May 30, 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 - Genetics has long been confused by the pleiotropic mutational effects associated with even clearly-annotated genes. To understand such genetic effects, we here examine HAP4, a well-studied transcription factor in *Saccharomyces cerevisiae* that forms 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). The clustered effects tend to be conserved in related yeasts while the distributed effects, which account for the pleiotropy of HAP4, are genetically non-heritable and evolutionarily volatile. There are substantial overlaps among clustered effects, but not distributed effects, of the four genes encoding the HAP2/3/4/5 tetramer, and this pattern holds for other biochemically characterized protein complexes and KEGG pathways. We further examine a large set of cell morphological traits of the gene deletion lines and obtain consistent results. Hence, an expanded framework for reverse genetics is suggested, in which genetic effects are further divided into evolutionarily selected and unselected ones, with the former supporting related biochemistry understandings and the latter accounting for cryptic pleiotropy.

It has been increasingly recognized that a gene can affect various seemingly unrelated traits\(^1\)\(^{-6}\), a phenomenon termed pleiotropy\(^7\)\(^,\)\(^8\). For instance, a mapping of quantitative trait loci (QTL) for 70 skeletal traits in mice showed that on average a QTL
affects about 10% of the traits\textsuperscript{9}. Attempts to understand pleiotropic genetic effects were primarily from mechanistic perspectives\textsuperscript{10-14}, by considering the focal gene’s multiple molecular functions or multiple cellular consequences associated with a single molecular function\textsuperscript{15}. Since biological phenomena would make little sense except in the light of evolution\textsuperscript{16}, a dissection of genetic effects from an evolutionary perspective may provide novel insights\textsuperscript{17}.

The yeast \textit{Saccharomyces cerevisiae} gene HAP4 is a non-essential transcription factor that has been subjected to extensive studies since its discovery 30 years ago\textsuperscript{18}. We deleted the open reading frame of HAP4 in \textit{S. cerevisiae} strain BY4741, and checked the expression trait of the other yeast genes (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\textsuperscript{19,20} (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 well characterized, with comparable fitness importance ($P = 0.1$, Mann-Whitney U-test; Fig. S1 and Table S2). 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).

Genetic effects observed in an isogenic background should be heritable in
population to be evolutionarily effective\textsuperscript{21}. We crossed two \textit{S. cerevisiae} strains to obtain a population of yeast segregants. Specifically, a wild-type strain BY4742(MATa\textsubscript{pha}), 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 defined a heritability score $HS = \frac{\text{Var(wt, null)}}{\text{Var(total)}}$, where $\text{Var(wt, null)}$ is the variance of the focal gene’s expression level between the two segregant groups, and $\text{Var(total)}$ is the total variance across the segregants. We found that the 65 clustered effects in general have much higher $HS$ than the 130 distributed effects ($P = 1.1 \times 10^{-3}$, Mann-Whitney U-test; Fig. 1D). Approximately 26.2\% (17/65) of the clustered effects are significantly heritable under a stringent statistical cutoff, while the number is 0.8\% (1/130) for the distributed effects ($P = 2.1 \times 10^{-8}$, Fisher’s exact test; Fig. 1E).

To test how the genetic effects are maintained by natural selection, we examined their evolutionary conservation. We obtained, under the same statistical cutoff as in BY4741, the deletion effects of HAP4 in \textit{Saccharomyces paradoxus} strain N17, a closely related yeast species diverged from \textit{S. cerevisiae} ~10 million years ago\textsuperscript{22} (Fig. 1F). Only 5.4\% (7/130) of the distributed effects found in \textit{S. cerevisiae} were also observed in the HAP4 deletion line of \textit{S. paradoxus} N17, while the number was 27.7\% (18/65) for the clustered effects ($P = 3.1 \times 10^{-5}$, Chi-square test). The difference is
robust as evidenced by plotting the expression responses of the 195 genes in N17(Dhap4) (Fig. 1G). We further examined intra-species conservation of the HAP4 deletion effects by looking at two other *S. cerevisiae* strains DBVPG1373 and GIL104, the former of which is 0.35% diverged from the strain BY4741 and the latter 0.07% diverged at the genomic level. Similarly, the 130 distributed effects were largely unreproducible in strains DBVPG1373(Dhap4) and GIL104(Dhap4), while the 65 clustered effects were much more conserved (*P* = 0.5 × 10^-4 for DBVPG1373(Dhap4) and *P* = 2.6 × 10^-5 for GIL104(Dhap4), Chi-square test; Fig. S2). As expected, both types of the deletion effects are more reproducible in more related yeasts (Fig. 1H).

We further confirmed that the results cannot be explained by different detectability of expression changes between the two gene sets by excluding those genes lowly expressed in wild-type BY4741 (Fig. S3). These data, together with the heritability analysis, suggest that the clustered effects of HAP4 tend to be selected in evolution; on the contrary, the distributed genetic effects appear non-evolutionary, likely representing context-dependent *ad hoc* responses of cell to the gene deletion.

HAP4 functions by forming a tetramer with HAP2, HAP3 and HAP5 as a result of evolution. We hypothesized that clustered genetic effects should support this biochemistry understanding better than distributed genetic effects, because the latter is non-evolutionary. To test the hypothesis, we deleted the other three genes that encode the tetramer, respectively, in *S. cerevisiae* BY4741, and measured the expression profiles of the deletion lines. We defined clustered effects and distributed effects for each of 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($\Delta$hap4) (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 HPA5 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($\Delta$hap4), 13 are direct targets of HAP4 (Fig. 2E)$^{27,28}$. Interestingly, there are 24-fold enrichment of direct targets in the clustered effects relative to the distributed effects ($P = 8.6 \times 10^{-6}$, 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) ($P = 4.6 \times 10^{-18}$, Hypergeometric test) (Fig. 2F). Hence, the CHIP data well support the distinction of the two effect types.

Collectively, these results are consistent with a recently proposed model that the
null phenotype of a gene can be ascribed to either the loss of native functions of the
gene, or the gain of spurious functions arising from the genetic manipulation due to
passive adjustments of the cell system\textsuperscript{24}. The key difference of the two function types
is their evolutionary effectiveness: native functions are historical and evolutionarily
selected while spurious functions are ahistorical and non-evolutionary\textsuperscript{29-32}.
Accordingly, the distributed effects examined here likely represent spurious functions
created by the HAP4 deletion, and the clustered effects could be in a large part ascribed
to the native functions of HAP4. It is intriguing how the two effect types defined by
GO could fit the two function types described in the model. We reason that
evolutionarily optimized native functions are likely to regulate specific pathways or
processes; losing them would thus cause clustered changes of the related genes, which
are detected by GO analysis. In contrast, spurious functions may affect the
transcriptome in a rather random way, resulting in distributed changes across diverse
cellular processes, most of which cannot be covered by overrepresented GO terms.
This may explain why GO clustering here could echo evolutionary effectiveness.
Regardless of the underlying logic, clustered genetic effects seem to be well
matched with related biochemistry understandings. This would help resolve a long-
standing challenge to molecular biology - the gap between genetic analysis and
biochemistry analysis\textsuperscript{33,34}; specifically, genes with intimate biochemical interactions do
not have common genetic effects and genes with common genetic effects do not show
intimate biochemical interactions\textsuperscript{35,36}. To test the generality 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\textsuperscript{26}. There are 54 protein complexes annotated
by a previous study suitable for our analysis\textsuperscript{37}. 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\textsuperscript{38}. 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\textsuperscript{39}.

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 biochemistry understandings better than clustered effects, highlighting the cryptic nature of them. Taken together, focusing on clustered effects is a readily operational approach to narrowing the gap between genetic analysis and biochemistry 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. With the help of a computer software as many as 405 quantitative traits can be obtained from cell wall and nuclear stained cell images. 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 great resource for studying genotype-phenotype relationships. 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%) for the non-conserved effects of HAP4 (*P* = 0.035, Fisher’s exact test; Fig. 4B). The
difference remains largely unchanged when only traits with small measuring noise are
considered (Fig. S8). Hence, the cell morphology data also support the role of
evolution in separating 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 aims to find causality only: it relies 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 unselected one, the former
of which would support related biochemistry understandings while the latter account
for cryptic pleiotropy. When expression traits are examined, GO clustering serves as
a proxy for evolutionary effectiveness, which offers a readily operational means to
delineate expression network using genetic data.

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 study43, the standard LiAc transformation method44 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.
paradoxus45; 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 study46. 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+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 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
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)\(^47\). 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](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](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)\(^48\) 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\(^49\) 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\(^50\). 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](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\(^51\) 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\(^52\).

### Heritability analysis

To estimate the heritability of the 195 HAP4 deletion effects, a population of 24
segregants were divided into two groups: 12 with the wild-type allele of HAP4 and 12 with the null allele of HAP4. The heritability score H was the variation between the two groups divided by the total variation of the population. To test the statistical significance of an H, 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, 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. The KEGG pathways of the yeast *S. cerevisiae* were downloaded from KEGG website ([https://www.genome.jp/kegg-bin/get_htext?sce00001](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. In brief, a single yeast colony was picked and cultured in YPD liquid overnight with shaking to the saturation phase. Then, 1.5uL culture was transferred to 100uL 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, concanaavalin 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. For each line three independent clones were available. For each clone there were 6-8 technical replicates in the staining and image capturing. The images were analyzed by CalMorph 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 larger 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 J. Yang, J. Zhang, W. Qian, and Y. Zhang.

References

1. Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Syst Biol* 1, 2005 0001, doi:10.1038/msb4100004 (2005).
2. Sonnichsen, B. *et al.* Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature* 434, 462-469, doi:10.1038/nature03353 (2005).
3. Mackay, T. F. Mutations and quantitative genetic variation: lessons from Drosophila. *Philos Trans R Soc Lond B Biol Sci* 365, 1229-1239, doi:10.1098/rstb.2009.0315 (2010).
Ohya, Y. et al. High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci U S A* **102**, 19015-19020, doi:10.1073/pnas.0509436102 (2005).

Sivakumaran, S. et al. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607-618, doi:10.1016/j.ajhg.2011.10.004 (2011).

Wang, Z., Liao, B. Y. & Zhang, J. Genomic patterns of pleiotropy and the evolution of complexity. *Proc Natl Acad Sci U S A* **107**, 18034-18039, doi:10.1073/pnas.1004666107 (2010).

Caspari, E. J. E. PLEIOTROPIC GENE ACTION. *6*, 1-18 (1952).

Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767-773, doi:10.1534/genetics.110.122549 (2010).

Wagner, G. P. et al. Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature* **452**, 470-472, doi:10.1038/nature06756 (2008).

Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12**, 204-213, doi:10.1038/nrg2949 (2011).

Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet* **29**, 66-73, doi:10.1016/j.tig.2012.10.010 (2013).

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483-495, doi:10.1038/nrg3461 (2013).

Foster, K. R., Shaulsky, G., Strassmann, J. E., Queller, D. C. & Thompson, C. R. Pleiotropy as a mechanism to stabilize cooperation. *Nature* **431**, 693-696, doi:10.1038/nature02894 (2004).

He, X. L. & Zhang, J. Z. Toward a molecular understanding of pleiotropy. *Genetics* **173**, 1885-1891, doi:10.1534/genetics.106.060269 (2006).

Ayala, F. J. "Nothing in biology makes sense except in the light of evolution": Theodosius Dobzhansky: 1900-1975. *J Hered* **68**, 3-10 (1977).

Chen, H., Wu, C. I. & He, X. The genotype-phenotype relationships in the light of natural selection. *Molecular biology and evolution*, doi:10.1093/molbev/msx288 (2017).

Forsburg, S. L. & Guarente, L. Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev* **3**, 1166-1178 (1989).

Lascaris, R. et al. Hap4p overexpression in glucose-grown Saccharomyces cerevisiae induces cells to enter a novel metabolic state. *Genome Biol* **4**, R3 (2003).

Buschlen, S. 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, doi:10.1002/cfg.254 (2003).

Pinzon, N. et al. microRNA target prediction programs predict many false positives. *Genome Res* **27**, 234-245, doi:10.1101/gr.205146.116 (2017).

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254, doi:10.1038/nature01644 (2003).

Liti, G. et al. Population genomics of domestic and wild yeasts. *Nature* **458**, 337-341, doi:10.1038/nature07743 (2009).

He, X. The Biology Complicated by Genetic Analysis. *Molecular biology and evolution* **33**, 2177-2181, doi:10.1093/molbev/msw111 (2016).

Bolotin-Fukuhara, M. Thirty years of the HAP2/3/4/5 complex. *Biochim Biophys Acta Gene Regul Mech* **1860**, 543-559, doi:10.1016/j.bbagrm.2016.10.011 (2017).
Kemmeren, P. et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**, 740-752, doi:10.1016/j.cell.2014.02.054 (2014).

Harbison, C. T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104, doi:10.1038/nature02800 (2004).

MacIsaac, K. D. et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 113, doi:10.1186/1471-2105-7-113 (2006).

Neander, K. Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of Science* **58**, 168-184 (1991).

Harbison, C. T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104, doi:10.1038/nature02800 (2004).

MacIsaac, K. D. et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 113, doi:10.1186/1471-2105-7-113 (2006).

Neander, K. Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of Science* **58**, 168-184 (1991).

Millikan, R. G. In Defense of Proper Functions. *Philosophy of Science* **56**, 288-302 (1989).

Amundson, R. & Lauder, G. V. Function without purpose. *Biology and Philosophy* **9**, 443-469, doi:10.1007/BF00850375 (1994).

Graur, D. 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, doi:10.1093/gbe/evt028 (2013).

Wong, S. L., Zhang, L. V. & Roth, F. P. Discovering functional relationships: biochemistry versus genetics. *Trends Genet* **21**, 424-427, doi:10.1016/j.tig.2005.06.006 (2005).

He, X., Qian, W., Wang, Z., Li, Y. & Zhang, J. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat Genet* **42**, 272-276, doi:10.1038/ng.524 (2010).

Costanzo, M. et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, doi:10.1126/science.aaa4239 (2016).

Benshop, J. J. et al. A consensus of core protein complex compositions for Saccharomyces cerevisiae. *Mol Cell* **38**, 916-928, doi:10.1016/j.molcel.2010.06.002 (2010).

Svejstrup, J. Q. Elongator complex: how many roles does it play? *Current Opinion in Cell Biology* **19**, 331-336, doi:10.1016/j.ceb.2007.04.005 (2007).

Litchfield, D. W. Protein kinase CK2: structure, regulation and role in cellular decisions of life and death. *The Biochemical journal* **369**, 1-15, doi:10.1042/BJ20021469 (2003).

Okada, H., Ohnuki, S. & Ohya, Y. Quantification of cell, actin, and nuclear DNA morphology with high-throughput microscopy and CalMorph. *Cold Spring Harb Protoc* **2015**, 408-412, doi:10.1101/pdb.prot078667 (2015).

Ho, W. C. & Zhang, J. The genotype-phenotype map of yeast complex traits: basic parameters and the role of natural selection. *Molecular biology and evolution* **31**, 1568-1580, doi:10.1093/molbev/msu131 (2014).

Nogami, S., Ohya, Y. & Yvert, G. Genetic complexity and quantitative trait loci mapping of yeast morphological traits. *Plos Genet* **3**, e31, doi:10.1371/journal.pgen.0030031 (2007).

Chen, P., Wang, D., Chen, H., Zhou, Z. & He, X. J. G. The non-essentiality of essential genes in yeast provides therapeutic insights into a human disease. *Genome Res* **26**, 1355-1362 (2016).

Gietz, R. D. & Schiestl, R. H. J. N. p. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31-34 (2007).

Scannell, D. R. 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, doi:10.1534/g3.111.000273 (2011).
Enyenihi, A. H. & Saunders, W. S. Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae. Genetics 163, 47-54 (2003).

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

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).

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).

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).

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).

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).
**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 heritability scores than the 130 distributed effects ($P = 0.0011$, Mann-Whitney U-test). Each dot represents an effect, and P-value measures the statistical significance of a heritability score, 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($\Delta$hap4) are examined with respect to their expression responses in *S. paradoxus* N17 ($\Delta$hap4). 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($\Delta$hap4); triangle: up-regulated in BY4741($\Delta$hap4))

(H): The rate of conservation in the three related yeasts for the 65 clustered effects and 130 distributed effects defined in BY4741($\Delta$hap4), 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 $\Delta$hap2, $\Delta$hap3, and $\Delta$hap5, 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($\Delta$hap4) 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 genetics. Statistically significant genetic effects defined in conventional framework are further separated into evolutionarily selected and unselected, with the former supporting related biochemistry understandings and the latter accounting for pleiotropy.
Fig. 1
Fig. 2

BY4741 Δhap2

BY4741 Δhap3

BY4741 Δhap5

A

B

C

D

E

F

Chromatin immuno-precipitation assay

Expectation

Observation

COR1

COX1

INH1

TIM11

ATP17

GCR7

RIP1

COX13

COX5A

ATP19

CYT1

ATP15

Rate of direct targets

overlapped

clustered

distributed

all genes

10/20

12/65

1/130

33/146
Fig. 3
Fig. 4

An expanded framework for reverse genetics

| Genetic mutation | Statistical test | Evolutionary decomposition |
|------------------|------------------|---------------------------|
| Mutant v.s. wildtype differences | Significant effects | Selected effects |
| | Insignificant effects | Unselected effects |
| | | Match biochemistry data |
| | | Underlie cryptic pleiotropy |

Conventional part | Expanded part
Supplementary Information of
“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

This file contains:
Legends of Tables S1 to S8
Figs. S1 to S8

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 heritability scores 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(Δ hap4), respectively. The horizontal dashed line shows adjusted P = 0.05 and vertical dashed lines show log2FC = ±0.58. (cyan: clustered effects; red: distributed effects; circle: 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)
Fig. S9. 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, KOA 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.
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 HAP4 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 22 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:soe00900) and 551.7 (PATH:soe04122), 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 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.