansky: 1900-1975. *J Hered* **68**, 3-10, doi:10.1093/oxfordjournals.jhered.a108767 (1977).
490 9 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature*
491 **489**, 57-74, doi:10.1038/nature11247 (2012).
492 10 Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Curr Biol* **22**, R898-899,
493 doi:10.1016/j.cub.2012.10.002 (2012).
494 11 Niu, D. K. & Jiang, L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem*
495 *Biophys Res Commun* **430**, 1340-1343, doi:10.1016/j.bbrc.2012.12.074 (2013).
496 12 Graur, D. *et al.* On the Immortality of Television Sets: "Function" in the Human Genome
497 According to the Evolution-Free Gospel of ENCODE. *Genome Biol Evol* **5**, 578-590,
498 doi:10.1093/gbe/evt028 (2013).
499 13 Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A* **110**, 5294-
500 5300, doi:10.1073/pnas.1221376110 (2013).
501 14 Kellis, M. *et al.* Reply to Brunet and Doolittle: Both selected effect and causal role elements can
502 influence human biology and disease. *Proc Natl Acad Sci U S A* **111**, E3366,
503 doi:10.1073/pnas.1410434111 (2014).
504 15 Brunet, T. D. & Doolittle, W. F. Getting "function" right. *Proc Natl Acad Sci U S A* **111**, E3365,
505 doi:10.1073/pnas.1409762111 (2014).
506 16 Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U*
507 *S A* **111**, 6131-6138, doi:10.1073/pnas.1318948111 (2014).
508 17 Griffiths, A. J. F. *An Introduction to genetic analysis*. 5th edn, (W.H. Freeman, 1993).
509 18 Forsburg, S. L. & Guarente, L. Identification and characterization of HAP4: a third component
510 of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev* **3**, 1166-1178 (1989).
511 19 Bolotin-Fukuhara, M. Thirty years of the HAP2/3/4/5 complex. *Biochim Biophys Acta Gene*
512 *Regul Mech* **1860**, 543-559, doi:10.1016/j.bbagrm.2016.10.011 (2017).
513 20 Buschlen, S. *et al.* The S. Cerevisiae HAP complex, a key regulator of mitochondrial function,
514 coordinates nuclear and mitochondrial gene expression. *Comparative and functional genomics*
515 **4**, 37-46, doi:10.1002/cfg.254 (2003).
516 21 Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of
517 yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254,
518 doi:10.1038/nature01644 (2003).
519 22 Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337-341,
520 doi:10.1038/nature07743 (2009).
521 23 He, X. The Biology Complicated by Genetic Analysis. *Molecular biology and evolution* **33**, 2177-
522 2181, doi:10.1093/molbev/msw111 (2016).
523 24 Kemmeren, P. *et al.* Large-scale genetic perturbations reveal regulatory networks and an

524 abundance of gene-specific repressors. *Cell* **157**, 740-752, doi:10.1016/j.cell.2014.02.054
525 (2014).
526 25 Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-
527 104, doi:10.1038/nature02800 (2004).
528 26 MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for Saccharomyces
529 cerevisiae. *BMC Bioinformatics* **7**, 113, doi:10.1186/1471-2105-7-113 (2006).
530 27 Neander, K. Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of*
531 *Science* **58**, 168-184 (1991).
532 28 Millikan, R. G. In Defense of Proper Functions. *Philosophy of Science* **56**, 288-302 (1989).
533 29 Amundson, R. & Lauder, G. V. Function without purpose. *Biology and Philosophy* **9**, 443-469,
534 doi:10.1007/BF00850375 (1994).
535 30 Chen, H., Wu, C. I. & He, X. The genotype-phenotype relationships in the light of natural
536 selection. *Molecular biology and evolution*, doi:10.1093/molbev/msx288 (2017).
537 31 Wong, S. L., Zhang, L. V. & Roth, F. P. Discovering functional relationships: biochemistry versus
538 genetics. *Trends Genet* **21**, 424-427, doi:10.1016/j.tig.2005.06.006 (2005).
539 32 He, X., Qian, W., Wang, Z., Li, Y. & Zhang, J. Prevalent positive epistasis in Escherichia coli and
540 Saccharomyces cerevisiae metabolic networks. *Nat Genet* **42**, 272-276, doi:10.1038/ng.524
541 (2010).
542 33 Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular
543 function. *Science* **353**, doi:10.1126/science.aaf1420 (2016).
544 34 Benschop, J. J. *et al.* A consensus of core protein complex compositions for Saccharomyces
545 cerevisiae. *Mol Cell* **38**, 916-928, doi:10.1016/j.molcel.2010.06.002 (2010).
546 35 Svejstrup, J. Q. Elongator complex: how many roles does it play? *Current Opinion in Cell Biology*
547 **19**, 331-336, doi:10.1016/j.ceb.2007.04.005 (2007).
548 36 Litchfield, D. W. Protein kinase CK2: structure, regulation and role in cellular decisions of life
549 and death. *The Biochemical journal* **369**, 1-15, doi:10.1042/BJ20021469 (2003).
550 37 Ohya, Y. *et al.* High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad*
551 *Sci U S A* **102**, 19015-19020, doi:10.1073/pnas.0509436102 (2005).
552 38 Okada, H., Ohnuki, S. & Ohya, Y. Quantification of cell, actin, and nuclear DNA morphology with
553 high-throughput microscopy and CalMorph. *Cold Spring Harb Protoc* **2015**, 408-412,
554 doi:10.1101/pdb.prot078667 (2015).
555 39 Ho, W. C. & Zhang, J. The genotype-phenotype map of yeast complex traits: basic parameters
556 and the role of natural selection. *Molecular biology and evolution* **31**, 1568-1580,
557 doi:10.1093/molbev/msu131 (2014).
558 40 Nogami, S., Ohya, Y. & Yvert, G. Genetic complexity and quantitative trait loci mapping of yeast
559 morphological traits. *Plos Genet* **3**, e31, doi:10.1371/journal.pgen.0030031 (2007).
560 41 Graur, D. & Li, W.-H. *Fundamentals of molecular evolution*. 2nd edn, (Sinauer Associates,
561 2000).
562 42 Chen, P., Wang, D., Chen, H., Zhou, Z. & He, X. J. G. r. The non-essentiality of essential genes in
563 yeast provides therapeutic insights into a human disease. *Genome Res.* **26**, 1355-1362 (2016).
564 43 Gietz, R. D. & Schiestl, R. H. J. N. p. High-efficiency yeast transformation using the LiAc/SS
565 carrier DNA/PEG method. *Nat. Protoc.* **2**, 31-34 (2007).
566 44 Scannell, D. R. *et al.* The Awesome Power of Yeast Evolutionary Genetics: New Genome
567 Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3-Genes Genom*

Genet **1**, 11-25, doi:10.1534/g3.111.000273 (2011).

45 Enyenihi, A. H. & Saunders, W. S. Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae. *Genetics* **163**, 47-54 (2003).

46 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

47 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

48 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

49 Qian, W., Ma, D., Xiao, C., Wang, Z. & Zhang, J. The Genomic Landscape and Evolutionary Resolution of Antagonistic Pleiotropy in Yeast. *Cell reports* **2**, 1399-1410, doi:https://doi.org/10.1016/j.celrep.2012.09.017 (2012).

50 Yu, G. C., Wang, L. G., Han, Y. Y. & He, Q. Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omics-a Journal of Integrative Biology* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

51 Yu, G. C. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976-978, doi:10.1093/bioinformatics/btq064 (2010).

589 **Figure legends**

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

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

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

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

600 **(d):** The 65 clustered effects have greater heritability scores than the 130 distributed
601 effects (P = 0.0011, Mann-Whitney U-test). Each dot represents an effect, and *P*-
602 value measures the statistical significance of a heritability score, with the vertical
603 dashed line showing adjusted P = 0.05.

604 **(e):** The proportions of deletion effects that are significantly heritable with adjusted P
605 < 0.05. Error bars represent SE.

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

607 **(g):** Conservation analysis of the HAP4 deletion effects. The 195 responsive genes
608 defined in BY4741(Δhap4) are examined with respect to their expression responses in
609 *S. paradoxus* N17 (Δhap4). The horizontal dashed line shows adjusted P = 0.05 and
610 vertical dashed lines show $\log_2$FC = ±0.58. (cyan: clustered effects; red: distributed
611 effects; cycle: down-regulated in BY4741(Δhap4); triangle: up-regulated in
612 BY4741(Δhap4))

613 **(h):** The rate of conservation in the three related yeasts for the 65 clustered effects and
614 130 distributed effects defined in BY4741(Δhap4), respectively. Error bars represent
615 SE.

616

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

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

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

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

629 **(f):** The proportion of direct target of HAP4 in different gene sets. Error bars represent
630 SE.

631

632 **Fig. 3. Clustered effects support related biochemistry understandings much**

633 **better than distributed effects in a variety of protein complexes and KEGG**
634 **pathways.**
635 **(a):** The clustered effects of genes encoding a protein complex in general overlap more
636 than their distributed effects. Each circle represents a complex, and the filled ones are
637 significant at a 99% confidence level estimated by random sampling. A total of 54
638 protein complexes are included here, with 24 cases showing at least twice more
639 overlapped clustered effects than overlapped distributed effects (below the line y =
640 0.5x). The numbers of effects have been normalized such that in each case the
641 overlaps of clustered effects and the overlaps of distributed effects can be directly
642 compared.
643 **(b):** The representative GO terms of the overlapped clustered effects of the elongator
644 holoenzyme complex. Only four genes encoding the complex, which are highlighted
645 in orange, have suitable expression data for the analysis. There are 10 overlapped
646 clustered effects of the four genes, which are over 300 times more than expected. The
647 expectation is estimated by random sampling of the distributed effects of the focal genes
648 to calculate overlaps. The orange circle each represents an overlapped clustered effect,
649 and the blue circles represent the enriched GO terms of the overlapped clustered effects
650 with the number inside showing the fold enrichment in the given term.
651 **(c):** The representative GO terms of the overlapped clustered effects of the protein
652 kinase CK2 complex. There are 21 overlapped clustered effects, 83.3 times more than
653 expected.
654 **(d):** The clustered effects of genes in the same KEGG pathway also tend to overlap
655 more than their distributed effects. Each circle represents a pathway, and the filled
656 ones are significant at a 99% confidence level. A total of 41 pathways are included,
657 with nine showing at least five times more overlapped clustered effects than overlapped
658 distributed effects (below the line y = 0.2x). The number of effects have been
659 normalized such that in each case the overlaps of clustered effects and the overlaps of
660 distributed effects can be directly compared.
661 **(e):** The representative GO terms of the overlapped clustered effects of the four genes
662 in the metabolic pathway sce00260. There are six overlapped clustered effects, over
663 100 times more than expected.
664 **(f):** The representative GO terms of the overlapped clustered effects of the four genes
665 in the genetic information processing pathway sce03010. There are 10 overlapped
666 clustered effects, 21.4 times more than expected.
667
668 **Fig. 4. Examination of cell morphological traits also supports the role of**
669 **evolution in separating genetic effects.**
670 **(a):** The yeast cell morphology outlined by coordinate points, lines and angles (only
671 some are shown) based on which a total of 405 quantitative traits can be derived by a
672 computer software.
673 **(b):** The traits affected by HAP4 in both *S. cerevisiae* and *S. paradoxus* (i.e., conserved
674 effects) are more likely to overlap with those affected by HAP2, HAP3 and HAP5 than
675 the traits affected by HAP4 only in *S. cerevisiae* (non-conserved effects) (P = 0.035,
676 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 genetics. 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 accounting for pleiotropy.

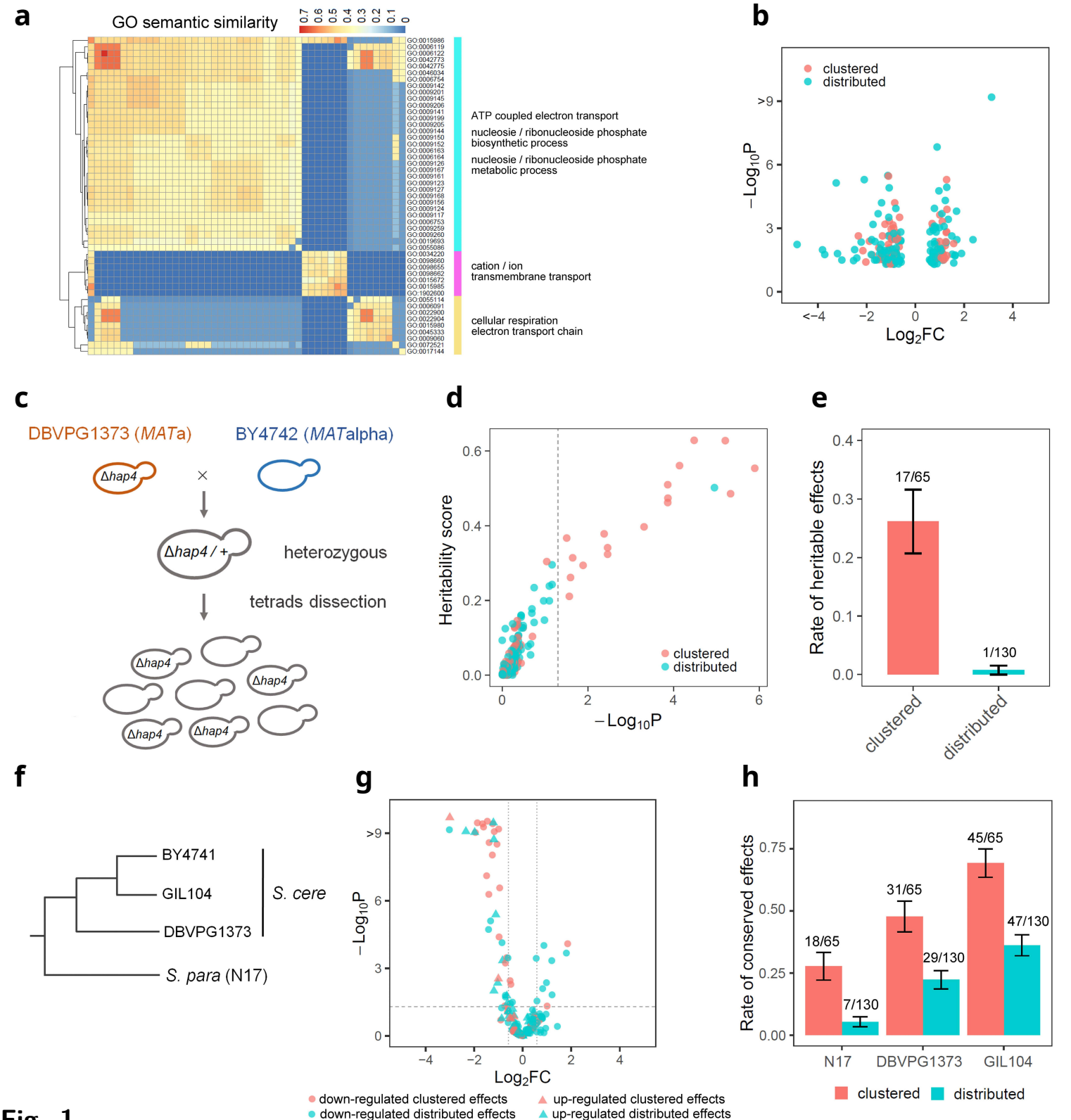

a
GO semantic similarity
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ATP coupled electron transport
nucleosie / ribonucleoside phosphate biosynthetic process
nucleosie / ribonucleoside phosphate metabolic process
cation / ion transmembrane transport
cellular respiration electron transport chain
b
clustered
distributed
−Log10P
Log2FC
c
DBVPG1373 (MATa)
BY4742 (MATalpha)
Δhap4
×
Δhap4 / +
heterozygous
tetrads dissection
d
Heritability score
−Log10P
clustered
distributed
e
Rate of heritable effects
17/65
1/130
clustered
distributed
f
BY4741
GIL104
DBVPG1373
S. cere
S. para (N17)
g
−Log10P
Log2FC
down-regulated clustered effects
down-regulated distributed effects
up-regulated clustered effects
up-regulated distributed effects
h
Rate of conserved effects
18/65
7/130
31/65
29/130
45/65
47/130
N17
DBVPG1373
GIL104
clustered
distributed

Fig. 1

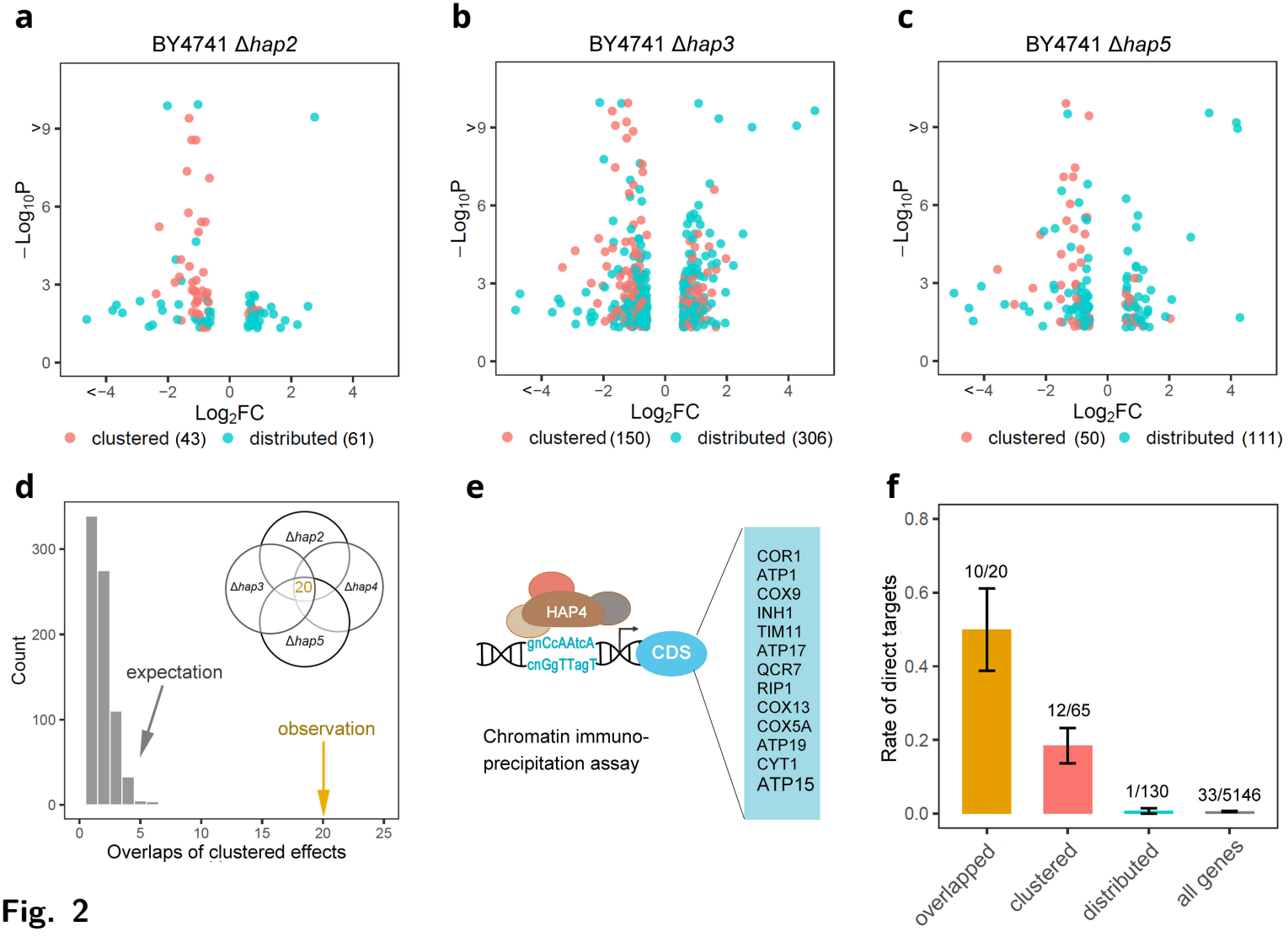
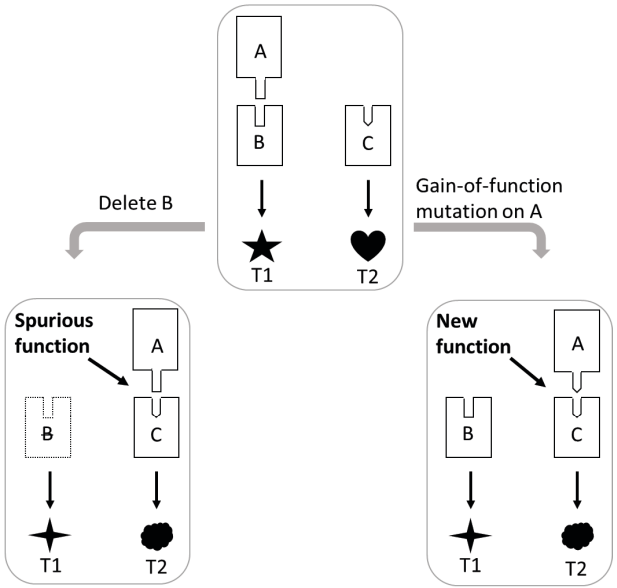

a
BY4741 Δhap2
−Log10P
Log2FC
clustered (43)
distributed (61)
b
BY4741 Δhap3
clustered (150)
distributed (306)
c
BY4741 Δhap5
clustered (50)
distributed (111)
d
Count
Overlaps of clustered effects
expectation
observation
Δhap2
Δhap3
Δhap4
Δhap5
20
e
HAP4
gnCcAAtcA
cnGgTTagT
CDS
Chromatin immuno-
precipitation assay
COR1
ATP1
COX9
INH1
TIM11
ATP17
QCR7
RIP1
COX13
COX5A
ATP19
CYT1
ATP15
f
Rate of direct targets
10/20
12/65
1/130
33/5146
overlapped
clustered
distributed
all genes

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***.

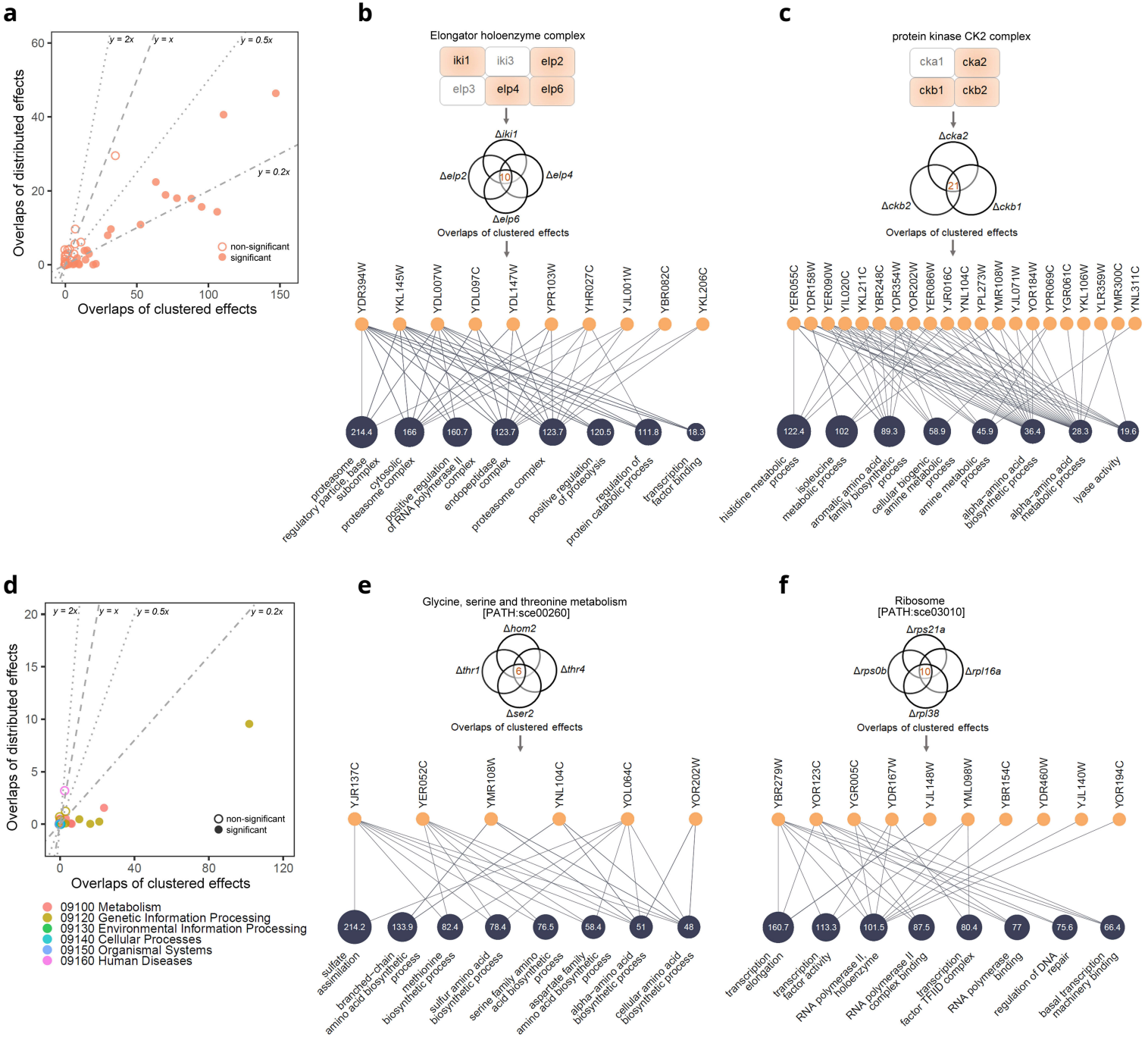

a
Overlaps of distributed effects
Overlaps of clustered effects
y = 2x
y = x
y = 0.5x
y = 0.2x
non-significant
significant
b
Elongator holoenzyme complex
iki1
iki3
elp2
elp3
elp4
elp6
Δiki1
Δelp2
Δelp4
Δelp6
10
Overlaps of clustered effects
YDR394W
YKL145W
YDL007W
YDL097C
YDL147W
YPR103W
YHR027C
YJL001W
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

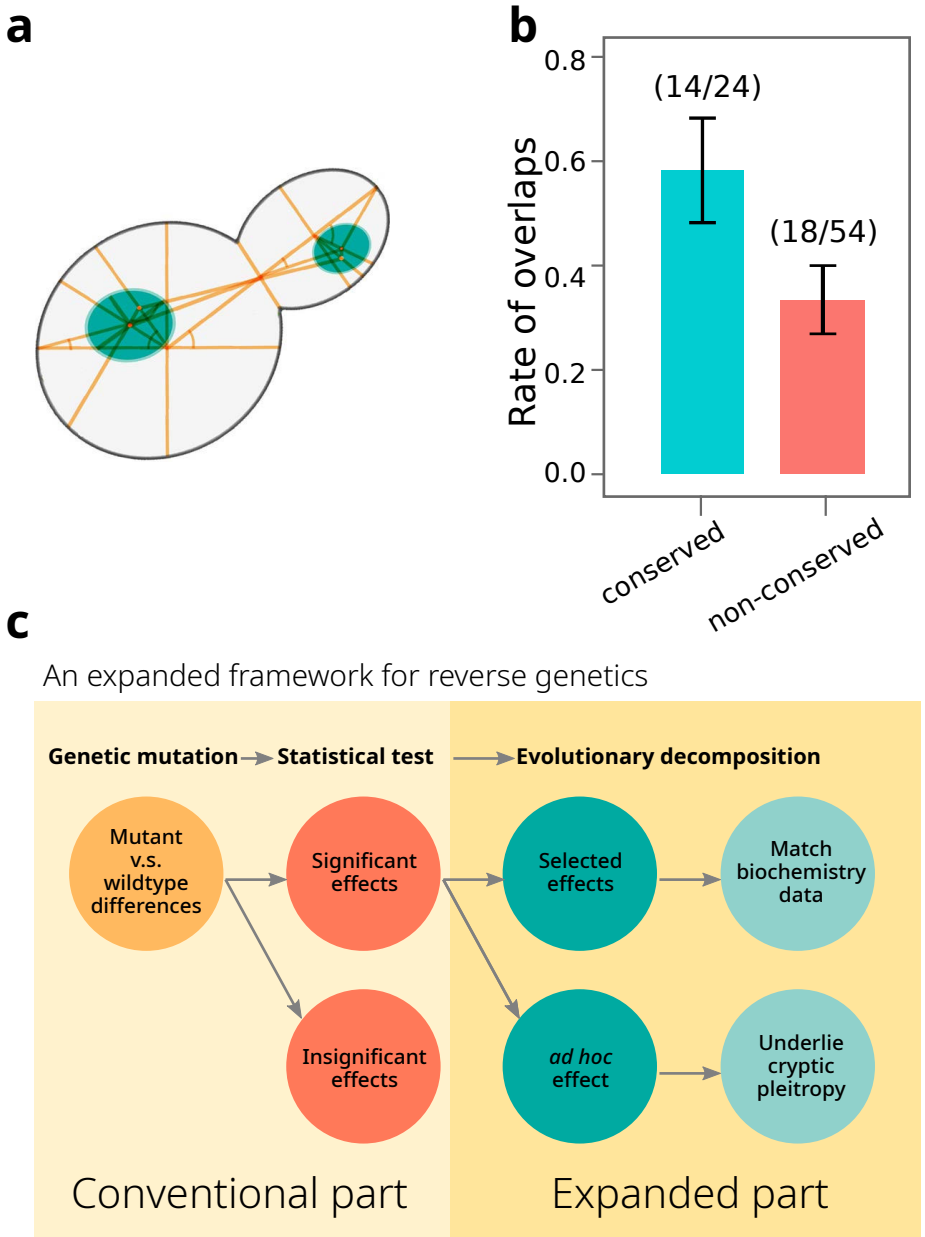

a
b
Rate of overlaps
0.8
0.6
0.4
0.2
0.0
(14/24)
(18/54)
conserved
non-conserved
c
An expanded framework for reverse genetics
Genetic mutation
Statistical test
Evolutionary decomposition
Mutant v.s. wildtype differences
Significant effects
Insignificant effects
Selected effects
ad hoc effect
Match biochemistry data
Underlie cryptic pleitropy
Conventional part
Expanded part

Fig. 4

688
689 **Supporting Information** of

690 "Evolutionary decomposition of genetic effects"

691

692 Li Liu[#], Mengdi Liu[#], Di Zhang, Shanjun Deng, Piaopiao Chen, Jing Yang, Yunhan Xie
693 & Xionglei He*
694
695 State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University,
696 Guangzhou 510275, China
697
698 **This file contains:**
699 Legends of Tables S1 to S8
700 Figs. S1 to S9
701
702
703 **Legends of supplementary tables**
704 **Table S1:** RNA-seq-based gene expression levels for each HAP4 deletion and wild-
705 type strain.
706 **Table S2:** RNA-seq-based expression changes (P-value and FC) of the 195
707 responsive genes in BY4741(Δhap4) and other deletion lines, with the heritability
708 scores also included.
709 **Table S3:** RNA-seq-based gene expression changes after deleting HAP2, HAP3,
710 HAP4 and HAP5, respectively, in BY4741.
711 **Table S4:** Summary of the analyses of protein complexes in this study.
712 **Table S5:** Summary of the analyses of KEGG pathways in this study.
713 **Table S6:** The affected morphological traits in a variety of gene deletion lines.
714 **Table S7:** Summary of the clustered effects and distributed effects defined in each
715 mutant that has public microarray data.
716 **Table S8:** Summary of the trait information of each diploid gene deletion or wild-
717 type yeast strain, with the 405 trait values, the number of examined cells, and the
718 number of replications included.
719

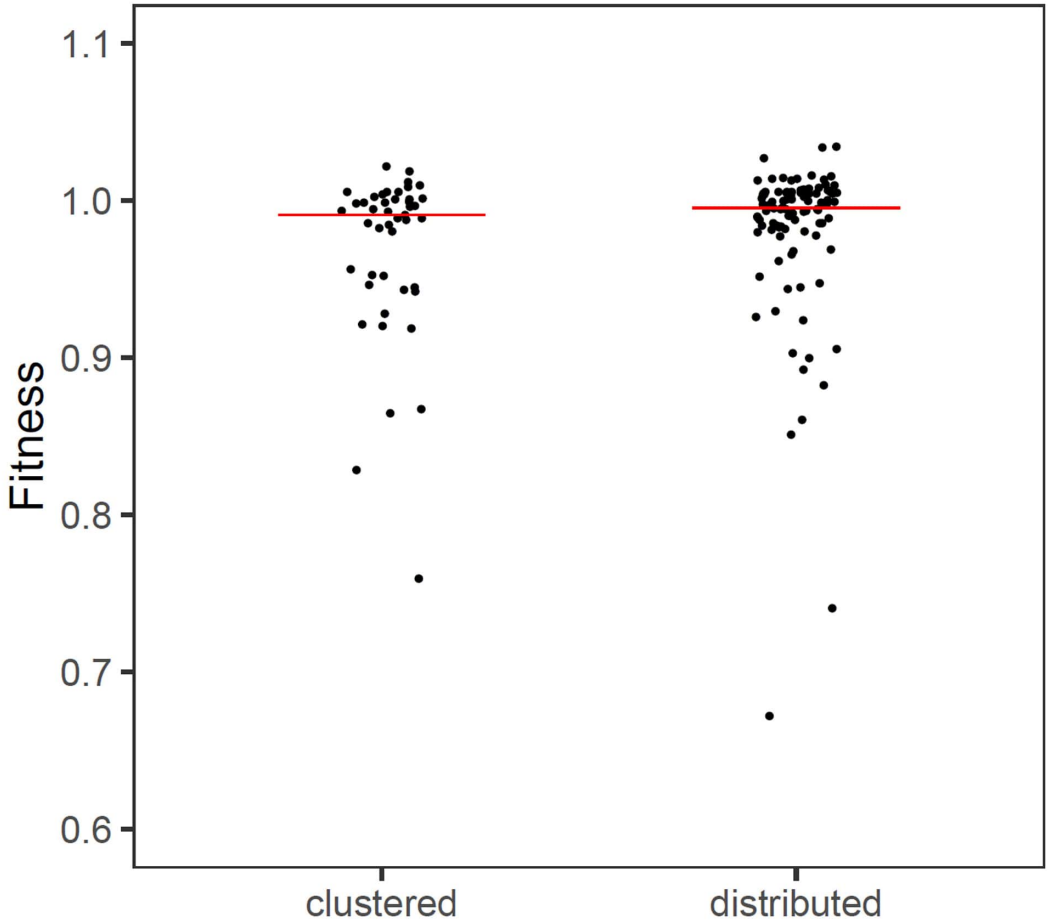


Fig. S1. Fitness importance is comparable between the clustered responsive genes and distributed responsive genes defined in BY4741(Δhap4) (*P* = 0.10, Mann-Whitney U-test). Fitness importance of a gene is measured by the relative growth rate of the gene deletion line to wild-type. The horizontal line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

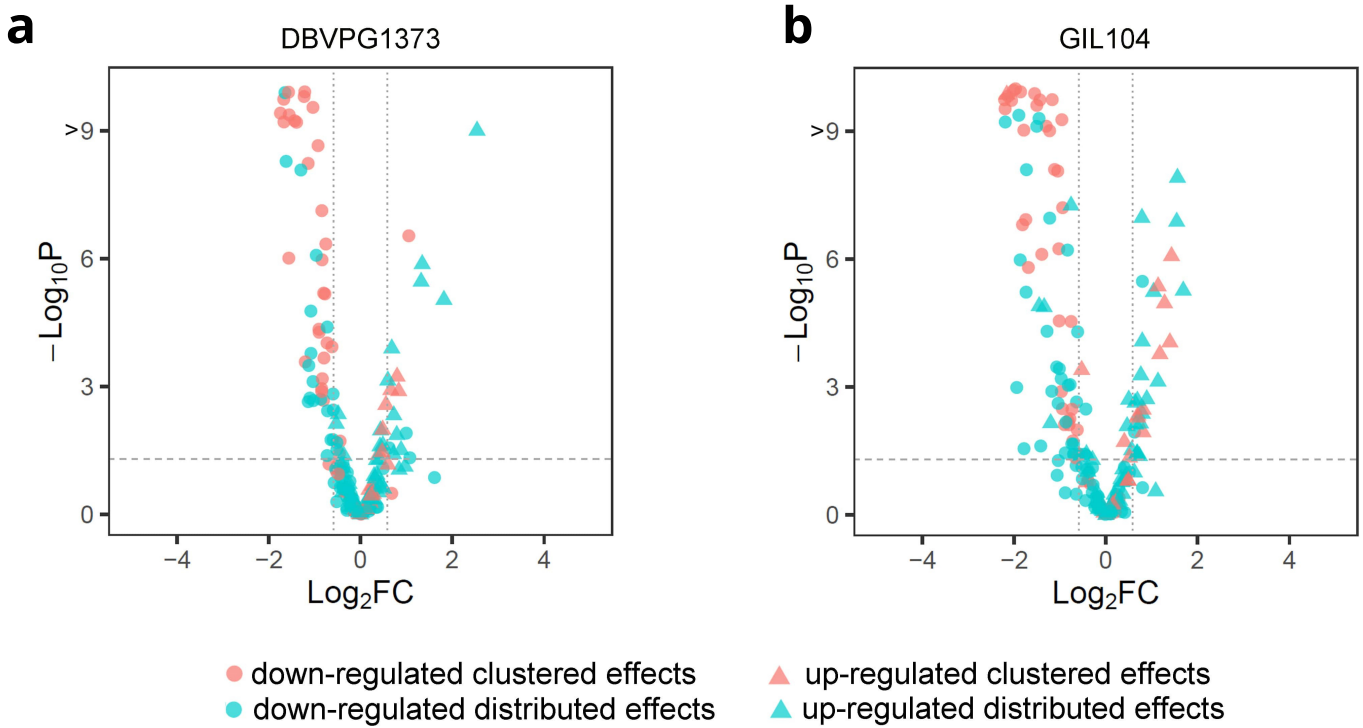


Fig. S2. The intra-species conservation analysis of the HAP4 deletion effects. The 195 responsive genes defined in BY4741(Δhap4) are examined with respect to their expression responses in *S. cerevisiae* DBVPG1373(Δhap4) and GIL104(Δhap4), respectively. The horizontal dashed line shows adjusted P = 0.05 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)).

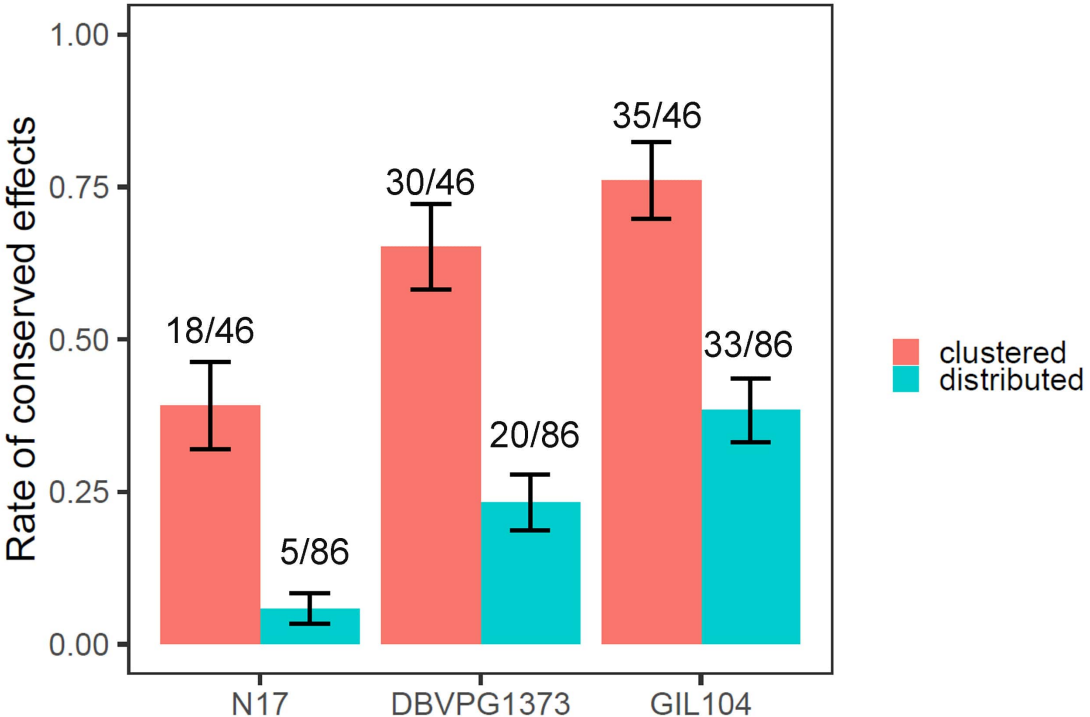


Fig. S3. Conservation analysis of the HAP4 deletion effects by considering only genes with a strong expression level in wild-type BY4741. This analysis is to address the concern that lowly expressed genes in wild-type tend not to have detectable down-regulation due to technical bias. Hence, for the 195 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.

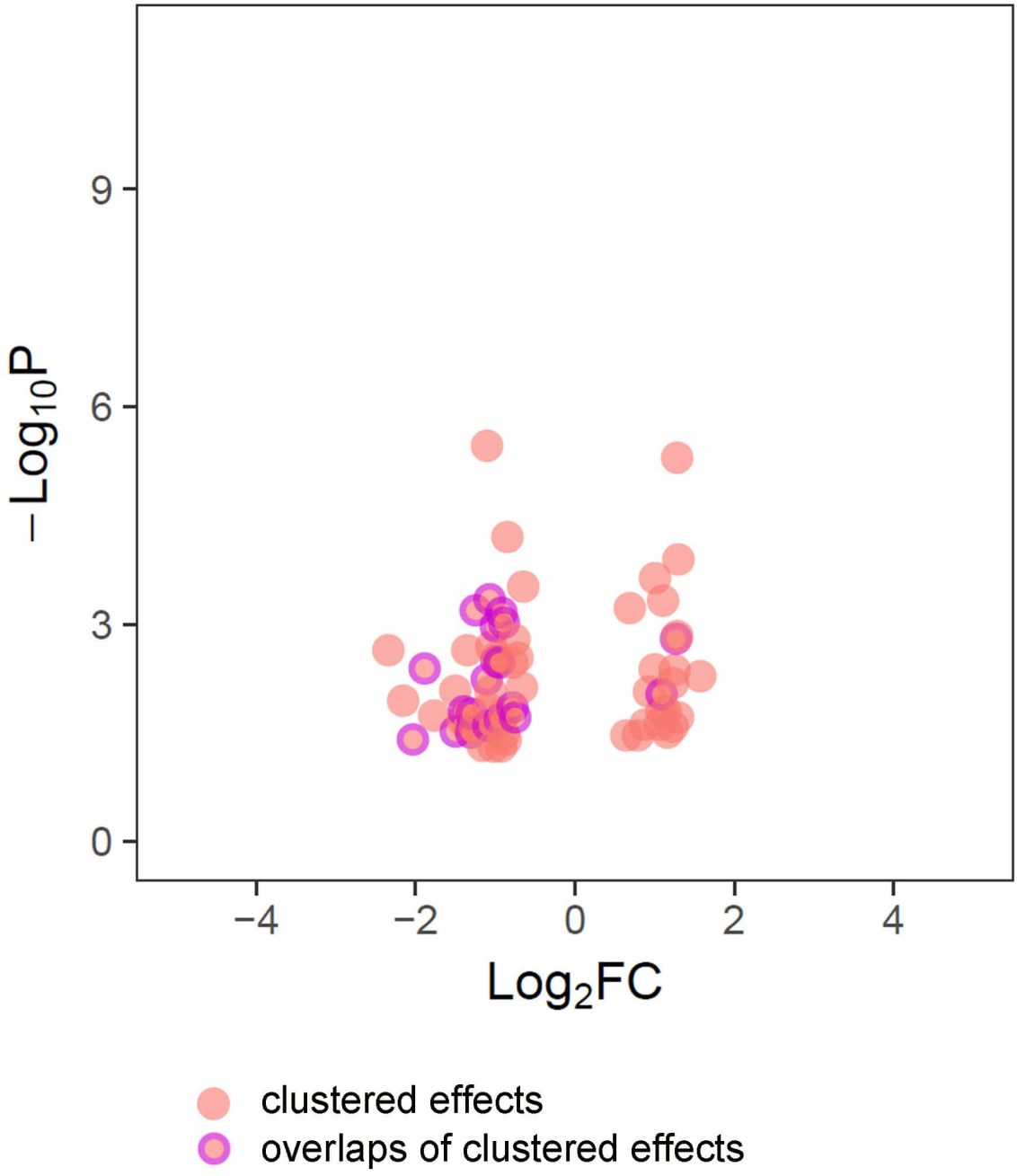
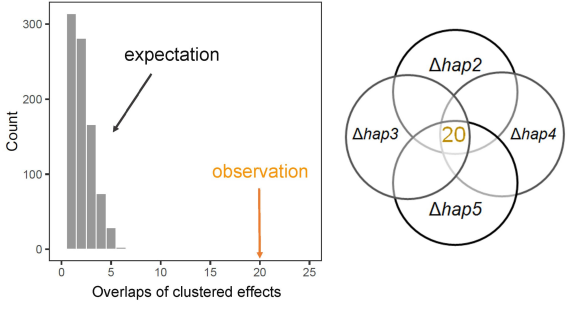


Fig. S4. The 20 overlapped clustered effects are comparable to the rest 45 (65-20) clustered effects defined in BY4741(Δhap4) with regard to their effect size. The differences are not statistically significant for both the P-values and fold changes observed in BY4741(Δhap4) (*P* = 0.75 and 0.51, respectively,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

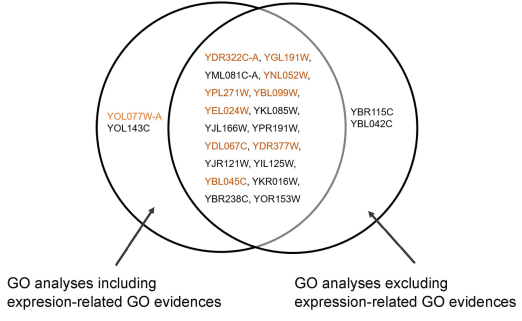
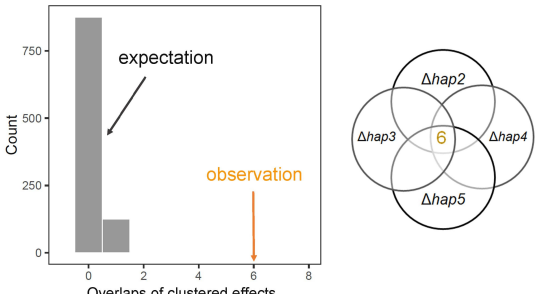


Fig. S5. The strong overlaps of clustered effects of the HAP2/3/4/5 complex genes are not biased by expression-related GO evidences for annotating the deletion effects. (A) Only the evidences IDA, HDA and IPI are used to re-define the clustered and distributed deletion effects of the four genes, respectively. (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

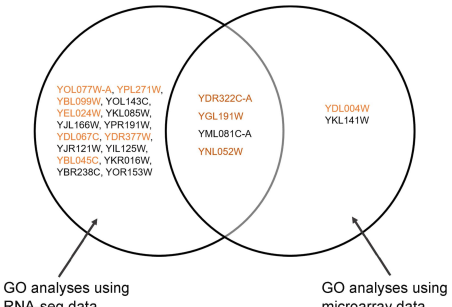


Fig. S6. The enrichment of overlapped clustered effects in HAP2/3/4/5 complex is reproduced by using public microarray data of the four gene deletion lines. (A) Microarray-based expression data are used to re-define the clustered and distributed deletion effects of the four genes, respectively. (B) There 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.

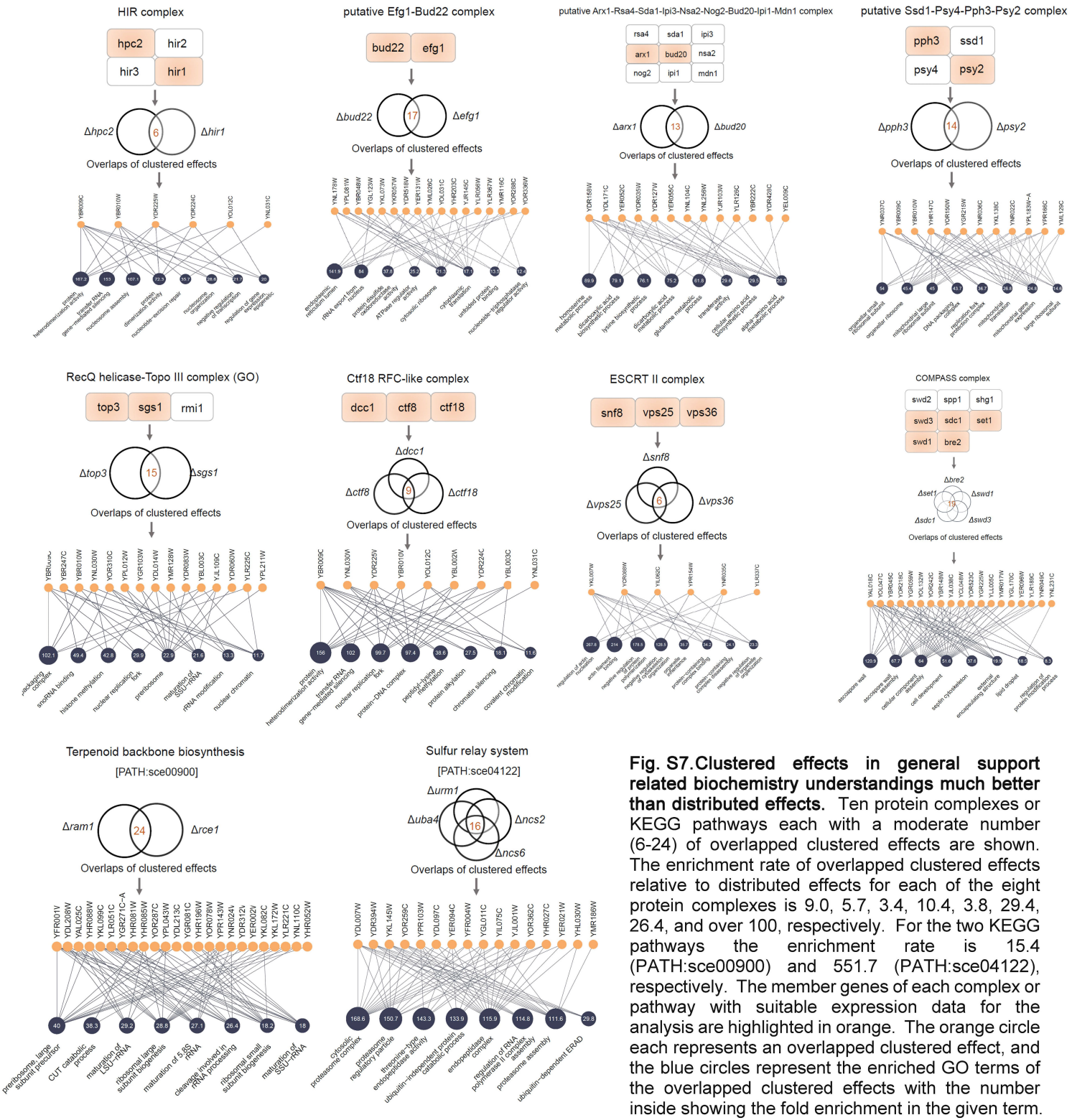


Fig. S7. Clustered effects in general support related biochemistry understandings much better than distributed effects. Ten protein complexes or KEGG pathways each with a moderate number (6-24) of overlapped clustered effects are shown. The enrichment rate of overlapped clustered effects relative to 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.

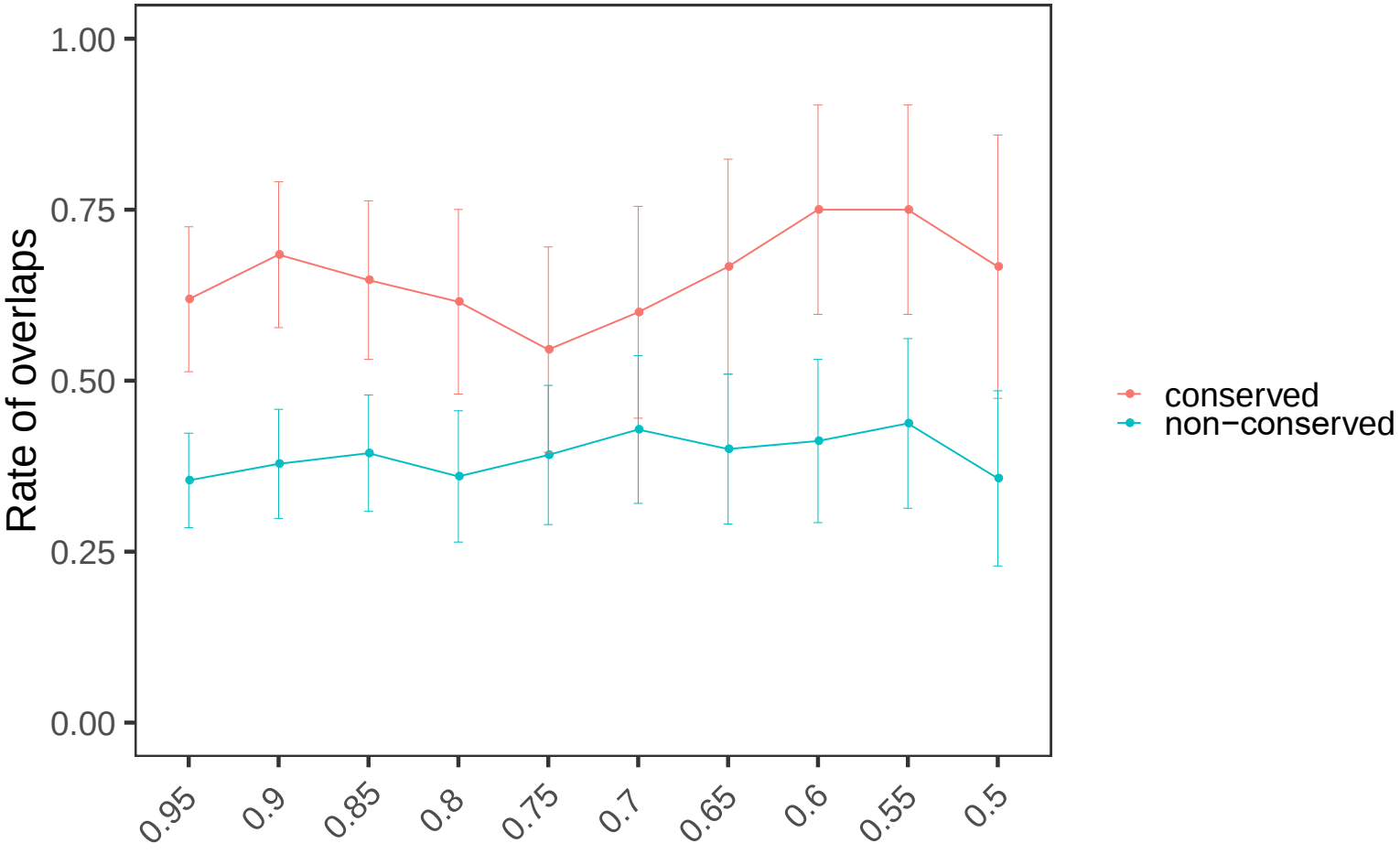


Fig. S8. The estimated rate of overlaps cannot be explained by correlated traits. The Pearson's R of all trait pairs is calculated using the trait values generated in ref. 4 for 4,718 yeast mutants. We then removed traits one by one from those with the highest absolute R until no two traits have R^2 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.

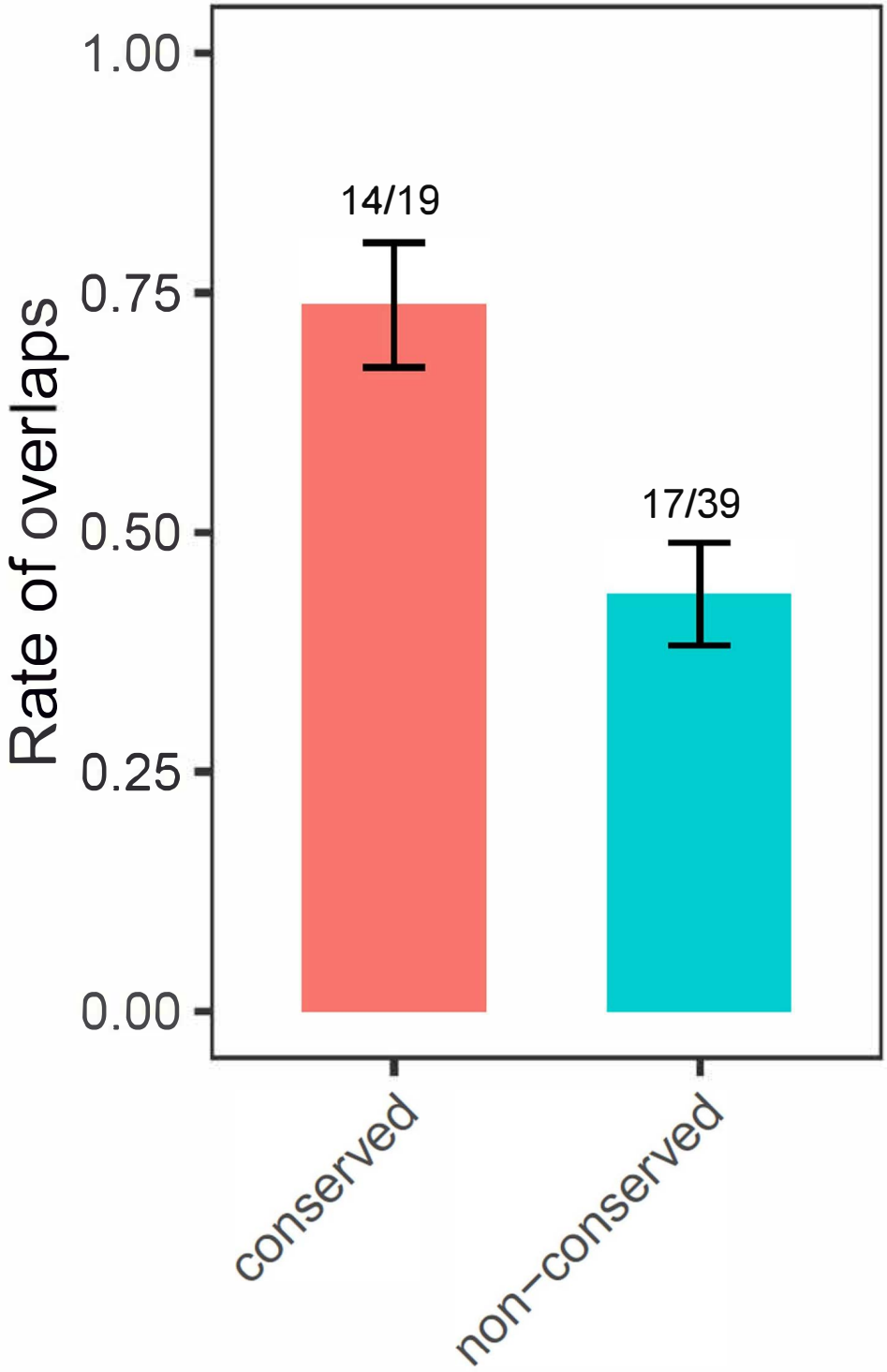


Fig. S9. The comparison in Fig. 4B is robust against trait measuring noise. To address the potential technical bias that traits with large measuring noise tend to be both non-conserved and non-overlapping, only traits with measuring CV < 0.1 across the replicates in wild-type BY4741 are considered. This 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*.