Genetic properties underlying transcriptional variability in *Escherichia coli*

Saburo Tsuru[1,*] & Chikara Furusawa[1,2,3,*]

[1]Universal Biology Institute, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2]Department of Physics, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[3]Center for Biosystems Dynamics Research (BDR), RIKEN, 6-2-3 Furuedai, Suita, Osaka 565-0874, Japan

*Corresponding authors:
Saburo Tsuru
E-mail: saburotsuru@gmail.com
Chikara Furusawa
chikara.furusawa@riken.jp

Abstract

The rate and direction of phenotypic evolution depend on the availability of phenotypic variants induced genetically or environmentally for selection. It is widely accepted that organisms do not display uniform phenotypic variation, where certain variants arise more frequently than others in response to genetic or environmental perturbations. Previous studies have suggested that gene regulatory networks governing phenotypes may channel both environmental and genetic influences. However, the properties of gene regulation that determine the rate and direction of phenotypic variation remain unknown. To address this issue, we characterized transcriptional variations in *Escherichia coli* in response to environmental and genetic perturbations. Based on the current understanding of transcriptional regulatory networks, we identified common genetic properties that account for gene-to-gene differences in the magnitude and directionality of transcriptional variation in both environmental and genetic causes. Our findings support the relevance of gene regulatory networks in shaping the phenotypic variability shared by different perturbations. This study provides insights into ability of gene regulatory network to learn about past environments in order to generalize to new environments.

Main

The rate and direction of phenotypic evolution depend not only on natural selection but also on the availability of phenotypic variants, whether induced genetically or environmentally, upon which selection may act[1]. Even environmentally induced phenotypic variability or plasticity has the potential to initiate and direct adaptive evolution through processes such as genetic assimilation[2]. It is widely acknowledged that organisms exhibit heterogeneous phenotypic variation, with certain variants emerging more frequently than others in response to genetic or environmental perturbations[3-6]. This bias in phenotypic variability can facilitate and constrain adaptive evolution across different phenotypic dimensions[7], contingent on the alignment between the bias in phenotypic variability and the direction of selection[8]. Thus, without a thorough understanding of the nature of phenotypic variability, our understanding of why evolution has progressed in the manner that it does or why some phenotypes evolve repeatedly while others are one-offs remains incomplete.

Bias in phenotypic variability against different perturbations may result from a common mechanism. Theoretical studies demonstrated that the bias in phenotypic variability is shared by different perturbations[9-11]. Consistently, environmentally induced phenotypic variations often mirror genetically induced phenotypic variations. Notably, there is similarity in the magnitude of variance, where traits sensitive to environmental perturbations tend to be sensitive to genetic perturbations[12-14]. Additionally, there is a consistent pattern of directional variation among traits, with organisms frequently exhibiting correlated variation between traits[15]. Environmental influences on traits often act in the same direction as genetic influences, shaping phenotypic variation in a consistent direction[6]. These similarities imply the existence of a common mechanism underlying the bias in phenotypic variability against any kinds of perturbations[16,17].

What molecular mechanisms bias the phenotypic variability and how? In biological systems, most traits arise from genetic interactions within networks represented by gene regulatory networks[18]. Genetic and environmental perturbations exert phenotypic effects via identical regulatory networks[19]. Consequently, phenotypic variance is expected to be influenced by the properties of the regulatory networks, regardless of the underlying causes of perturbations[8,10]. Despite a growing number of studies demonstrating empirical evidence supporting the concordance between environmentally induced and genetically induced phenotypic variations, it remains largely unknown how the real regulatory networks account for the observed bias[8].

To address this issue, we characterized the variability of transcriptome profiles in *Escherichia coli* in response to different environmental and genetic perturbations. *E. coli* is an ideal model organism for this investigation because of its well-characterized gene regulatory network. Our analysis revealed that genes showing higher transcriptional variability against environmental perturbations tended to have higher sensitivity to genetic perturbations. In addition, the major directions, or the top principal components, of correlated variation among thousands of genes were shared by different perturbations. Most importantly, we demonstrate that the structure of transcriptional regulation accounts for these

common biases. These empirical findings support the relevance of the gene regulatory networks in explaining the bias in phenotypic variability shared across different perturbations. This study sheds light on how gene regulatory networks learn about past environments in order to generalize to new environments.

Results and Discussion

Similarity in transcriptional variance between environmental and genetic perturbations

We initially analyzed two public transcriptome datasets to evaluate the transcriptional variability across different perturbations. To this end, we curated transcriptome profiles from the PRECISE database[20,21], focusing on a single strain of *E. coli*, K12 MG1655, cultured under different environmental conditions (160 profiles, 76 unique environmental conditions) (Supplementary table 1), termed the Env dataset (Fig. 1a). Additionally, to evaluate transcriptional variability in response to genetic perturbations via natural evolution, we used transcriptome profiles from 16 natural strains of *E. coli* cultured under an identical environmental condition[22], termed the Evo dataset (Fig. 1b). These isolates spanned the most common phylogroups and lifestyles of *E. coli*. Importantly, genetic mutations among these strains might reflect their diverse natural habitats, potentially confounding the estimation of transcriptional variability in response to genetic perturbations independent of environmental perturbations. To mitigate this, we conducted a mutation accumulation (MA) experiment[23] using a mutator strain of *E. coli*, K12 MG1655, lacking the *dnaQ* gene[24,25] (Fig. 1c). Here, mutations primarily accumulate through genetic drift owing to small population bottlenecks, thus minimizing positive selection. We established two founder colonies (founders A and B) and propagated single colonies on lysogeny broth (LB) agar plates (yielding three MA lineages per founder). Consequently, we obtained transcriptome profiles from six MA lineages cultured in LB broth using the reliable DNA microarray technique[26] (Supplementary Fig. 1), termed the Mut dataset (Supplementary tables 2,3). Importantly, each dataset was generated within single laboratories using a common protocol, minimizing technical variation within each dataset. We used quantile-normalized and $\log_2$-transformed transcriptome data for the analysis.

The means and standard deviations of expression levels were computed for each gene across each dataset (Fig. 1d for six representative genes). While standard deviation might directly measure transcriptional variability, it can be influenced by the mean[27], leading to a confounding effect in estimating similarities in transcriptional variability between datasets. Notably, the standard deviations exhibited a statistically significant mean dependency for all datasets (Fig. 1e). To alleviate this influence, we calculated the vertical deviation of each standard deviation from a smoothed spline of the running median of the standard deviations (DM), which served as a measure of the magnitude of transcriptional variability. The DM values of each gene in the dataset was designated as DM_{env}, DM_{evo}, DM_{mut} for the Env, Evo, and Mut datasets, respectively (Fig. 1f, Supplementary table 4).

Comparing DM values between datasets revealed positive correlations (Spearman's rank correlation coefficient (R) =0.43–0.56, $p<0.05$) (Fig. 2a). These results indicate that genes with higher sensitivity to environmental perturbations tend to have higher sensitivity to genetic perturbations in terms of their expression levels. This phenomenon of canalization/decanalization could occur not only at the gene level but also at more complex phenotypes subject to natural selection. Canalization has been observed in the functional categories of genes in yeast[28]. To explore this possibility in bacteria,

we utilized a refined subset of Gene Ontology (GO) terms as functional categories (**Methods**) (Supplementary table 5), using the mean DM values of genes within each GO term as a measure of transcriptional variability. Consequently, we observed positive correlations between the datasets even at the functional category level (Fig. 2b). These results indicate frequent associations between canalization/decanalization of genes and their functions.

Using gene set enrichment analysis (GSEA[29]) with GO terms for both higher and lower DM values, we identified several common GO terms across all datasets (Fig. 2c, Supplementary Fig. 2). Notably, flagellum-dependent cell motility and amino acid catabolic processes were consistently enriched with higher DM values in all three datasets. In contrast, biological processes related to cell-wall construction and maintenance were consistently enriched at lower DM values, indicating that the expression levels of genes within these categories were relatively robust to perturbations. We obtained relevant results using the functional categories defined by the KEGG pathways (Supplementary Fig. 3,4, Supplementary table 6). These results suggest a certain common mechanism underlying the heterogeneous transcriptional variability associated with functional categories.

Transcriptional regulators influencing transcriptional variability

To explore the underlying mechanism, we examined the influence of transcriptional regulators on transcriptional variability. First, we conducted GSEA based on known regulatory interactions (Supplementary table 7) to identify transcriptional regulators that affect target genes exhibiting higher and lower DM values, termed TR_{high} and TR_{low}, respectively. We analyzed two golden standard databases, RegulonDB[30] and EcoCyc[31], collecting known regulatory interactions between genes. We focused on the transcription factors and sigma factors that regulate more than four target genes. For the Env dataset, we identified 30 transcriptional regulators enriched in higher DM_{env}, such as FlhDC and Fur, and termed them TR_{env} (Fig. 3a). In contrast, we identified only four transcriptional regulators enriched in the lower DM_{env}. Similar analyses were performed for DM_{evo} and DM_{mut} (Supplementary Fig. 5). Consequently, we identified 29 (TR_{evo}) and three transcriptional regulators enriched in higher and lower DM_{evo}, respectively. Similarly, 28 (TR_{mut}) and two transcriptional regulators were enriched in the higher and lower DM_{mut}, respectively. Thus, fewer transcriptional regulators were enriched at lower DM values, suggesting that regulation by transcriptional regulators tends to increase the transcriptional variability of target genes, with some regulatory motifs, such as negative feedback, potentially repressing fluctuations in gene expression, as a minor exception[32].

Next, we explored the similarity between TR_{env}, TR_{evo}, and TR_{mut}. We found that 13 transcriptional regulators, referred to as TR_{com}, were common among the three datasets, corresponding to more than one-third of the transcriptional regulators enriched in each dataset (Fig. 3b). These results suggest that the target genes regulated by TR_{com} exhibited higher transcriptional variability in response to both environmental and genetic perturbations, indicating that TR_{com}s function as a common mechanism to bias transcriptional variability against various perturbations.

Network signatures of transcriptional regulators influencing transcriptional variability

How do TR_{com}s enhance transcriptional variability of target genes against a variety of perturbations both inside and outside of the cells? We found that TR_{com}s are global regulators with a large number of target genes, including TR_{com}s themselves (Fig. 3c). Based on these findings, we speculated that TR_{com}s might regulate each other frequently, which could increase the probability of the propagation of different perturbations to their target genes. To test this hypothesis, we explored the regulatory structure of a network comprising TR_{com}s alone (Fig. 3d). We found that most TR_{com}s had both an in-degree and an out-degree (11 out of 13 regulators). FliA had no out-degrees but four in-degrees. Only IscR was isolated with no regulatory interactions within TR_{com}s. This network structure indicates that the network within TR_{com}s is not top-down structured but has many mutual interactions (activations and inhibitions) within TR_{com}s. The edge density within the TR_{com}s (Fig. 3e) was remarkably high (0.17), with statistical significance ($p<0.0001$), compared to the edge density of random expectations (0.019). These results indicate that the TR_{com}s form a relatively densely connected network that is proficient in propagating the influence of perturbations within this group.

Next, we explored the effects of TR_{com}s on transcriptional variability. First, we compared the DM values of the target genes with TR_{com}-containing regulation with those of the target genes with TR_{com}-less regulation (Fig. 3f). The target genes of TR_{com}s tended to have higher DM values in all datasets than other genes with the same number of unique regulators. We found none or very weak positive correlations between the DM values of the genes without regulations by TR_{com}s and the number of unique regulators that the genes received in all datasets (Spearman's $R=\sim0.13$). On the other hand, the target genes of TR_{com}s exhibited relatively high positive correlations in all datasets (Spearman's $R=0.23–0.35$, $p<0.05$). These results indicate that the recruitment of TR_{com}s is more effective in enhancing transcriptional variability. To further investigate this hypothesis, we explored the relationship between DM values and the number of unique TR_{com}s among the genes with a fixed number of regulators. In the case of the genes that received five unique regulators (Fig. 3g), we found a positive correlation between DM_{env} and the number of unique TR_{com}s among the five regulators (Spearman's $R=0.29$, $p<0.05$). We performed similar analyses for all the datasets and found pervasive positive correlations across broad classes of the number of unique regulators (Fig. 3h). These results highlight the role of TR_{com}s in enhancing transcriptional variability through dense mutual regulation within TR_{com}s and co-regulation by multiple TR_{com}s, likely because both increase *trans*-target sizes[33] that affect the expression levels of the genes regulated by TR_{com}s in response to both environmental and genetic perturbations.

To understand how TR_{com}s access a variety of perturbations, we explored the relationship between TR_{com}s and the known molecular machinery for sensing stimuli both inside and outside the cells. First, we performed bow-tie decomposition of the TRN to identify the subnetworks to which the TR_{com}s belonged (Fig. 4a–d, Supplementary table 8). The *E. coli* TRN was formally dissected into seven sectors defined by Yang's bow-tie structure[34] based on the directed regulatory links between

genes[35]. The primary sectors are IN, OUT and the greatest Strongly Connected Component (SCC). The SCC sector represents the subnetwork, in which each node (TR) is reachable by any other node following the direction of regulation from the TR to its target. The IN sector comprises TRs that are not included in the SCC and can reach TRs in the SCC. The OUT sector comprises nodes (TRs or non-regulatory genes) not included in SCC but reachable by TRs in SCC. The numbers of nodes in these primary sectors in *E. coli* TRN were seven TRs, 62 TRs and 2715 genes including the TRs for IN, SCC, and OUT, respectively (Fig. 4e). Interestingly, TR_{com}s (12 of 13 TR_{com}s) and TR_{high}s (26 of 48 TR_{high}s in Env, Evo, and Mut) were significantly enriched in SCC (Hypergeometric test, $q < 0.05$, Fig. 4d). This implies that TR_{com}s can receive broad perturbations from approximately 70 genes in the IN and SCC sectors. We noted that TR_{com}s in SCC still showed a higher edge density (0.19) than other TRs in SCC ($p<0.005$), supporting the especially dense connections within the TR_{com}s.

Next, we investigated how TR_{com}s sense different intracellular and extracellular stimuli. *E. coli* employs at least 145 sensory regulators to transmit various stimuli to the TRN (Fig. 4b, Supplementary table 8). These regulators, termed SRs, change their regulatory activity depending on their interactions with corresponding effectors, as summarized by Martínez-Antonio *et al*[36]. For visualization, we defined the STIMULI block as an aggregated group of the known effectors that interact with SRs, such as nutrients, stressors and metabolites. We found that the IN and SCC sectors contained 39 SRs (27% of the total SRs, Fig. 4f), which was reasonable agreement with the random expectation based on the number of TRs in these sectors. Interestingly, more than half of TR_{com}s (8 of 13) were SRs. These results indicate that TRs in the SCC sectors, TR_{com}s included, not only serve as intermediate managers in information processing but also play a role in sensing different kinds of stimuli. These characteristics of TR_{com}s, particularly their presence in SCC with many SRs, may enable TR_{com}s to access a variety of stimuli both inside and outside the cells.

Similarity in direction of transcriptional variance between environmental and genetic perturbations

Next, we explored the major directions of correlated transcriptional variation among thousands of genes. Theoretically, genetically caused phenotypic variance is expected to be the greatest along the major axis or the principal component of environmentally caused phenotypic variance[10]. To identify such principal components (PC) and test this prediction, we performed principal component analysis (PCA) for each dataset (Fig. 5a, Supplementary Figs. 6–8). The first three principal components (PC1–3) explained 46%, 53%, and 57% of the total variance in the Env, Evo, and Mut datasets, respectively, indicating their significance as the major axes of transcriptional variation for each dataset. Correlation analysis based on the loadings of these major axes revealed similarities in their directions across the datasets (Fig. 5b). We used the absolute values of Spearman's R because of the relative nature of the signs of the principal components. Specifically, PC1 in the Env dataset exhibited relatively strong correlations with PC1 and PC2 in the Evo dataset (Spearman's $R = 0.45$ and 0.41, respectively, $p<0.05$) and with PC1 and PC2 in the Mut dataset (Spearman's $R = 0.34$ and 0.41, respectively, $p<0.05$).

Similarly, PC1 in the Evo dataset showed a relatively strong correlations with PC1–3 in the Env and Mut datasets. In addition, we assessed the extent to which the principal components in the Env dataset explained the transcriptional variance in the other two datasets. Overall, all principal components in the Env dataset explained up to 37% and 39% of the total variance in the Evo and Mut datasets, respectively (Fig. 5c). Moreover, the cumulative variance plot displayed a convex upward curve for both the Evo and Mut datasets. These findings support the notion that the major directions of transcriptional variation in response to environmental perturbations are considerably similar to those observed in response to genetic perturbations, despite a large number of genes.

A directional bias in transcriptional variations is expected to reflect transcriptional regulations[8]. To validate this prediction, we explored the transcriptional regulators that contribute to the major directions. Utilizing the loadings of PC1–3 in each dataset, we conducted GSEA (Fig. 5d,e). We found the transcriptional regulations enriched in the positive and negative directions of PC1–3 in the Env dataset (Fig. 5d). For instance, activation by FadR and FlhDC or inhibition by DksA and PdhR was enriched in the positive direction of PC1. Conversely, activation by RcdA and IscR or inhibition by CecR and MarA was enriched in the negative direction of PC1. We conducted a similar GSEA for the other two datasets (Supplementary Fig. 9), identifying 31 regulations common across all three datasets (Fig. 5e), including 11 regulations by TR_{coms}, which is consistent with the notion that TR_{com}s are global regulators collectively controlling numerous targets. Overall, these findings support the idea that the similarity in the direction of transcriptional changes in response to different types of perturbations arises from shared transcriptional regulation.

Furthermore, functional enrichment in GO terms was observed in the top three PCs across all three datasets (Fig. 5f, Supplementary Figs. 10–12). The most major direction in Env (PC1) revealed enrichment of translation, cell division, and cell-wall-related biological processes in one direction (positive sign), and responses to starvation and amino acid catabolic/biosynthetic processes in the opposite direction (negative sign). We termed these opposing functional families growth and starvation families, respectively, because of their close association with cellular growth and starvation states. The correlation between PC1 in Env and PC1 and PC2 in Evo and Mut suggested that approximately half to one-thirds of the functional categories in both families were enriched in these PCs, with opposite directions between the families retained across datasets. Importantly, the mutually opposite directionality observed in these functional families tended to be preserved across different perturbations, suggesting a consistent response pattern at the functional category level (Fig. 5g).

Similarity of phenotypic variability across various perturbations

Heterogeneous phenotypic variability may result from the generalization of past environments[9-11]. In addition, several studies suggest that evolution tends to produce global canalization that buffers phenotypes against not only genetic and environmental perturbations[17], but also perturbations caused by stochastic molecular noise[37,38]. These findings imply a similarity in phenotypic variability across the different scenarios. To address this, we analyzed five additional datasets for *E. coli*, termed Strs_hp,

Strs_he, Strs, Prtn, and Noise_ch (Fig. 6a, Supplementary Figs. 13–21, Supplementary information Appendices I,II). The phenotypic variabilities of the datasets, either at the transcriptional or protein level, were termed DM_{strs_hp}, DM_{strs_he}, DM_{strs}, and DM_{noise_ch}, respectively. In short, DM_{strs_hp} and DM_{prtn} show variability in their expression levels (mRNA level for Strs_hp and protein level for Prtn) in response to various environmental perturbations. DM_{strs_he} and DM_{strs} represent transcriptional variability in response to genetic mutations accumulated through adaptive laboratory evolution under various stressors. DM_{noise_ch} represents the protein noise of clonal populations under identical environmental and genetic conditions. We found that these DM values were positively correlated with those of the three main datasets (Env, Evo, and Mut) (Fig. 6b). Notably, the correlations with DM_{env} show relatively strong in all pairs (Spearman's $R = 0.31–0.68$, $p<0.05$). Next, we conducted PCA to identify the directionality of each dataset. To understand the similarity in the major directions across the datasets, we compared PC1–3 of these additional datasets with those of the three main datasets (Fig. 6c). Correlation analysis using the loadings of PC1–3 revealed that PC1s of the additional datasets tended to correlate with PC1s or PC2s of the three main datasets. In particular, PC1 in the Env dataset showed the strongest correlation with the PC1s of the additional datasets. These results indicated similar trends in gene-to-gene differences in both the magnitude and major directions of expressional variations across various perturbations, suggesting global canalization/decanalization across different perturbations[17,37]. These results also imply that the heterogeneous transcriptional variability can influence the rates and directions of evolution not only through neutral processes, as shown in the MA experiment, but also through complex processes, including adaptive processes (Supplementary information Appendices II,III).

Conclusions

In this study, we characterized the transcriptional variability of *E. coli* against different perturbations and explored the regulatory commonalities between genetic and environmental perturbations. We identified common global transcriptional regulators that underlie the shared variability across different perturbations. These transcriptional regulators interact relatively densely with each other and strongly connect with many sensory regulators, forming a network structure that is favorable to facilitate the frequency of propagation of various stimuli both inside and outside the cells. In addition, the more regulated genes by these influential regulators showed greater changes in expression effectively. PCA revealed that these global transcriptional regulators simultaneously induced changes in the expression levels of multiple genes and contributed significantly to the formation of major directions of genome-wide changes in expression levels. Although our analysis may have been conducted using limited methods and focused only on the network structure of transcriptional regulators, our results provide implications for the molecular mechanisms behind non-uniform phenotypic variability.

Methods

Bacterial strains and media

We constructed a mutator strain of *E. coli* MG1655 by deleting the *dnaQ* gene. Deletion was performed using conventional λ RED recombination[39]. In brief, a kanamycin resistance gene flanked by FLP recognition sites was amplified from *E. coli* BW25113 *ΔmoaA::kan*, a strain of the Keio collection[40], by PCR with chimeric primers containing homology arms for the *dnaQ* region. We used the same chimeric primers as those used to knock out *dnaQ* in a previous study[40]. The PCR fragment was used for lambda RED-mediated homologous recombination with pKD46. The obtained strain after obviating pKD46 was called as MG1655 *ΔdnaQ* and used for the MA experiment. We used YENB broth[41] supplemented with 60 mM arabinose (FUJIFILM Wako, Osaka, Japan, 012-04581) and 50 μg/mL ampicillin (FUJIFILM Wako, Osaka, Japan, 012-23303) to prepare electrocompetent cells for λ RED recombination. LB agar plates (Miller, BD Difco, 244620) supplemented with 25 μg/mL kanamycin (FUJIFILM Wako, Osaka, Japan, 115-00342) were used as the selective media after electroporation. We used LB agar and broth for the MA experiments and transcriptome analysis of the mutants, respectively.

MA experiments

The MA experiments were performed as previous studies[25]. In brief, we used two independent colonies of MG1655 *ΔdnaQ* as founder strains, founder A and B, to alleviate a potential bias of the mutations before the MA procedure that might affect the accumulated mutations of the descendants during the MA procedures[42]. We initially established six independent lineages for each founder in preparation for extinction during MA experiments. The MA experiments were conducted by transferring randomly chosen single colonies to fresh LB agar plates daily. The plates were incubated at 32 °C. The MA procedure was repeated for 30 times. The MA lineages after the 10th and final rounds were stored at –80 °C. Three randomly selected three MA lineages from each founder were subjected to transcriptome analyses (Supplementary table 2).

Transcriptome profiling using microarray technology

The detailed procedure for transcriptome profiling has been described in previous study[26]. All strains obtained by the MA experiments were grown in 200 μL of LB medium in 96-well microplates at 32 °C. The cell density was monitored using a 1420 ARVO microplate reader (PerkinElmer Inc., Waltham, Massachusetts, USA). The cell broth was sampled at the exponential phase, when OD_{600} reached in the 0.072–0.135 range. The cell broth was immediately added to an equal volume of ice-cold ethanol containing 10% (w/v) phenol. The cell pellets were obtained by centrifugation at 20,000 × g at 4 °C for 5 min. Total RNA was extracted using an RNeasy Mini kit with on-column DNase I digestion for 15 min (Qiagen, Hilden, Germany) according to the manufacturer's protocol. Purified RNA samples were analyzed for quality control using an Agilent 2100 Bioanalyzer and an RNA 6000 Nano kit (Agilent Technologies). Microarray experiments were performed using a custom-designed Agilent

330 8 × 60 K array for the *E. coli* W3110 strain, comprising 12 probes for each gene. Purified total RNA
331 samples (100 ng) were labeled with Cyanine3 (Cy3) using a Low Input Quick Amp WT labeling kit
332 (One-color; Agilent Technologies). The Cy3-labeled cRNA samples were checked for amount
333 (>825 ng) and specific activity (>15 pmol/μg) using a NanoDrop ND-2000 (Thermo Scientific).
334 Subsequently, the labeled cRNA (600 ng) was fragmented and hybridized to a microarray for 17 h
335 while rotating at 10 rpm at 65 °C in a hybridization oven (Agilent Technologies). The hybridized
336 microarray was washed and scanned according to the manufacturer's instructions. Microarray images
337 were analyzed using Feature Extraction version 10.7.3.1 (Agilent Technologies). Duplicate biological
338 data of the founder strains under the same culture conditions were obtained to confirm technical noise.
339 The compendium of transcriptome data obtained from the MA experiment is called the Mut dataset.
340

341 **Compilation of the TRN in *E. coli***

342 We compiled the reported TRN using all interactions with experimental evidence from RegulonDB
343 11.0[30] and EcoCyc[31] for both transcription factors and sigma factors (Supplementary table 7). For
344 transcription factors, the mode of regulation (activation or repression) was also included. We set the
345 mode of regulation by the sigma factor to activation. When multiple modes were reported for a given
346 regulatory mechanism between a transcription factor and its target gene, they were designated as
347 unknown.
348

349 **Analysis of network architecture of TRN using bow-tie decomposition**

350 First, we analyzed the architecture of TRN in *E. coli* based on bow-tie decomposition[35]. We followed
351 the formal definition of the bow-tie structures proposed by Yang *et al*[34]. Their bow-tie structure has
352 seven sectors. The three main sectors are IN, OUT, and SCC. The SCC sector consists of the largest
353 subnetwork in which each node (gene or transcriptional regulator) is reachable by any other node,
354 following the direction of regulation (activation and/or inhibition). The IN sector represents a group
355 of nodes that are not included in the SCC but can reach nodes in the SCC. The OUT sector represents
356 a group of nodes not included in the SCC that can be reached by the nodes in the SCC. The remaining
357 nodes, other than those in the three main sectors, can be further categorized into TUBES,
358 INTENDRILS, OUTTENDRILS, and OTHERS, according to reachability.

359 To understand how the internal and external perturbations of the cells propagate TRN, we
360 embedded links between the internal/external effectors and TRs, as follows. In addition to the bow-tie
361 structure, we also defined the STIMULI block. The STIMULI block does not consist of a gene or
362 regulator, but consists of intracellular and extracellular effectors, such as nutrients, stressors,
363 metabolites, and stimulating transcriptional regulators by allosteric interactions and so forth. We used
364 the compendium of known links between effectors and TRs collected by Martínez-Antonio *et al*[36].
365 These target regulators are defined as SR. SRs change their regulatory activity depending on whether
366 they bind to effectors. The networks obtained are listed in Supplementary table 8.

367 **Data analysis of the three main datasets (Env, Evo and Mut)**
368 In addition to the Mut dataset, we analyzed two datasets of transcriptome in *E. coli*. To estimate the
369 variability in the expression levels of *E. coli* MG1655 in response to environmental changes, we used
370 an updated version of PRECISE[20,21], which contains the RNA-seq compendium of *E. coli* MG1655
371 obtained from a single laboratory using a standardized protocol. In particular, the transcriptome
372 profiles of PRECISE 2.0 were downloaded from iModulonDB[43]. We filtered the transcriptome data of
373 any mutant or evolved strains derived from the MG1655 strain or from other strains to avoid mutational
374 effects on transcriptional variance. Consequently, 160 profiles under 76 unique environmental
375 conditions were used (Supplementary table 1). The environmental conditions were mainly
376 combinations of seven base media (LB, M9, W2, CAMHB, Tris-maleic acid minimal medium, M9
377 without phosphate, RPMI supplemented with 10% LB), 12 carbon sources (acetate, D-ribose, fructose,
378 galactose, glucarate, gluconate, glucose, glycerol, N-acetylglucosamine, pyruvate, sorbitol, xylose),
379 three nitrogen sources (NH_4Cl, cytosine, glutamine), two electron donors (O_2, KNO_3), three trace
380 elements (aebersold trace element mixture, sauer trace element mixture with or without $MgSO_4$), 35
381 supplements (cytidine, thiamine, leucine, tyrosine, HCl, lactic acid, adenosine, 2, 2′-dipyridyl, $FeCl_2$,
382 NaCl, paraquat, glutathione, adenine, dibucaine, methionine, salicylate, acetoacetate/LiCl, H_2O_2, LiCl,
383 ceftriaxone, ciprofloxacin, glutamate, glycine, L-tryptophan, meropenem, threonine, trimethoprim-
384 sulfamethoxazole, KCl, phenylalanine, uracil, L-arginine, pyruvate, $CuSO_4$, ethanol, $ZnCl_2$), two
385 temperatures (37 °C, 42 °C), three pHs (5.5, 7, 8.5) and two culture types(batch, chemostat). The
386 filtered compendium was called the Env dataset. To estimate the variability in expression levels during
387 the evolution of *E. coli* in nature, we analyzed the RNA-seq compendium of natural isolates of *E. coli*
388 during growth in the exponential phase in LB broth, where the RNA-seq profiles were obtained in a
389 single laboratory using a common protocol[22]. We analyzed the RNA-seq profiles of 16 natural isolates
390 that covered all five primary phylogroups of *E. coli*. The compendium of this dataset was called the
391 Evo dataset.

392 We used 2622 genes that were common to the three datasets (Env, Evo, and Mut)
393 (Supplementary table 3). For all three datasets, the $\log_2$-transformed gene expression levels were used
394 in subsequent analyses. Expression levels were subsequently quantile-normalized. The means and
395 standard deviations of the expression levels of each gene were calculated for each dataset. For the Mut
396 dataset, we first calculated the means and variances of expression levels among the MA lineages of
397 the final rounds, which were derived from the same founder. Subsequently, we averaged the mean and
398 variance of the MA lineages derived from different founders. We obtained the means and standard
399 deviations of the expression levels of MA lineages that underwent 30 rounds of the MA procedure. We
400 found that the standard deviations were dependent on the mean expression levels for every dataset,
401 which can confer a confounding effect on the correlation analysis between the datasets. To compensate
402 for this dependency, we calculated the vertical distance of each standard deviation from a smoothed
403 running median of standard deviations, referred to as the DM, following the method outlined in a
404 previous study[44,45]. For each gene, we calculated the DM for each dataset and named it DM_{env}, DM_{evo}

and DM_{mut} for the Env, Evo, and Mut datasets, respectively. We also calculated variability at the functional category level based on the DM values. Functional categories were defined by Gene Ontologies (GO)[46,47]. We constructed a subset (82) of GO by removing redundant GO terms based on semantic similarity among GO terms using the GOSemSim[48], AnnotationDbi[49], rrvgo[50] R packages with default parameters. Alternatively, we used the KEGG pathway database[51]. GO terms and KEGG pathways that consisted of fewer than five genes or more than 300 genes were omitted. The mean DM values among the genes belonging to a given GO term or KEGG pathway were used as a measure of variability at the functional category level (Supplementary tables 5,6).

Gene set enrichment analysis (GSEA)[29] was performed using the clusterprofiler R package[52]. We used the refined subset of GO terms and KEGG pathways as functional categories. We filtered out the GO terms and KEGG pathways that consisted of fewer than five genes or more than 300 genes. For GSEA of transcriptional regulators, we also filtered out the transcriptional regulators that regulated fewer than five target genes or 1000 genes. The q-values were calculated using the Benjamini-Hochberg (BH) procedure[53]. Transcriptional regulators whose target genes were significantly enriched at higher DM in each dataset were called TR_{env}, TR_{evo} and TR_{mut} for the Env, Evo, and Mut datasets, respectively. The 13 transcriptional regulators present in all three groups were called TR_{com}.

The edge density of a given gene set belonging to TRN was calculated as the total number of edges (i.e. regulations) within the gene set divided by the total number of possible edges within the gene set. Self-loops (i.e. autoregulations) were ignored. To compare the edge density within TR_{com}s with that within randomly chosen regulators, we generated 10,000 subsets by drawing the same number of TR_{com}s (i.e. 13) transcriptional regulators at random from a pool of transcriptional regulators. The transcriptional regulators that regulated fewer than five target genes were omitted.

PCA was conducted using the prcomp function in R. For the Mut dataset, we used all transcriptional profiles, including isolates at rounds 10 and 30, as well as the founders. In PCA, the *k*-th principal component, PC_k, is represented as a linear combination of loadings *a*, and expression levels *x*, as follows:

$$PC_{k,D} = \sum_{i}^{N} a_{k,i,D} x_{i,D}$$

where *D*, *N* and *i* represent the datasets (ex. Env), total number of genes, and *i*-th gene, respectively. Correlation analysis was performed for the loadings of the top three PCs, that is, PC1–3 between the datasets. The correlation was evaluated using absolute values of the Spearman's rank correlation coefficients (R), as the signs of the PCs are arbitrary.

Data analysis of the other omics datasets (Strs_hp, Strs_he, Strs, Prtn and Noise_ch)

The Strs_hp and Strs_he datasets were obtained from transcriptome profiles reported by Horinouchi *et al*[54]. In Strs_hp, *E. coli* MDS42[55], a derivative of MG1655, was cultured under different stress conditions, and transcriptome profiles were obtained using the DNA microarray technique

(Supplementary Fig. 15a). The transcriptome profiles obtained from the same stressors used in the Env dataset were removed. Transcriptome profiles for eight stressful conditions (cobalt chloride, sodium carbonate, l-malate, methacrylate, crotonate, methylglyoxal, *n*-butanol and cetylpyridinium chloride) and one stress-free condition (modified M9 medium) were analyzed (Supplementary table 9). Thus, the difference in transcriptome profiles within the obtained dataset, called the Strs_hp dataset, represents transcriptional variability in response to environmental perturbations. For Strs_he, transcriptome profiles were obtained for mutants resistant to each of the different stressors cultured under stressor-free conditions using DNA microarrays (Supplementary Fig. 15b). These resistant mutants were obtained through adaptive laboratory evolution experiments using MDS42 as the ancestral strain (Supplementary Fig. 15c). We removed the transcriptional profiles obtained from the mutants that experienced the same stressors used in the Env dataset. We analyzed the transcriptional profiles of the mutants obtained under the eight stress conditions described above (Supplementary table 9). Thus, the difference in transcriptional profiles within the resultant dataset, called the Strs_he dataset, reflected the transcriptional variability against genetic mutations accumulated through adaptive evolution. Both datasets contained expression levels of 3741 genes (Supplementary Fig. 15d).

The Strs dataset was obtained from the transcriptome profiles reported by Maeda and Iwasawa *et al*[26]. In the Strs dataset, the different MDS42 mutants isolated from adaptive laboratory evolution experiments in the presence of different stressors were subjected to DNA microarray analysis (Supplementary Fig. 18a–c). The evolutionary experiment was aimed at obtaining resistance mutants against each of the different stressors using MDS42 as an ancestral strain. The transcriptome profiles of the mutant and ancestral strains were obtained under stressor-free conditions, similar to those of Strs_hp. We removed the transcriptome profiles of mutants that evolved under the same stressors used in the Env dataset (example. hydrogen peroxide, meropenem). We analyzed the transcriptional profiles of the mutants obtained under the 47 stress conditions as detailed in Supplementary table 10. Thus, the difference in transcriptome profiles within the obtained dataset, called the Strs dataset, contained transcriptional variability against genetic mutations that accumulated through adaptive evolution. In addition to the transcriptome profiles of the resistant mutants, Maeda and Iwasawa *et al*.[26] identified several genes that were commonly mutated in multiple evolved lines among independently evolved lines under the same stress conditions. Among these commonly mutated genes, we analyzed those encoding transcriptional regulators, called commonly mutated transcriptional regulators, as listed in Supplementary table 11.

For the Strs_hp, Strs_he and Strs datasets, the mRNA expression levels obtained using DNA microarrays were $\log_2$-transformed and quantile-normalized, similar to the Mut dataset.

The Prtn dataset was obtained from proteomic profiles reported by Schmidt *et al*[56]. Proteomic profiles with the growth rate were determined for *E. coli* BW25113 cultured under 22 different environmental conditions (including different carbon sources). The difference in proteomic profiles within the resultant dataset, called the Prtn dataset, represented the proteomic variability (2002

proteins) in response to environmental perturbations (Supplementary Fig. 13). The absolute abundance of proteins (in copies/cell) was $\log_2$-transformed.

For all four datasets (Strs_hp, Strs_he, Strs, and Prtn), the means and standard deviations of the expression levels of each gene were calculated across either the mutant or environment in each dataset (Supplementary Figs. 13,15,18). All datasets showed that the standard deviations depended on the mean expression levels. To compensate for this dependency, we calculated the DM values for each gene as the vertical distance of each standard deviation from a smoothed running median of the standard deviations for each dataset as detailed above. Consequently, we obtained DM_{strs_hp}, DM_{strs_he}, DM_{strs} and DM_{prtn} for Strs_hp, Strs_he, Strs, and Prtn, respectively. PCA was performed as described previously (Supplementary Figs. 14,17,19).

The Noise_ch dataset was obtained from the cell-to-cell heterogeneity of protein expression levels reported by Taniguchi *et al*[57] (Supplementary Fig. 20). Briefly, a library of yellow fluorescent protein (YFP) fusion strains of *E. coli* K12 was cultured under identical conditions, and each chromosomal copy of the gene was labeled with YFP fusion. YFP fluorescence was measured using microscopy at the single-cell level, which yielded the fluorescence distribution for each gene (1018 genes). For each gene, the protein noise was defined as the variance divided by the square of the mean, which was identical to the coefficient of variation (CV^2)[57]. Therefore, the protein noise in the Noise_chr dataset represents isogenic cell-to-cell heterogeneity in protein expression levels. It is well known that protein noise is mainly composed of two sources: intrinsic noise dominating lowly expressed proteins and extrinsic noises dominating highly expressed proteins. Owing to the Poisson-like statistical nature of intrinsic noise, the protein noise depends on the mean expression levels of the lowly expressed genes (Supplementary Fig. 20c). To compensate for this dependency, we calculated the DM values for each gene as the vertical distance of each protein noise ($\log_{10}$-transformed CV^2) from a smoothed running median of the protein noise; this conventional compensation is widely used for protein noise[44,58]. Consequently, we obtained DM_{noise_ch}, as a measure of variability against stochastic perturbations (Supplementary Fig. 20d). Using the noise limits defined by Taniguchi *et al*., we determined the expression level at which the intrinsic noise limit was equal to the extrinsic noise limit. Genes with expression levels below this threshold were defined as lowly expressed, whereas those with expression levels beyond this threshold were defined as highly expressed. We compared DM_{noise_ch} with the DM values of the other datasets for lowly and highly expressed genes separately.

Data visualization

All figures were generated in R[59] and the following packages were used. The phylogenetic tree of the 16 natural isolates of *E. coli* was visualized using the ape[60] and ggtree[61] packages. Pairwise phylogenetic distances were based on core genome alignments and were obtained from a previous study[22]. Network visualization was performed using the ggraph[62], igraph[63] and tidygraph[64] packages. The Venn diagram was visualized using the ggVenndiagram[65], gplots[66] and venneuler[67] packages. Other illustrative plots were generated using the ggplot2[68] and ggpubr[69] packages.

Data availability

All microarray data are available in the National Center for Biotechnology Information's Gene Expression Omnibus functional genomics data repository under the accession number GSE260863.

Code availability

Custom scripts used in the manuscript are available at https://github.com/tsuruubi/Variability_E_coli.

Acknowledgements

The authors thank H. Koike for performing MA experiments. The authors acknowledge H. Kotani for providing technical assistance with the DNA microarray experiments. This work was supported by, the Japan Society for the Promotion of Science (JSPS) KAKENHI (18H02427, 22H05403 to S.T.;17H06389 to C.F. and S.T.; 22K21344 to C.F.) and the Japan Science and Technology Agency (JST) ERATO (JPMJER1902 to S.T. and C.F.).

References

1. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits* (Sinauer Sunderland, MA, 1998).
2. Levis, N. A. & Pfennig, D. W. Evaluating 'Plasticity-First' Evolution in Nature: Key Criteria and Empirical Approaches. *Trends Ecol Evol*. **31**, 563-574 (2016).
3. Darwin, C. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life* (London : John Murray, 1859).
4. Waddington, C. H. *The strategy of the genes* (Routledge, 1957).
5. Smith, J. M. et al. Developmental constraints and evolution: a perspective from the Mountain Lake conference on development and evolution. *The Quarterly Review of Biology*. **60**, 265-287 (1985).
6. Noble, D. W. A., Radersma, R. & Uller, T. Plastic responses to novel environments are biased towards phenotype dimensions with high additive genetic variation. *Proc Natl Acad Sci U S A*. **116**, 13452-13461 (2019).
7. Arnold, S. J., Pfrender, M. E. & Jones, A. G. The adaptive landscape as a conceptual bridge between micro- and macroevolution. *Genetica*. **112-113**, 9-32 (2001).
8. Uller, T., Moczek, A. P., Watson, R. A., Brakefield, P. M. & Laland, K. N. Developmental Bias and Evolution: A Regulatory Network Perspective. *Genetics*. **209**, 949-966 (2018).
9. Parter, M., Kashtan, N. & Alon, U. Facilitated variation: how evolution learns from past environments to generalize to new environments. *PLoS Comput Biol*. **4**, e1000206 (2008).
10. Draghi, J. A. & Whitlock, M. C. Phenotypic plasticity facilitates mutational variance, genetic variance, and evolvability along the major axis of environmental variation. *Evolution*. **66**, 2891-2902 (2012).
11. Kouvaris, K., Clune, J., Kounios, L., Brede, M. & Watson, R. A. How evolution learns to generalise: Using the principles of learning theory to understand the evolution of developmental organisation. *PLoS Comput Biol*. **13**, e1005358 (2017).
12. Waddington, C. H. Genetic assimilation of an acquired character. *Evolution*, 118-126 (1953).
13. Wagner, G. P., Booth, G. & Bagheri-Chaichian, H. A Population Genetic Theory of Canalization. *Evolution*. **51**, 329-347 (1997).
14. Rifkin, S. A., Houle, D., Kim, J. & White, K. P. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature*. **438**, 220-223 (2005).
15. Denver, D. R. et al. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat Genet*. **37**, 544-548 (2005).
16. Cheverud, J. M. A Comparison of Genetic and Phenotypic Correlations. *Evolution*. **42**, 958-968 (1988).
17. Meiklejohn, C. D. & Hartl, D. L. A single mode of canalization. *Trends in Ecology & Evolution*. **17**, 468-473 (2002).
18. Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional

organization. *Nat Rev Genet*. **5**, 101-113 (2004).

19. Voordeckers, K. et al. Identification of a complex genetic network underlying *Saccharomyces cerevisiae* colony morphology. *Mol Microbiol*. **86**, 225-239 (2012).
20. Sastry, A. V. et al. The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat Commun*. **10**, 5536 (2019).
21. Lamoureux, C. R. et al. A multi-scale transcriptional regulatory network knowledge base for *Escherichia coli*. *bioRxiv*. Preprint at https://www.biorxiv.org/content/biorxiv/early/2022/11/18/2021.04.08.439047.full.pdf, 2021.2004.2008.439047 (2022).
22. Rousset, F. et al. The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. *Nat Microbiol*. **6**, 301-312 (2021).
23. Halligan, D. L. & Keightley, P. D. Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*. **40**, 151-172 (2009).
24. Fijalkowska, I. J. & Schaaper, R. M. Mutants in the Exo I motif of Escherichia coli dnaQ: defective proofreading and inviability due to error catastrophe. *Proc Natl Acad Sci U S A*. **93**, 2856-2861 (1996).
25. Tsuru, S. et al. Genomic confirmation of nutrient-dependent mutability of mutators in *Escherichia coli*. *Genes Cells*. **20**, 972-981 (2015).
26. Maeda, T. et al. High-throughput laboratory evolution reveals evolutionary constraints in *Escherichia coli*. *Nat Commun*. **11**, 5970 (2020).
27. Singh, G. P. Coupling between noise and plasticity in *E. coli*. *G3 (Bethesda)*. **3**, 2115-2120 (2013).
28. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nat Genet*. **38**, 830-834 (2006).
29. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. **102**, 15545-15550 (2005).
30. Tierrafria, V. H. et al. RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12. *Microb Genom*. **8** (2022).
31. Keseler, I. M. et al. The EcoCyc Database in 2021. *Front Microbiol*. **12**, 711077 (2021).
32. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature*. **405**, 590-593 (2000).
33. Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J. & Hartl, D. L. Genetic properties influencing the evolvability of gene expression. *Science*. **317**, 118-121 (2007).
34. Yang, R., Zhuhadar, L. & Nasraoui, O. Bow-tie decomposition in directed graphs. in *14th International Conference on Information Fusion* 1-5 (IEEE, 2011).
35. Ghosh Roy, G., He, S., Geard, N. & Verspoor, K. Bow-tie architecture of gene regulatory networks in species of varying complexity. *J R Soc Interface*. **18**, 20210069 (2021).

36. Martinez-Antonio, A., Janga, S. C., Salgado, H. & Collado-Vides, J. Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends Microbiol*. **14**, 22-27 (2006).
37. Kaneko, K. Relationship among phenotypic plasticity, phenotypic fluctuations, robustness, and evolvability; Waddington's legacy revisited under the spirit of Einstein. *J Biosci*. **34**, 529-542 (2009).
38. Furusawa, C. & Kaneko, K. Formation of dominant mode by evolution in biological systems. *Phys Rev E*. **97**, 042410 (2018).
39. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A*. **97**, 6640-6645 (2000).
40. Baba, T. et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*. **2**, 2006 0008 (2006).
41. Sharma, R. C. & Schimke, R. T. Preparation of electrocompetent *E. coli* using salt-free growth medium. *Biotechniques*. **20**, 42-44 (1996).
42. Niccum, B. A., Lee, H., MohammedIsmail, W., Tang, H. & Foster, P. L. The Spectrum of Replication Errors in the Absence of Error Correction Assayed Across the Whole Genome of *Escherichia coli*. *Genetics*. **209**, 1043-1054 (2018).
43. Rychel, K. et al. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res*. **49**, D112-D120 (2021).
44. Newman, J. R. et al. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*. **441**, 840-846 (2006).
45. Uchida, Y., Shigenobu, S., Takeda, H., Furusawa, C. & Irie, N. Potential contribution of intrinsic developmental stability toward body plan conservation. *BMC Biol*. **20**, 82 (2022).
46. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. **25**, 25-29 (2000).
47. Gene Ontology, C. et al. The Gene Ontology knowledgebase in 2023. *Genetics*. **224** (2023).
48. Yu, G. et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*. **26**, 976-978 (2010).
49. Pagès, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. *R package version 1.64.1*, https://bioconductor.org/packages/AnnotationDbi (2023).
50. Sayols, S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. *microPublication Biology* (2023).
51. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28**, 27-30 (2000).
52. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*. **2**, 100141 (2021).
53. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful

Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. **57**, 289-300 (1995).

54. Horinouchi, T. et al. Prediction of Cross-resistance and Collateral Sensitivity by Gene Expression profiles and Genomic Mutations. *Sci Rep*. **7**, 14009 (2017).
55. Posfai, G. et al. Emergent properties of reduced-genome *Escherichia coli*. *Science*. **312**, 1044-1046 (2006).
56. Schmidt, A. et al. The quantitative and condition-dependent *Escherichia coli* proteome. *Nat Biotechnol*. **34**, 104-110 (2016).
57. Taniguchi, Y. et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*. **329**, 533-538 (2010).
58. Silander, O. K. et al. A genome-wide analysis of promoter-mediated phenotypic noise in *Escherichia coli*. *PLoS Genet*. **8**, e1002443 (2012).
59. R Core Team. R: A Language and Environment for Statistical Computing, 2023).
60. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. **35**, 526-528 (2019).
61. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. **8**, 28-36 (2017).
62. Pedersen, T. L. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. *R package version 2.2.1*, https://CRAN.R-project.org/package=ggraph (2024).
63. Csardi, G. & Nepusz, T. The igraph software. *Complex syst*. **1695**, 1-9 (2006).
64. Pedersen, T. L. tidygraph: A Tidy API for Graph Manipulation. *R package version 1.3.1*, https://tidygraph.data-imaginist.com (2024).
65. Gao, C. H., Yu, G. & Cai, P. ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram. *Front Genet*. **12**, 706907 (2021).
66. Warnes, G. R. et al. gplots: Various R Programming Tools for Plotting Data. *R package version 3.1.3.1*, https://CRAN.R-project.org/package=gplot (2024).
67. Wilkinson, L. Venn and Euler Diagrams, https://www.rforge.net/venneuler/ (2024).
68. Wickham, H. ggplot2: Elegant Graphics for Data Analysis, https://ggplot2.tidyverse.org (2016).
69. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. *R package version 0.6.0*, https://rpkgs.datanovia.com/ggpubr/ (2023).

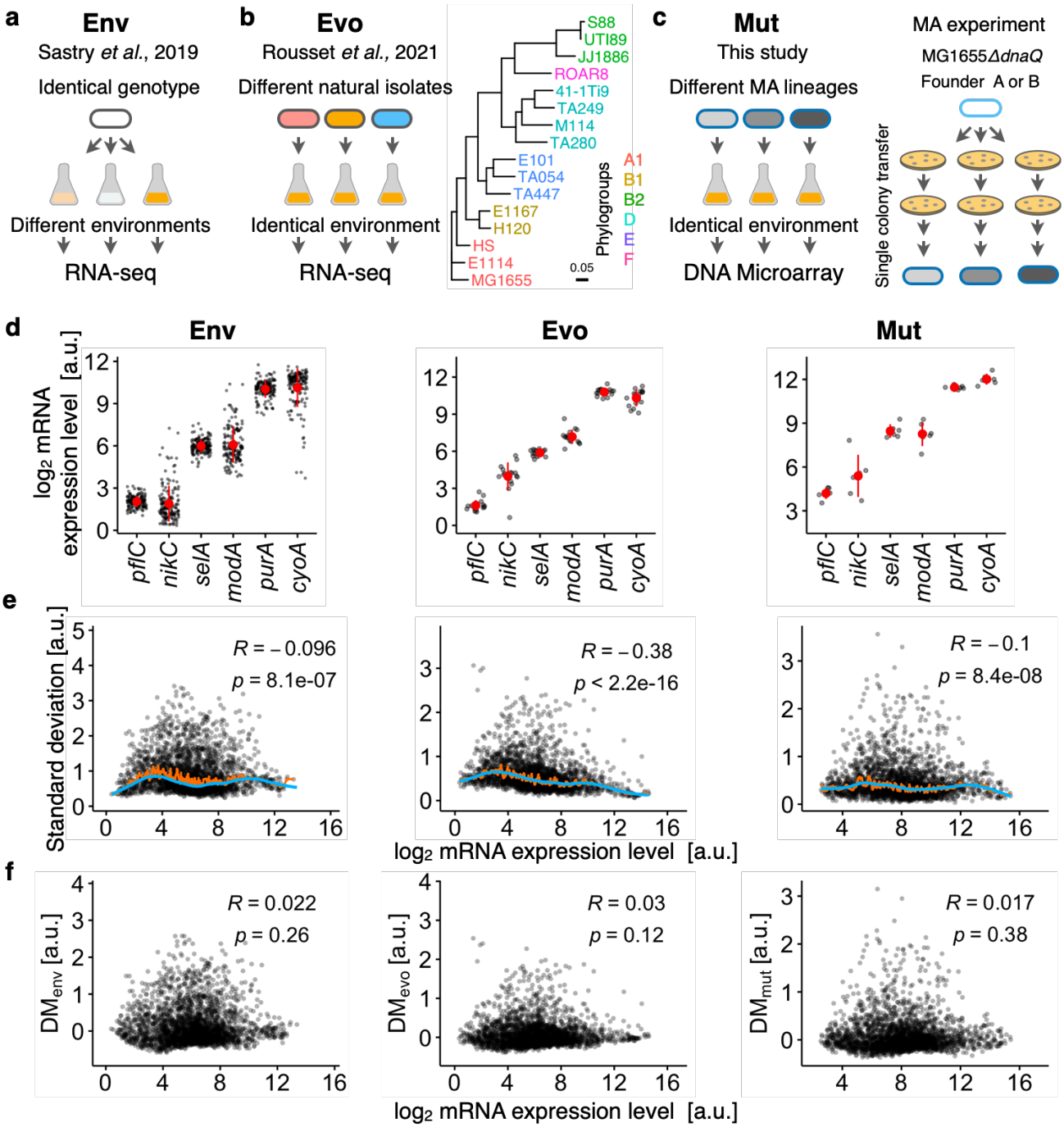
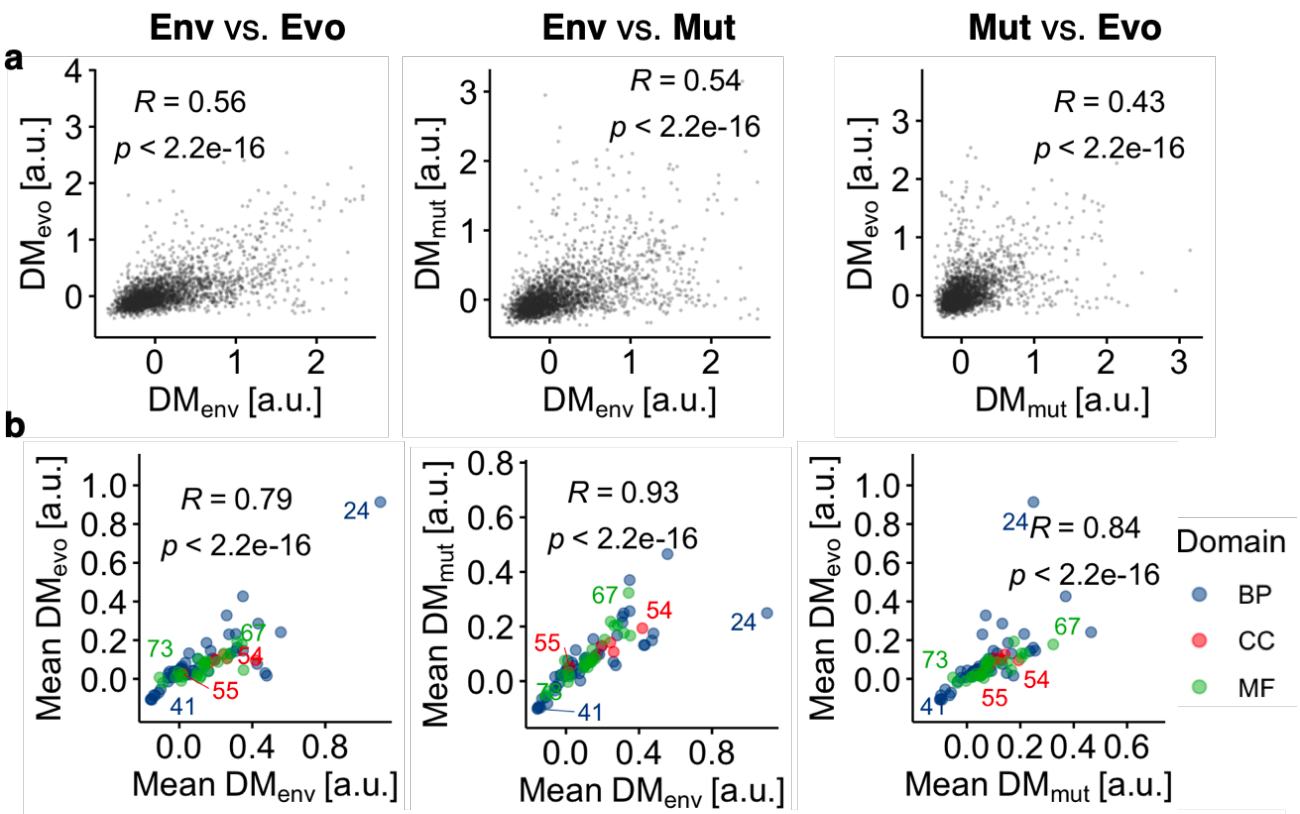


Fig. 1: Transcriptional variability in *E. coli*.

a, A schematic diagram of the Env dataset. The transcriptome profiles of *E. coli* MG1655 were obtained under various growth conditions[20,21]. **b**, Schematic of the Evo dataset. The transcriptome profiles of 16 natural isolates of *E. coli* were obtained under a common growth condition[22]. The phylogenetic tree was based on the core genome alignment. **c**, Schematics of the MA and Mut dataset. The MA experiment was initiated by two founders of the mutator strain, *E. coli* MG1655*ΔdnaQ*. Single colonies were randomly picked and transferred daily to fresh LB agar plates for 30 rounds. Three MA lineages from rounds 10 and 30 for each founder were subjected to DNA microarray analysis. Transcriptome profiles were obtained during the exponential growth phase in LB. **d–f**, Transcriptional

variability in Env, Evo, and Mut datasets (from left to right). The transcriptome profiles of the MA lineages in the final (30th) round were used for the Mut dataset. **d**, Expression levels of six representative genes. The expression levels were $\log_2$-transformed and quantile-normalized for each dataset. Each grey circle corresponds to a different bacterial culture. Red closed circles and error bars represent means and standard deviations, respectively. **e**, Relationship between means and standard deviations in the expression levels of 2622 genes present in all three datasets. Each grey circle corresponds to a different gene. The variability metric for each gene was defined as the vertical deviation from a smoothed spline (sky blue) calculated from the running median (dark orange) of the standard deviations. The calculated distance to the median is called DM, with subscripts of the corresponding datasets. For the Mut dataset, the standard deviations were obtained by calculating the variances in expression levels separately in each of the two groups distinguished by the founders (founders A and B) and subsequently averaging them. **f**, Relationship between DMs and mean expression levels. The insets represent Spearman's rank correlation coefficients (R) and p-values.



Label	Domain	ID	Term Name
24	BP	GO:0071973	bacterial-type flagellum-dependent cell motility
41	BP	GO:0009273	peptidoglycan-based cell wall biogenesis
54	CC	GO:0055052	ABC transporter complex, substrate-binding subunit-containing
55	CC	GO:0022625	cytosolic large ribosomal subunit
67	MF	GO:0051539	4 iron, 4 sulfur cluster binding
73	MF	GO:0004518	nuclease activity

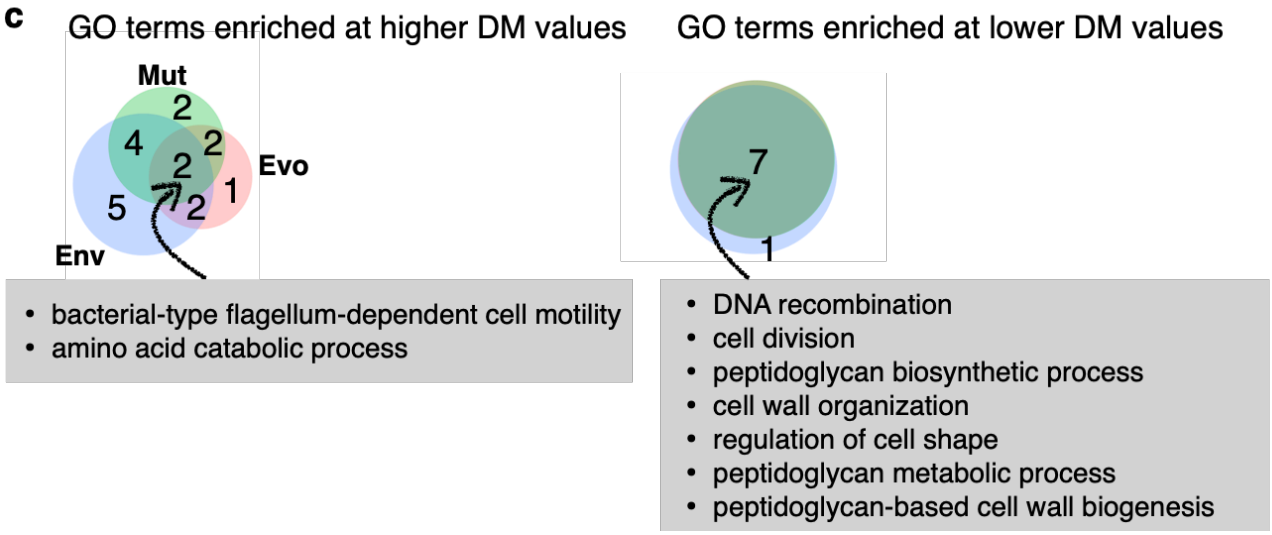


Fig. 2: Positive correlations of transcriptional variability between datasets.

a, Pairwise correlations are shown for DMs at the gene levels between the Env, Evo, and Mut datasets.

b, Correlations of DMs at the functional category level between pairs of datasets. The genes were grouped into functional categories based on their GO terms. To reduce the redundancy of GO terms, a custom subset (82) of GO terms was used as detailed in the **Methods** section. Each dot corresponds to a different GO term. The mean DM was calculated by averaging the DM of the genes belonging to each GO term. The three domains of GO–biological processes (BP), cellular components (CC), and molecular functions (MF)–are labeled in different colors. Spearman's R and p-values are shown in the panels. Six representative GO terms are labeled with different numbers, as detailed at the bottom. **c**, Venn diagram of GO terms enriched in higher (left) and lower (right) DM values in each dataset. The numbers indicated in the diagram show the number of GO terms enriched in the corresponding set.

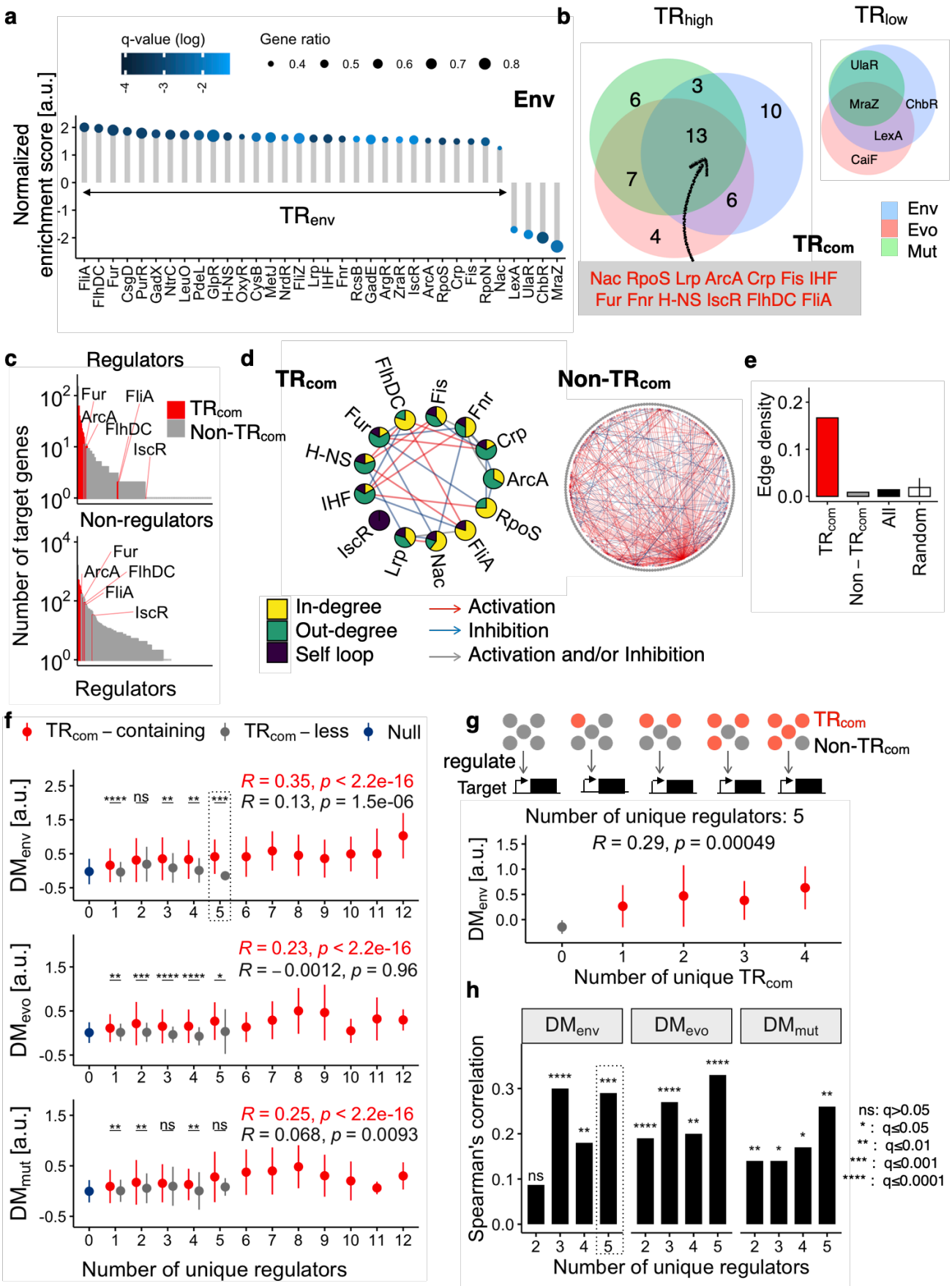


Fig. 3: Transcriptional regulators and transcriptional variability.

a, Transcriptional regulators (TR) enriched in genes with higher and lower DM_{env}. GSEA identified statistically significant TRs ($q<0.05$). Positive and negative enrichment scores indicate enrichment at higher and lower DM_{env}, respectively. The gene ratio was calculated as the number of core target genes contributing to enrichment scores divided by the number of target genes. TRs with positive enrichment scores are labeled as TR_{env}. The q-values (adjusted p-values) were computed using the BH procedure. **b**, Venn diagram of the enriched TRs in each dataset. TR_{evo} and TR_{mut} represent the TRs with positive enrichment scores in the Evo and Mut datasets, respectively. The intersection of the three sets represents TR_{com}. **c**, Rank plot of the number of target genes in the TRs. The TRs were sorted in decreasing order. Target genes were divided into regulators (transcription factors and sigma factors, upper panel) and non-regulators (lower panel). The red and grey bars represent TR_{com}s and the other TRs (Non-TR_{com}s), respectively. The labels are subsets of TR_{com}. **d**, Circular representation of *E. coli* transcriptional regulatory network (TRN) within TR_{com}s (left) and Non-TR_{com}s (right). The red, blue, and grey lines represent activation, inhibition, and activation and/or inhibition, respectively. Pie charts indicate the percentage of each degree within the TR_{com} nodes. The edges corresponding to the self-loops were omitted from the visualization. **e**, Edge densities within regulatory groups. Random groups were generated by drawing 13 TRs at random 10,000 times from a pool of TRs, where TRs that regulate fewer than five target genes were omitted. Error bars represent standard deviation. Edge density was calculated as the total number of edges within the regulators divided by the total number of possible edges within the regulators, excluding the self-loops. **f**, Relationship between DM values and the number of unique regulators. The top, middle, and bottom panels show the DM_{env}, DM_{evo} and DM_{mut}, respectively. Red, grey, and blue symbols represent target genes receiving TR_{com}-containing regulation, TR_{com}-less regulation, and genes without known regulators, respectively. Means and standard deviations are shown as the points and the error bars, respectively. Data points for TR_{com}-containing and TR_{com}-less regulations were restricted to those comprising more than four target genes. Spearman's R was calculated for data points 0–5 on the x-axis for both TR_{com}-containing and TR_{com}-less regulations. Asterisks represent the statistical significance levels (q-values) of the Wilcoxon test between TR_{com}-containing and TR_{com}-less genes in each class. The target genes within the dotted square are used in panel **g**. **g**,**h**, The correlation between DM values and the number of TR_{com} regulations. **g**, Relationship between DM_{env} and the number of TR_{com}s for target genes regulated by any of the five regulators. Spearman's R and p-values are shown. **H**, Spearman's R for DM_{env} (left), DM_{evo} (center), and DM_{mut} (right). The enclosed square corresponds to panel **g**.

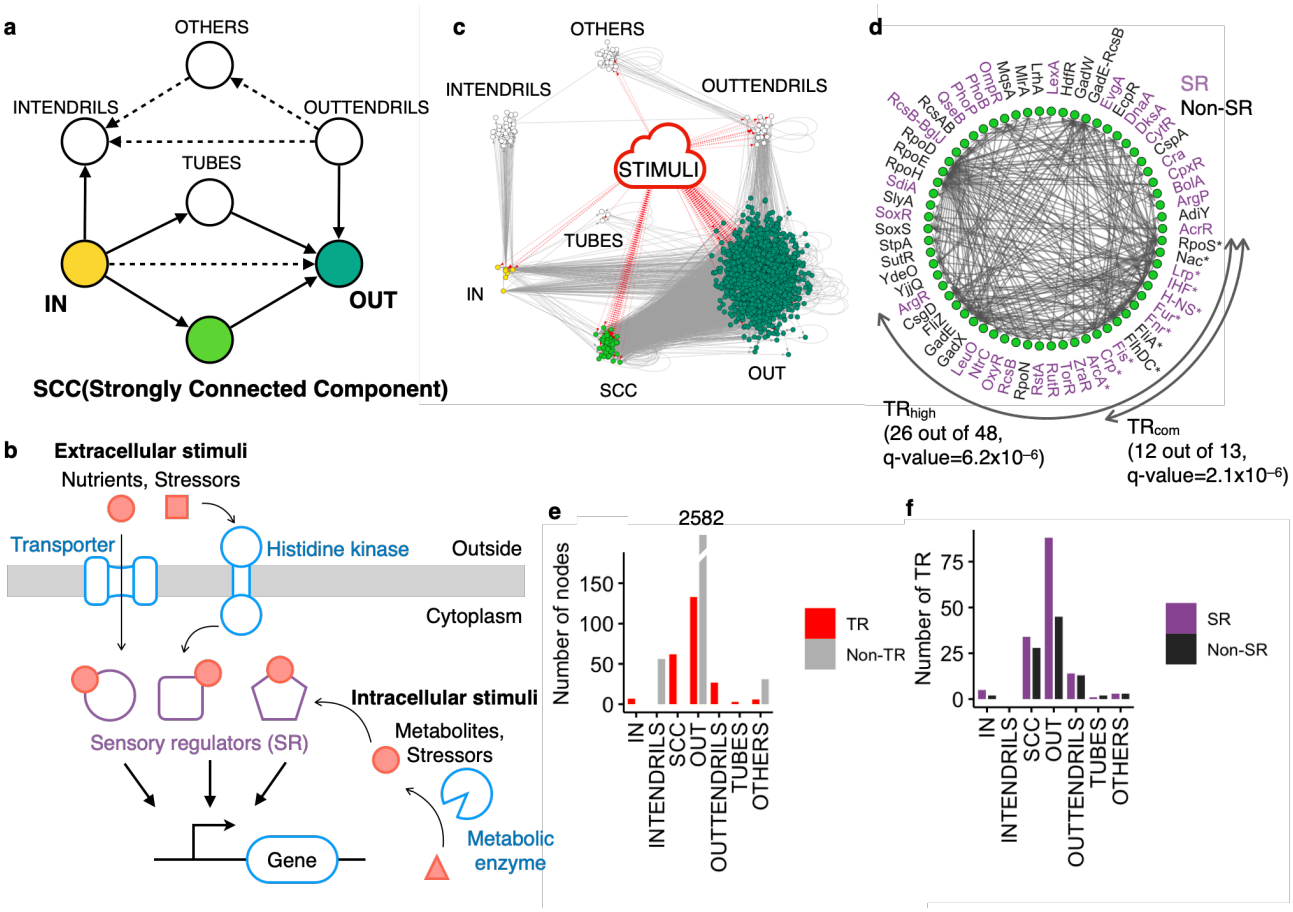
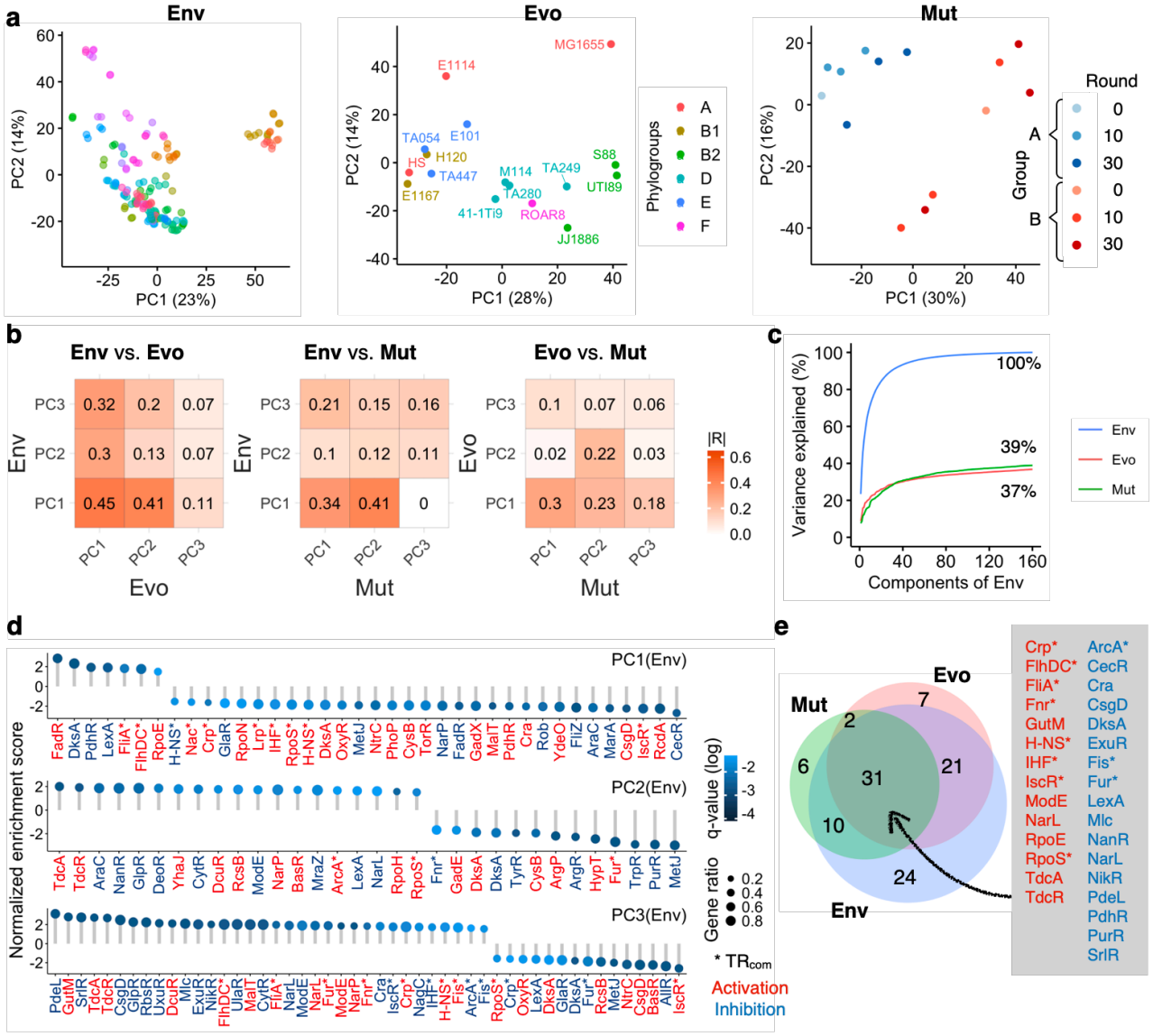


Fig. 4: Bow-tie decomposition of the gene regulatory network in *E. coli*.
a, Yang's bow-tie structure[34] is delineated into seven sectors, with the main sectors highlighted: IN (yellow), OUT (light green), and the strongly connected component (SCC, blue-green). Arrows indicate the direction of regulation, with solid arrows defining each sector and dashed arrows representing possible regulations without violating the sector definitions. **b**, Schematic illustrating how sensory regulators (SR, open purple symbols) influence the expression levels of target genes by interacting with intracellular and extracellular effectors (red closed symbols). **c**, Bow-tie representation of the gene regulatory network in *E. coli* based on panel **a**. Each node represents a TR or target gene. Grey arrows denote regulation (activation and/or inhibition). The STIMULI block represents an aggregated group of known effectors interacting with SRs, with red dotted arrows indicating the interactions between the effectors in STIMULI and SRs. **d**, Regulatory network within the SCC. SR and non-SR groups are colored purple and black, respectively, with asterisks representing TR_{com}. **e**, Distribution of nodes in each sector. TR and non-TR groups are colored red and grey, respectively. IN, SCC, TUBES, and OUTTENDRILS consist solely of TR, whereas INTENDRILS contains only non-TRs by definition. **f**, Number of TR in each sector. SR and non-SR groups are colored purple and black, respectively.



f Representative GO terms enriched in the **positive** direction of PC1 in Env (growth family)

Domain	ID	Term Name
MF	GO:0140640	catalytic activity, acting on a nucleic acid
MF	GO:0003735	structural constituent of ribosome
BP	GO:0006412	translation
BP	GO:0140694	non-membrane-bounded organelle assembly
BP	GO:0001510	RNA methylation

Representative GO terms enriched in the **negative** direction of PC1 in Env (starvation family)

Domain	ID	Term Name
CC	GO:0055052	ABC transporter complex, substrate-binding subunit-containing
BP	GO:0008652	amino acid biosynthetic process
BP	GO:0009401	phosphoenolpyruvate-dependent sugar phosphotransferase system
BP	GO:0006006	glucose metabolic process
BP	GO:0042594	response to starvation

Fig. 5: Alignment of genetic and environmental correlations in transcriptional variability.

a, First and second principal components (PC) of Env, Evo and Mut arranged from left to right. **b**, Correlation analysis for loadings of the first three PCs among the three datasets. The color scale and numbers in the tiles represent the absolute values of Spearman's R. **c**, Cumulative variance explained by the PCs defined by the Env dataset. Overall, the Env dataset explained 37% and 39% of the variance in the Evo and Mut datasets, respectively. **d**, TRs were significantly enriched ($q<0.05$) in greater loadings, eithter positive or negative, of the first three PCs in the Env dataset. GSEA was used to identify regulations, distinguishing between activatory (red) and inhibitory (blue) regulation. Asterisks denote regulations by TR_{com}s. **e**, Venn diagram illustrating the TRs enriched in the first three PCs of each dataset. GSEA was performed for each dataset, as shown in panel **d**. Numbers indicate the number of regulations in the corresponding set. The detailed regulations at the intersection of the three sets are shown on the right. **f**, GO terms enriched with greater loadings in the positive (top) and negative (bottom) directions in PC1 of the Env dataset (five representatives). The former and latter GO terms were categorized into the growth and starvation families. **g**, Number (left) and fraction (right) of GO terms enriched in each sign of each PC in each dataset. GSEA was conducted for each PC in each dataset. The fraction was calculated by dividing the number of GO terms in each family by the total number of GO terms in each PC sign. A full list of the enriched GO terms is presented in Supplementary Fig. 10–12.

a

Dataset	Perturbation	mRNA/protein	Variability	Reference
Strs_hp	Environmental	mRNA	DM_{strs_hp}	Horinouchi et al (2017)
Strs_he	Genetic	mRNA	DM_{strs_he}	Horinouchi et al (2017)
Strs	Genetic	mRNA	DM_{strs}	Maeda & Iwasawa et al (2020)
Prtn	Environmental	protein	DM_{prtn}	Schmidt et al (2016)
Noise_ch	Molecular noise	protein	DM_{noise_ch}	Taniguchi et al (2010)

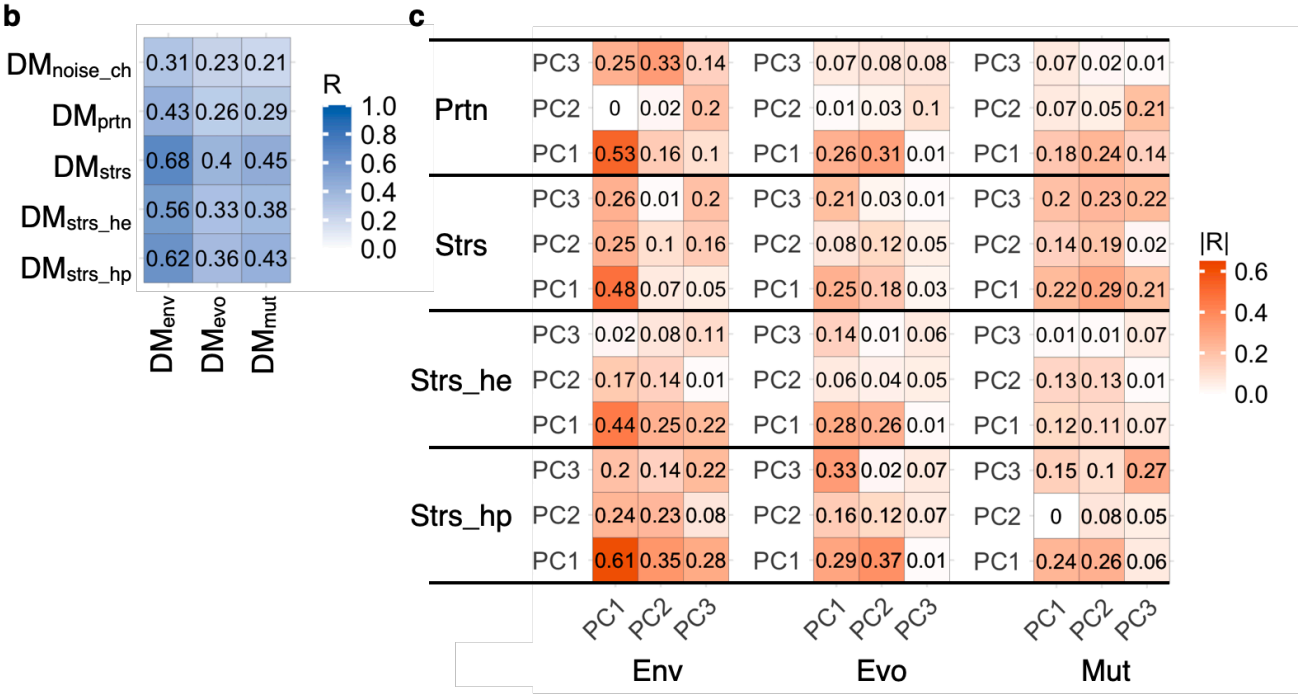


Fig. 6: Similarity of phenotypic variability across different datasets

a, Datasets compared to three main datasets. **b**, Correlation analysis of the variability across different datasets. The color bar and numbers in the tiles represent Spearman's R. Correlation analysis for DM_{noise_ch}, was conducted only for highly expressed proteins with extrinsic noise dominance. The results of the lowly expressed proteins are detailed in Supplementary information Appendices II. **c**, Correlation analysis of loadings of the first three PCs between the datasets. The color bar and numbers in the tiles represent absolute values of Spearman's R.