Evolution of gene knockout strains of E-coli reveal regulatory architectures governed by metabolism

McCloskey, Douglas; Xu, Sibe; Sandberg, Troy E.; Brunk, Elizabeth; Hefner, Ying; Szubin, Richard; Feist, Adam M.; Palsson, Bernhard O.

Published in:
Nature Communications

Link to article, DOI:
10.1038/s41467-018-06219-9

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Evolution of gene knockout strains of *E. coli* reveal regulatory architectures governed by metabolism

Douglas McCloskey¹,², Sibei Xu¹, Troy E. Sandberg, Elizabeth Brunk, Ying Hefner, Richard Szubin, Adam M. Feist,² & Bernhard O. Palsson

Biological regulatory network architectures are multi-scale in their function and can adaptively acquire new functions. Gene knockout (KO) experiments provide an established experimental approach not just for studying gene function, but also for unraveling regulatory networks in which a gene and its gene product are involved. Here we study the regulatory architecture of *Escherichia coli* K-12 MG1655 by applying adaptive laboratory evolution (ALE) to metabolic gene KO strains. Multi-omic analysis reveal a common overall schema describing the process of adaptation whereby perturbations in metabolite concentrations lead regulatory networks to produce suboptimal states, whose function is subsequently altered and re-optimized through acquisition of mutations during ALE. These results indicate that metabolite levels, through metabolite-transcription factor interactions, have a dominant role in determining the function of a multi-scale regulatory architecture that has been molded by evolution.
B

iological response to gene loss can be evaluated on multiple time-scales. The immediate response to genetic perturbation is studied by measuring an organism’s phenotypic response to a gene knockout (KO)\(^1\)-\(^4\). For example, entire KO strain collections have been generated and used to define essential genes\(^5\)-\(^8\). Besides assessing gene function, gene knockouts can be studied at the systems level through the integration of multi-omics data sets (i.e., metabolomics, fluxomics or network reaction rates, proteomics, and transcriptomics) to better understand the regulatory architecture that relies on the gene product. For example, it has been found that perturbations to the metabolic network are rapidly compensated for by flux re-routing caused by adjustments made at the regulatory level that re-tune enzyme level\(^1\)-\(^9\). Specifically, these studies found that regulatory changes (and in particular, changes in metabolite levels) occurred in proximity to the network lesion that a gene KO created. However, the extent to which distant regulatory changes relative to the location of the network lesion occurred was not discussed\(^1\)-\(^9\). In addition, the adaptive consequences of gene loss were not investigated.

The adaptive response to genetic perturbation is studied by measuring changes in physiological function after perturbation and during adaptation\(^10\)-\(^12\), and then characterizing the mutations that are required for the organism to regain the ability to grow optimally under the given conditions\(^13\)-\(^26\). For example, it has been shown in bacteria and yeast that the likelihood of accumulating compensatory mutations is a function of the fitness cost of the KO\(^22\)-\(^24\). Importantly, compensatory mutations often require the rewiring of existing regulatory networks to regain fitness, thus revealing the role of the lost gene in the regulatory architecture of the biological system\(^26\). Despite the potential to reveal novel insights into the regulatory architecture, to the best of our knowledge, a comprehensive systematic study looking at the rewiring of the regulatory network in response to gene loss has not been performed.

Previous work implemented a novel experimental design that involved gene knockouts (KOs) and adaptive laboratory evolution (ALE) in a pre-evolved *Escherichia coli* K-12 MG1655 strain (Fig. 1) to reveal detailed and mechanistic KO-specific adaptive responses to the loss of a gene\(^27\)-\(^30\). Here, bioinformaticians were implemented to reveal commonalities of how biological systems and specifically regulatory networks respond and adapt to gene KO at a systems level. First, the experimental design was confirmed through control evolutions of the pre-evolved strain. Second, multivariate statistical data decomposition methods found that the dominant modes of the data involved the drive towards regaining optimal fitness, while independent replicate evolutions revealed diversity in the adaptive paths selected in pursuit of optimal fitness. In this context, “optimal” indicates the biochemical state that allows for the maximal growth rate that the organism can achieve given the current environmental and genetic conditions. Third, biochemical pathway integration with multi-omic data sets revealed a common model of adaptive evolution. In this schema, network perturbation from gene KO altered metabolic flux, leading to perturbations in metabolite concentrations, which in turn triggered regulatory network responses altering gene expression. Gene expression responses were subsequently modified through mutations selected for during adaptation that improved fitness via ameliorated metabolic flux.

Results

Evolution experiment implementation. A wild-type *E. coli* K-12 MG1655 strain previously evolved under glucose minimal media at 37 °C\(^31\) (denoted as “Ref”) was used as the starting strain in order to isolate biological changes caused by adaption to the loss of a gene product from those caused by adaption to the growth conditions of the experiment (Fig. 1e). Ref was a non-mutator strain and had the fewest number of mutations among the replicate adaptive laboratory evolution (ALE) endpoints generated.

Perturbations consisting of five separate metabolic gene KOs that were predicted to result in large metabolic rearrangements based on computational metabolic network analysis (see Methods, Supplementary Data 1) were implemented in Ref. Genes (see Methods) encoding enzymes for the reactions of GND (\(gnd\), 6-phosphogluconate dehydrogenase), GLCptsp (genes \(ptsH\), \(ptsl\), and \(crr\) corresponding to enzymes HPr, EI, and EIA, respectively), SUCDi (genes \(sucA\), \(sucB\), \(sucC\), and \(sucD\) corresponding to the enzyme succinate dehydrogenase), TPI (\(tpiA\), triosephosphate isomerase), and PGI (\(pgi\), phosphoglucone isomerase) were removed to generate strains uGnd, uPtsHcrr, uSdhCB, uTpiA, and uPgi, respectively (denoted “un-evolved knockout strains” or “uKO”). GND generates D-ribulose-5-phosphate (rup-D), which is used in nucleotide biosynthesis, and re-charges NADPH, which is used for biosynthesis, in the final step of the oxidative Pentose Phosphate Pathway (oxPPP). \(ptsH\), \(ptsl\), and \(crr\) are primary components of the phosphotransferase system (PTS), which is the primary route for carbon import in *E. coli*, and aids in conserving energy by utilizing phosphoenolpyruvate (pep) to phosphorylate glucose instead of ATP. SUCDi couples the TCA cycle to respiration by charging and donating quinones to the electron transport chain (ETC) via Complex II. TPI avoids bifurcation of lower glycolysis by isomerizing dihydroxyacetone phosphate (dhap) to glyceraldehyde-3-phosphate (g3p) for subsequent enzymatic convert to pyruvate (pyr) via upper glycolysis. PGI converts glucose 6-phosphate (g6p) to fructose 6-phosphate (f6p) in the first committed step through upper glycolysis, thus controlling the flux split between the oxPPP and upper glycolysis.

Replicates of the five knockout strains, as well as Ref, were simultaneously evolved on glucose minimal media at 37 °C in an automated ALE platform\(^31\), \(^32\) denoted “evolved knockout strains” or “eKO” where \(i\) denotes the replicate number. The number of replicate endpoints were the following: 2 for “evolved reference strain” (denoted eRef), 3 for eGnd, 4 for ePtsHcrr, 3 for eSdhCB, 4 for eTpiA, and 8 for ePgi. Intracellular metabolite levels, gene expression levels, flux levels, and mutations (i.e., system components) were measured for the ref, uKO, and eKO strains during exponential growth. Intracellular metabolite levels consisted of close to 100 absolute and relative quantitative amounts of metabolites from glycolysis, the pentose phosphate pathway, the TCA cycle, energy and redox metabolism, cofactors, nucleotide metabolism, and amino acid metabolism\(^33\), \(^34\). Gene expression levels consisted of relative fold changes from global RNA sequencing. Flux levels consisted of absolute intracellular flux values computed by metabolic flux analysis (MFA) using a genome-scale model from 13C isotope-labeling experiments\(^35\), \(^36\). Mutations consisted of DNA sequencing mapped onto the reference *E. coli* K-12 MG1655 genome.

Reference strain evolution confirmed the experimental design. An insignificant flux change and the fewest number of network changes were found in eRef strains compared to all eKO strains (Fig. 1e). The average numbers of significant component changes per eRef replicate at the metabolite, transcript, and flux levels were 2.0 ± 0.0, 35.0 ± 5.7, and 0.0 ± 0.0 (ave ± stdev, \(n = 2\)), respectively. These changes in systems components were far fewer than in any of the other eKO strains, where the minimum
number of corresponding changes were 19, 341, and 158 (the average number of corresponding changes were 27.7 ± 7.7, 1051.6 ± 513.7, and 307.9 ± 123.2 (ave ± stdev, \(n = 24\))). The average number of mutations per eRef replicate was also the lowest of all lineages, and were primarily found in cell wall biosynthesis genes. The average number of mutations per eRef was 6.5 ± 0.7, while the average number of mutations per all other eKO strains was 12.8 ± 4.5 (ave ± stdev). Overall, these findings demonstrated that the use of a pre-evolved strain minimized the number of confounding component changes.

Evolution to optimal fitness was captured by the data. Multivariate statistical analysis was performed on the data sets generated. Partial least squares discriminatory analysis (PLS-DA) revealed that the primary adaptive response to the gene KO involved a drive towards recovery of the optimal state (i.e., system re-optimization), followed by a secondary adaptive response that described unique alternate states that could be found at the newly evolved state. For almost all cases analyzed, the first most explanatory mode of PLS-DA (Fig. 2) separated Ref and eKO strains from the uKO strain (74% of eKOs from all data types and lineages, see Methods). This result indicated that the primary
mode of the data accounted for a dominant transition between the Reference state, perturbed state, and evolved fitness states (i.e., captured systems fitness properties). This result was also reflected in the system component profiles themselves where the majority of component levels were restored or partially restored to reference levels (Fig. 3). For almost all cases analyzed, the second most explanatory mode of PLS-DA separated the Ref and eKO strains. This result indicated that the secondary mode of the data accounted for alternate evolved states (i.e., capturing systems diversity, or a ‘plateau’ in the evolutionary landscape). These
types. Overall, the metabolite levels were the most distributed. This characteristic was further reflected in the unique distribution of component profiles between each of the eKOs.

Component profiles reveal systematic variations. In order to dissect the drive towards fitness (mode 1) and generation of diversity (mode 2) further, changes in each system component (i.e., metabolite, transcript, and flux level) between Ref, uKO, and eKO strains were grouped into six profiles (Fig. 3a, see Methods): novel, overcompensated, partially restored, reinforced, restored, and unrestored. The distribution between these six profiles for each component type are shown with horizontal bar charts in Fig. 3b–d. Several trends were found based on these six profiles.

First, the occurrence of profiles varied between omics data types. Overall, the metabolite levels were the most distributed between the six profiles (i.e., had the least deviation). The ave ± stdev of the relative standard deviation (RSD) between profiles ($n = 12$, $+$ and $-$ directions for each of the six profiles) and across lineage ($n = 22$) was $39.9 \pm 14.1$, $132.1 \pm 45.9$, and $84.0 \pm 12.7\%$ for metabolites, transcripts, and fluxes, respectively. In contrast, the transcript levels were dominated by the restored profile, and flux levels were dominated by the restored and unrestored profile. For example, the $pgi$ lineages had an ave ± stdev of restored profiles of $59.0 \pm 5.0$, $80.1 \pm 8.3$, and $66.9 \pm 3.1\%$ for metabolites, transcripts, and fluxes, respectively. The more even metabolite distribution compared to the transcript levels or flux levels indicated that the changes in metabolite levels were less constrained than the gene expression and fluxes.

Second, distribution amongst the profiles varied between KOs. The lineages with the greatest initial loss of fitness had a greater percentage of novel, overcompensated, reinforced, and unrestored profiles than the lineages with a smaller initial loss of fitness. This
difference was most evident for the transcript levels (ave ± stdev of 2.7 ± 0.4, 8.2 ± 3.7, 31.3 ± 23.4, 20.3 ± 10.9, and 18.6 ± 1.8%, and fitness change across evolution of 11.9 ± 3.9, 11.1 ± 2.9, 365.2 ± 20.0, 337.8 ± 73.8, and 244.3 ± 7.1% for the gnd, sdhCB, pgi, ptsHicrr, and tipA lineages; Pearson’s R = 0.94, P-value < 0.02, 6. Supplementary Fig. 1). This observation suggests that the larger the loss in fitness, the greater the number of Innovative (as opposed to restorative) network changes required to regain fitness. Future work with larger sample sizes will be needed to confirm this trend.

Third, the distribution amongst profiles also varied between evolved strain lineages. For example, the eight pGpi endpoints had varying levels of fitness (ave ± stdev of 0.68 ± 0.006, 0.61 ± 0.015, 0.65 ± 0.008, 0.72 ± 0.009, 0.64 ± 0.008, 0.69 ± 0.018, 0.67 ± 0.006, 0.69 ± 0.015 h⁻¹), and noticeable differences in the distribution of profiles among endpoints. This highlighted the biochemical differences in evolved network configurations during adaptation to overcome the perturbation.

Finally, a decoupling between degree of fitness change and degree of -omics data change was apparent. The tipA, pgi, and ptsHicrr lineages incurred the largest loss and recovery of fitness while the gnd and sdhCB lineages incurred only minimal changes in fitness. However, major changes in all -omics data measured between Ref and uKO and between uKO and eKO strains were found in all lineages (Figs. 2 and 3). Interestingly, major changes often occurred in common system components. Major changes could be traced to either perturbed metabolites that act as allosteric or transcriptional regulators (which is consistent with previous studies) or mutations that resulted in alterations to gene expression and protein production could occur as a result of perturbations in relatively few key regulators.

The system component profiles were mapped to the biochemical network of E. coli and analyzed to develop a general framework for understanding evolution at the molecular level. It is important to highlight that the component profiles described above were in all of the analyses presented below. The component profiles were assigned based on statistical criteria. They provided a unitless metric to compare and map multiple data types when quantitative relationships between data types have not been fully established. The component profiles also provided robustness by basing the analysis on change in values between states (i.e., ref, uKO, and eKO) instead of the absolute value found in any one state.

**Changed flux distribution was most prevalent during ALE.**

Changes in pathway usage between the Ref, uKO, and eKO strains were calculated, and differences between the flux distribution in the uKO and eKO strains were grouped into changed flux distribution (i.e., the pathway usage was changed) or changed flux capacity (i.e., the same pathway was used but at a higher flux level, see Methods for extended definitions, Fig. 4a–d). Changed flux distribution was found to be more prevalent than changed flux capacity. Changed flux distribution was found to occur 55.6% of the time, while a change in flux capacity was found to occur 22.0% of the time across all perturbations and lineages (Fig. 4e). The remaining 22.4% of cases were unaffected.

For example, flux was initially re-routed through the Entner–Doudoroff (ED) pathway in uGnd (Fig. 4f) in order to generate ribose through the non-oxidative Pentose Phosphate pathway (non-oxPPP). The ED pathway has a net yield of one ATP, NADH, and NADPH per molecule of glucose, whereas glycolysis has a net yield of two ATP and NADH. Instead, the eGnd strains limited the use of the oxidative pentose phosphate pathway (oxPPP) and increased the flux capacity through the higher energy and redox equivalent producing pathway of glycolysis. Further examples are given in Fig. 4. These results
Traced to known TRN responses were linked to corresponding changes in regulatory lism, iron regulation, oxidative stress, DNA repair, and other that were associated with carbon metabolism, nitrogen metabo-

tion pro

Perturbed metabolite levels triggered TRN responses in uKOs. Transcriptional regulatory network (TRN) responses in uKOs that were associated with carbon metabolism, nitrogen metabo-

ism, iron regulation, oxidative stress, DNA repair, and other stress responses that control the majority of known functions in E. coli were linked to corresponding changes in regulatory metabolite levels (see Methods). Perturbed metabolite levels were traced to known TRN responses by mapping measured metabolite profiles to metabolite-activated transcription factors (TFs). The relationship (i.e., positive or negative) between a metabolite profile, a TF that interacts with the metabolite, and the expression profiles of the transcription units (TUs) regulated by the TF (see Methods, Fig. 5) were compared. Strong evidence (i.e., statistically significant gene expression pattern for genes that are regulated by a single TF, see Methods) for changed TF activation profiles (analogous to the system component profiles, Fig. 2) were identified for 75 TFs (Supplementary Data 2, Fig. 5). These included 7 global TFs (i.e., CRP, Fis, IHF, ArcA, Lrp, FNR, and HNS) and 68 pathway-specific TFs (see Methods). The activation profiles of 15 TFs (which included the 7 global TFs and the 8 pathway-specific TFs ArgR, CpxR, Cra, Fur, NsrR, OxyR, PhoB, and TyrR) were changed across all lineages. The remaining 60 TFs appeared to be changed in a perturbation and lineage-specific manner.

Interestingly, TF activation and TF gene expression was not coincidental (ave ± std 5.4 ± 3.8, 4.1 ± 2.6, and 70.5 ± 6.1% agreement, disagreement, no significant change in expression profile per lineage, respectively, Supplementary Data 3). This result indicated that changed TF activation was mostly attributed to changed concentrations in their metabolic activators as opposed to changed TF gene expression levels. Similar observations have been made for sigma factors and the expression levels of sigma factor DNA binding operons in response to a key rpoB mutation, where alterations in the binding of the regulator subsequently altered gene expression of regulated operons. For example, a changed CRP activation was found in all lineages due to elevated levels of cAMP in the uKOs. CRP was not differentially expressed in any of the lineages, but restored cAMP levels were mirrored by restored gene expression of TUs solely regulated by CRP-cAMP (Supplementary Data 5). ArcA provided another example for global TF activation without a significant gene expression change. The restored activation profile of ArcA and several other iron–sulfur cluster homeostasis TFs found in all lineages could be linked to changes in TCA cycle intermediates as well as quinone pools (e.g., gnd and sdhCB). The ArcAB two-component system in particular modulates genes in response to changes in respiratory conditions that are communicated via the intermembrane quinone pools.

Pathway-specific TF activation was also identified in the uKOs. A change in activation of the PurR regulator was found in pgi and several other lineages due to changed levels of purine degradation products. Specifically, the purR dimer binds hypoxanthine and guanine, and regulates genes involved in purine metabolism. The concentration profiles of hypoxanthine and/or guanine matched the expression profile for purR-target genes, while the expression profile for purR itself did not (Supplementary Data 4). In another example, the change in activation of TyrR in many lineages was found to be attributed to the change in levels of L-tyrosine and L-phenylalanine. The component profile of L-tyrosine and L-phenylalanine and modulates genes involved in aromatic amino acid production and transport. The component profile of L-tyrosine was found to match the expression of aroF. The component profile of L-tyrosine and aroF was also consistent with TyrR activation by L-tyrosine and regulation of aroF gene expression, which indicates that aroF gene expression was modulated by L-tyrosine levels via TyrR. Expression of aroF is controlled only by TyrR. Another example of pathway-specific TF activation involved the use of small regulatory RNA. A sugar phosphate toxicity response was generated by abnormal elevations in glucose 6-phosphate (g6p) and an imbalance of the glycolytic intermediates in uPgi. SgrR is thought to bind hexose phosphates and induce the expression of the small RNA sgrS. The expression profile for purR-target genes, while the expression profile for purR itself did not (Supplementary Data 4).

The common perturbation of TFs by small molecules indicated that the initial flux distribution of the uKO strains following perturbation were often suboptimal, and required a change primarily in flux distribution and secondarily in flux capacity in order to restore fitness in the eKO strains.

**Perturbed metabolite levels triggered TRN responses in uKOs.** Transcriptional regulatory network (TRN) responses in uKOs that were associated with carbon metabolism, nitrogen metabolism, iron regulation, oxidative stress, DNA repair, and other stress responses that control the majority of known functions in E. coli were linked to corresponding changes in regulatory metabo-

litism, iron regulation, oxidative stress, DNA repair, and other stress responses that control the majority of known functions in E. coli were linked to corresponding changes in regulatory metabolite levels (see Methods). Perturbed metabolite levels were traced to known TRN responses by mapping measured metabolite profiles to metabolite-activated transcription factors (TFs). The relationship (i.e., positive or negative) between a metabolite profile, a TF that interacts with the metabolite, and the expression profiles of the transcription units (TUs) regulated by the TF (see Methods, Fig. 5) were compared. Strong evidence (i.e., statistically significant gene expression pattern for genes that are regulated by a single TF, see Methods) for changed TF activation profiles (analogous to the system component profiles, Fig. 2) were identified for 75 TFs (Supplementary Data 2, Fig. 5). These included 7 global TFs (i.e., CRP, Fis, IHF, ArcA, Lrp, FNR, and HNS) and 68 pathway-specific TFs (see Methods). The activation profiles of 15 TFs (which included the 7 global TFs and the 8 pathway-specific TFs ArgR, CpxR, Cra, Fur, NsrR, OxyR, PhoB, and TyrR) were changed across all lineages. The remaining 60 TFs appeared to be changed in a perturbation and lineage-specific manner.

Interestingly, TF activation and TF gene expression was not coincidental (ave ± std 5.4 ± 3.8, 4.1 ± 2.6, and 70.5 ± 6.1% agreement, disagreement, no significant change in expression profile per lineage, respectively, Supplementary Data 3). This result indicated that changed TF activation was mostly attributed to changed concentrations in their metabolic activators as opposed to changed TF gene expression levels. Similar observations have been made for sigma factors and the expression levels of sigma factor DNA binding operons in response to a key rpoB mutation, where alterations in the binding of the regulator subsequently altered gene expression of regulated operons. For example, a changed CRP activation was found in all lineages due to elevated levels of cAMP in the uKOs. CRP was not differentially expressed in any of the lineages, but restored cAMP levels were mirrored by restored gene expression of TUs solely regulated by CRP-cAMP (Supplementary Data 5). ArcA provided another example for global TF activation without a significant gene expression change. The restored activation profile of ArcA and several other iron–sulfur cluster homeostasis TFs found in all lineages could be linked to changes in TCA cycle intermediates as well as quinone pools (e.g., gnd and sdhCB). The ArcAB two-component system in particular modulates genes in response to changes in respiratory conditions that are communicated via the intermembrane quinone pools.

Pathway-specific TF activation was also identified in the uKOs. A change in activation of the PurR regulator was found in pgi and several other lineages due to changed levels of purine degradation products. Specifically, the purR dimer binds hypoxanthine and guanine, and regulates genes involved in purine metabolism. The concentration profiles of hypoxanthine and/or guanine matched the expression profile for purR-target genes, while the expression profile for purR itself did not (Supplementary Data 4). In another example, the change in activation of TyrR in many lineages was found to be attributed to the change in levels of L-tyrosine and L-phenylalanine. The component profile of L-tyrosine and L-phenylalanine and modulates genes involved in aromatic amino acid production and transport. The component profile of L-tyrosine was found to match the expression of aroF. The component profile of L-tyrosine and aroF was also consistent with TyrR activation by L-tyrosine and regulation of aroF gene expression, which indicates that aroF gene expression was modulated by L-tyrosine levels via TyrR. Expression of aroF is controlled only by TyrR. Another example of pathway-specific TF activation involved the use of small regulatory RNA. A sugar phosphate toxicity response was generated by abnormal elevations in glucose 6-phosphate (g6p) and an imbalance of the glycolytic intermediates in uPgi. SgrR is thought to bind hexose phosphates and induce the expression of the small RNA sgrS. The expression profile for purR-target genes, while the expression profile for purR itself did not (Supplementary Data 4).

The common perturbation of TFs by small molecules indicated that the initial flux distribution of the uKO strains following perturbation were often suboptimal, and required a change primarily in flux distribution and secondarily in flux capacity in order to restore fitness in the eKO strains.
Component profiles revealed competing layers of regulation. Cells contain multiple levels of counteracting regulatory mechanisms that often overlap. For example, a relatively low agreement between changes in gene expression profiles and metabolic flux profiles (i.e., gene–protein–reaction association, GPR) within each lineage was found (Supplementary Data 2). Specifically, an average agreement of 27.5% (stdev = 17.4%, n = 22) and average disagreement of 11.5% (stdev = 6.8%, n = 22) was found. A similarly low agreement between types of literature-derived regulation were found (Supplementary Data 2). These findings are consistent with previous work and can be explained by the actions of multiple and competing layers of regulation.

Disagreements were found to categorize into three main groups: (1) counteracting regulatory mechanisms, (2) evidence for inaccurate or incomplete knowledge of regulatory networks, and (3) changes to regulation introduced through fixed mutations. Evidence of competing layers of regulation for 89 regulators (i.e., any biological component that can affect a change in another component, e.g., TF or small-molecule) across 5887 regulated entities (i.e., any biological component that is subject to regulation, e.g., TU or enzyme) were found. Evidence of inaccurate or incomplete knowledge of the regulatory network in 38 regulators across 631 regulated entities were found (Supplementary Data 3). While it is unfeasible to investigate each discrepancy here, specific examples are given that illustrate the above three mechanisms.

In an example of counteracting regulatory mechanisms, a hierarchy of TF control over gene expression was recapitulated. The activation profile of Fur was found to conflict with its consensus activation profile of the pyrD promoter in all of the pgi lineages, whereas the PurR activation profile was found to agree with pyrD expression profiles. This indicated that pyrD expression was dominated by PurR regulation. In another example, the restored activation of sgrS found in the pgi lineages and a novel activation of sgrS found in the ptsHlcrr endpoints 1 and 3 negated the transcription factor regulation of sgrS target genes. In another example, the activation profile of cAMP-CRP was found to conflict with its consensus activation profile.
on the gapA promoter in all of the tpiA lineages, whereas the Cra activation profile of gapA was found to be consistent with gapA expression data (Fig. 4d) and the global regulator of nitrogen metabolism,72 was found to dominate over the regulation of the methylglyoxal operon by ppGpp74,75.

Mutations selected during ALE changed many global regulators. For example, 17% of mutations affected regulators of carbon transport and metabolic processes that appear to offset the activation of operons induced by CRP-cAMP. These included mutations to galR, maltT, and crr in the ePgi strains that appeared to negate repression of galR controlled operons. The mutations may give the evolved strains an additional route to import and catabolize glucose because the galactose importer also has the ability to import glucose albeit with lesser affinity than galactose. In addition, the mutation may have improved the fitness of the ePgi strains by increasing the availability of phosphoenolpyruvate (pep) for aromatic amino acid production. Interestingly, mutations in galR or at the galR operon in ePtsHcr2/04 and in ePgiA01/03 also resulted in the upregulation of GalR controlled genes. The prevalence of galR mutations may indicate that expression of the gal regulon may aid in increasing fitness when the ability to import glucose is impaired or the levels of pep are inadequate for aromatic amino acid production. Additional mutations that affected carbon transport processes included ptsG, galR, nagC in the ePtsHcr strains, and ptsG, galR, and nagA, nagC, and nagE in the ePgiA strains.

A series of mutations were also identified that altered protein homeostasis networks, two-component systems, small RNA networks, and the sigma factor networks. These included mutations that altered the Lon protein homeostasis network in ePgi and the two-component system RcsA/RcsB in ePtsHcr that targeted pathways involved in cell motility, acid resistance, and cell wall biosynthesis. Mutations that altered the SPF small RNA networks, RpoC core RNA polymerase unit, and RpoD sigma factor networks in ePgi were found. Alterations to stress response systems that included Soxx/SoxR in pgI and PhoB/PhoR in tpiA involving oxidative stress and phosphate stress, respectively, were also found.

Mutations were also identified that changed the regulation of pathway-specific TFs. These occurred in a KO-specific manner, and appeared to optimize specific pathways at the regulatory level. For example, the expression of the methylglyoxal pathway in ePgi strains were altered to more efficiently convert methylglyoxal to lactate through mutations that altered methylglyoxal detox pathway gene expression. These examples of global and pathway-specific regulatory shifts indicated that mutations that affect hubs in complex regulatory networks are common in adaptive evolution37, and provide a fitness advantage by rewiring regulatory network responses that may no longer be optimal for fitness.

Rarer were mutations that introduced innovations that appeared to target-specific metabolite imbalances. For example, the levels of nadph, which is used to drive biosynthesis, was affected in many of the KOs. Mutations were found in the transhydrogenases in several of the ePgi strains and in all of the eGnd strains to compensate for an overproduction and underproduction of NADPH, respectively. A mutation found in the active site of seven of the eight ePgi endpoints in isocitrate dehydrogenase appeared to alter cofactor specificity to allow for the use of nadh.

Discussion
Taken together, the combination of study design, automated ALE, multi-omic data sets, and statistics and bioinformatics revealed common mechanisms of adaptation whereby imbalances in metabolite levels from altered fluxes triggered a multitude of network responses that were readjusted by mutations selected for during evolution (Fig. 7). The mutations that fixed during adaptation acted to rewire many existing hardwired responses and/or introduce novel network functions that addressed the imbalances that the initial KO lesion created. The findings of this
study represent a step towards developing a fundamental understanding of how cells mechanically adapt to gene loss from a systems perspective that accounts for proximal and distal relationships in the metabolic and regulatory network. Novel mechanisms and inconsistencies, revealed through adaptation, between measurement and known regulatory mechanisms identified in the case studies present opportunities for future discovery (Supplementary Data 2 and 4). Specific avenues of exploration may include the effect of regulation acting on different time-scales (i.e., transcriptional vs. allosteric regulation) or the effect of RNA and protein stability and degradation that were not addressed in this study.

Methods

Biological material. A glucose, 37 °C, evolved E. coli MG1655 (ATCC 700926) served as the starting strain. Lambda-red-mediated polymorphism (SNP). MG1655 (ATCC 700926) served as the starting strain. Lambda-red-mediated DNA mutagenesis was used to create the knockout strains. Knockouts were confirmed by PCR and DNA resequencing. Genes gnd, ptsH, ptsI, crr, sdhC, sdhA, sdhD, sdhC, tpiA, and pgi encoding for the reactions of 6-phosphogluconate dehydrogenase (GND), phosphotransferase sugar import (GLCptspp), succinate dehydrogenase complex (SUCDi), triosephosphate isomerase (TIPI), and phosphoglucone isomerase (PGI) were removed. PPC was also deleted, but resulted in an auxotrophy for asp-1, and was not included in the study. Genes aceE, aceF, zwf, and atpI-A encoding for the reactions of PDH, G6PDH2r, and ATPS4rpp could not be removed using the method of Datsenko et. al. All cultures were grown in a 50 mL of working volume in a 50 mL autoclaved tube. The cultures were maintained at 37 °C on a heat block and aerated using magnets.

Materials and reagents. Uniformly labeled 13C glucose and 1-13C glucose were from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA). Unlabeled glucose and other reagents were from Sigma-Aldrich (St. Louis, MO). LC-MS/MS reagents were from Honeywell Burdick & Jackson® (Muskegon, MI), Fisher Scientific (Pittsburgh, PA) and Sigma-Aldrich (St. Louis, MO).

Reaction knockout selection. iJO1366 was used as the metabolic model for E. coli metabolism; GLPK (version 4.57) was used as the linear program solver. MCMC sampling was used to predict the flux distribution of the optimized reference strain. Uptake, secretion, and growth rates were constrained to the measured average value ± SD. Potential reaction deletions were ranked by (1) averaged sampled flux, (2) the number of immediate upstream and downstream metabolites that could be measured, (3) the number of genes required to produce a functional enzyme. Reactions involved in sampling loops, that were spontaneous, were computationally or experimentally essential, or were not actively expressed under the experimental growth conditions were not included in the analysis. Also, reactions that would require more than one genetic alteration to abolish activity.
to correlate peak height ratio to absolute concentration. Quality Control samples that were composed of all biological replicates were run twice a day to check the consistency of periodically re-calibrated standards. All biological replicates were run with minimal alterations was used for all MFA estimations and confidence intervals were used as lower and upper bound reaction constraints for further constraint-based analyses. MFA derived constraints that violated optimality were discarded and re-sampled. The descriptive statistics (i.e., mean, median, interquartile ranges, min, max, etc.) for each reaction for each model were calculated from 5000 points sampled from 5000 steps using optGP sampler (version 1.1.3), which resulted in an approximate sampling resolution of 0.5 for all models. A permuted P-value < 0.05 and geometric fold-change of sampled flux values > 0.001 were used to determine differential flux levels, differential metabolite utilization levels, and differential subsystem utilization levels between models. Demand reactions and reactions corresponding to unassigned, transport; outer membrane porin, transport; inner membrane, inorganic ion transport and metabolism, transport; outer membrane, nucleotide salvage pathway, oxidative phosphorylation were excluded from differential flux analysis. The geometric fold-change of the mean between models and the reference model were used for hierarchical clustering; the median, interquartile ranges, min, and max values of each sampling distribution for each reaction and model were used as representative samples for downstream statistical analyses.

**Transcriptomics.** Total RNA was sampled from triplicate cultures (3 mL of cell broth at an OD₆₀₀ ~1.0) and immediately added to 2 volumes Qiagen RNA-protect Bacteria Reagent (6 mL) vortexed for 5 s, incubated at room temperature for 5 min, and immediately centrifuged for 10 min at 17,500 RPMs. The supernatant was decanted and the cell pellets were stored at approximately -80 °C. Cell pellets were then incubated with Readylyse Lysozyme, Superselase, Proteinase K, and 20% SDS for 20 min at 37 °C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns. The RNA was quantified using a NanoDrop and checked for quality using an RNA-nano chip on a bioanalyzer. The RNA was then used as input for downstream statistical analyses. A KAPA Stranded RNA-Seq Kit (Kapa Biosystems KK8401) was used following the manufacturer’s protocol to create sequencing libraries with an average insert length of ~300 bp for two biological replicates. Libraries were run on an Illumina HiSeq 2500 system with an average insert length of ~300 bp for two biological replicates. Libraries were run on an Illumina HiSeq 2500 system with an average insert length of ~300 bp for two biological replicates. Libraries were run on an Illumina HiSeq 2500 system with an average insert length of ~300 bp for two biological replicates.

**Metabolomics.** Uniformly labeled E. coli cell extracts were used as internal standards. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis. The same batch of internal standards with an RSD ≥ 50% were not used for analysis.
approach as coded in the R package Amelia II (version 1.7.4, 1000 imputations). Remaining missing values were filled using the minimum expression level of the data set. Normalized FPKM values for gene expression were log2 normalized to generate an approximately normal distribution prior to any statistical analysis. All replicates for a given condition were found to have a pairwise Pearson correlation coefficient of 0.95 or greater.

DNA resequencing. Total DNA was sample from an overnight culture (1 mL of cell broth at an OD600 of ~2.0) and immediately centrifuged for 5 min at 8000 RPMs. The supernatant was decanted and the cell pellet was frozen in the –80°C. Genomic DNA was isolated using a Nucleospin Tissue kit (Macherey Nagel 749592.50) following the manufacturer’s protocol, including treatment with RNase A. Resequencing libraries were prepared using a Nextera XT kit (Illumina FC-131-1024) following the manufacturer’s protocol. Libraries were ran on a MiSeq (Illumina).

DNA resequencing reads were aligned to the *E. coli* reference genome (U00996.2. genbank) using Bresq (version 0.26.0) as populations. Mutations with a frequency of <0.1, P-value >0.1, or quality score <6.0 were removed from the analysis. In addition, genes corresponding to crl, insertion elements (i.e., insH1, insB1, and insA), and the rhs and rse gene loci were not considered for analysis due to repetitive regions that appear to cause frequent miscalcs when using Bresq. mRNA and peptide sequence changes were predicted using BioPython (https://github.com/biopython/biopython.git). Large regions of DNA (minimum of 200 consecutive indices) where the coverage was two times greater than the average coverage of the sample were considered duplicated.

Structural analysis. Corresponding PDR files for genes with a mutation of interest were downloaded from PDR90,100. Structural models for genes for which there were no corresponding PDR files were taken from I-TASSER generated homology models101 or generated using the I-TASSER protocol102. The BioPython predicted sequence changes and important protein features as listed in EcoCyc103 were visualized and annotated using VMD104.

System component statistical feature identification analyses. Network components (i.e., RNA-seq, metabolomics, fluxomics, genomics) were pre-processed as described above, and subjected to a feature identification analysis pipeline. Network components for each lineage were first subjected to a differential test (ref vs. KO, KO vs. endpoints, ref vs. endpoints, and endpoints vs. endpoints). The criteria for significance were: unprocessed data types are as weighted fold change > 0.01 and 0.5 < fold_change < 2.0 as calculated from a t-test of the g-log normalized metabolite concentrations. Transcriptions: q-value >0.05 (FDR corrected P-value) and abs(log2(fold-change)) > 1.0 as calculated from re-sampled flux distributions that were constrained by the 95% confidence intervals derived from estimated MFA flux bounds (demand reactions and reactions in subsystems corresponding to unassigned, transport; outer membrane; porin, transport; inner membrane, inorganic ion transport and metabolism, transport; outer membrane, nucleotide salvage pathway, oxidative phosphorylation were excluded). Mutations: frequency >0.1 (i.e., mutation found in the reference strain and in repetitive regions were excluded). Components that met the significance criteria for any combination of comparisons from the differential test were used in the pairwise PLS-DA analyses and profile matching. Counts of significant components for each lineage were based on components that met the significance criteria for Ref vs. eKO, or uKO vs. eKO. Network components for each lineage were subjected to pairwise PLS-DA analyses (ref vs. KO, KO vs. endpoints, ref vs. endpoints, and endpoints vs. endpoints). The components with a loadings magnitude within the top 25% of all components and correlation coefficient >0.88 for different combinations of comparison were selected using pairwise PLS-DA analysis.

Network components for each lineage were subjected to profile matching. System component levels between Ref, eKO, and uKO were correlated (Pearson’s R) to six profiles in both positive and negative directions. novel−→novel+, overcompensation−→overcompensation+, partially restored−→partially restored+, reinforced−→reinforced+, restored−→restored+, unrestored−→unrestored+, profiles were expressed in terms of 1−0, 0−1, 1−0, 0−1, 1−0, 1−0, 0−1, 0−1, 1−0, 1−0, 1−0, 0−1, and 0−1−1. System components were binned into profiles when a Pearson correlation coefficient >0.88 was calculated. Only negligible changes in the assignment of profiles were found when using absolute or relative component units (e.g., mmol/gDCW−→log2(FC vs. ref)) or different correlation methods (i.e., Spearman).

System component statistical sample trend analysis. Components identified from the differential tests (except for metabolomics) were used for sample trend analyses. Hierarchical clustering was used to diagnose sample groupings and distances between samples (distance metric of Euclidean and linkage method of complete). Principal component analysis (PCA) as encoded in the R package pcaMethods105 (version 1.64.0, univariate scaling, centering, SVC PCA) was then used as the representative unsupervised method to project samples into component space, and confirm the relative magnitude and direction of component weights. PCA models were first constructed for the reference, knockout, and endpoint for each of the lineages to confirm that the primary component best separated the reference and endpoint from the knockout, and that the second component best separated the knockout from the endpoint. PCA models were then constructed for the reference, knockout, and all endpoints for each network perturbation. The PCA models were validated using cross validation (CV type of k-fold, default 5 segment with 5 CV runs per segment with minimum number of segments equal to the number of samples). Partial Least Squares Discriminatory Analysis (PLS-DA) was implemented using the R package plsRglm (version 2.5, univariate scaling, centering, Canonical Powered Partial Least Squares (cppLS) PLS-DA was used to project replicate samples into component space. PLS-DA models were first constructed for the reference, knockout, and endpoint for each lineage to confirm that the primary component best separated the reference and endpoint from the knockout, and that the second component best separated the reference and knockout from the endpoint. PLS-DA models were then constructed for the reference, knockout, and all endpoints for each network perturbation. The PLS-DA models were validated using cross validation (default 10 segments with minimum number of segments equal to the number of samples).

Metabolite, flux, and gene set enrichment analyses. Metabolite and gene set enrichment analyses were conducted using the subsystem categories of JGI36. Flux and metabolite flux sum set enrichment analyses were conducted using the subsystem categories of iDM2015. A P-value < 1e−3 (hypergeometric test) was used as a threshold for enrich for subsystems. Gene set enrichment analysis differentially expressed genes was also performed using R package topGO107 with GO annotations for E. coli108. A P-value <0.05 (Fischer statistic, parent–child algorithm109) was used to test for enriched biological processes and molecular functions.

Network distance and graph analyses. The inverse mean values from sampled flux distributions that were constrained by the 95% confidence intervals derived from estimated MFA flux bounds (demand reactions and reactions in subsystems corresponding to unassigned, transport; outer membrane; porin, transport; inner membrane, inorganic ion transport and metabolism, transport; outer membrane, nucleotide salvage pathway, oxidative phosphorylation were excluded). Mutations: frequency >0.1 (i.e., mutation found in the reference strain and in repetitive regions were excluded). Components that met the significance criteria for any combination of comparisons from the differential test were used in the pairwise PLS-DA analyses and profile matching. Counts of significant components for each lineage were based on components that met the significance criteria for Ref vs. eKO, or uKO vs. eKO. Network components for each lineage were subjected to pairwise PLS-DA analyses (ref vs. KO, KO vs. endpoints, ref vs. endpoints, and endpoints vs. endpoints). The components with a loadings magnitude within the top 25% of all components and correlation coefficient >0.88 for different combinations of comparison were selected using pairwise PLS-DA analysis.

Network components for each lineage were subjected to profile matching. System component levels between Ref, eKO, and uKO were correlated (Pearson’s R) to six profiles in both positive and negative directions. novel−→novel+, overcompensation−→overcompensation+, partially restored−→partially restored+, reinforced−→reinforced+, restored−→restored+, unrestored−→unrestored+, profiles were expressed in terms of 1−0, 0−1, 1−0, 0−1, 1−0, 1−0, 0−1, 0−1, 1−0, 1−0, 1−0, 0−1, and 0−1−1. System components were binned into profiles when a Pearson correlation coefficient >0.88 was calculated. Only negligible changes in the assignment of profiles were found when using absolute or relative component units (e.g., mmol/gDCW−→log2(FC vs. ref)) or different correlation methods (i.e., Spearman).

Biomass to network component correlation analysis. EcoCyc108 subsystem categories for the following biomass producing pathways were used in the analysis: amines and polyanimes biosynthesis, amino acids biosynthesis, nucleosides and nucleotides biosynthesis, fatty acid and lipid biosynthesis, cofactors, prosthetic groups, electron carriers, energy metabolism, cell structures biosynthesis, and carbohydrates biosynthesis. Gene identifiers from these pathways were mapped onto iDM2014 via the GPR relationship to identify biomass producing reactions and metabolites. The analysis was conducted at the level of individual lineages using the system component profiles of restored−→novel+, overcompensation−→partially restored−→reinforced−→unrestored, growth inhibition (i.e., growth promoting) and negatively correlated (correlation coefficient < -0.88, Pearson, r) with growth (i.e., growth inhibiting). The number of significant component compositions were divided by the number of measured biomass components, and expressed as a percent. A direct pairwise correlation between metabolism component units (e.g., mmol/gDCW) and the growth rate (units of log2(FC vs. ref)) between the reference strain, knockout, and endpoints for all or each knockout condition for comparison (data not shown). Components that were
positively correlated (correlation coefficient > 0.88, Pearson, r) with growth rate or negatively correlated (correlation coefficient > 0.88, Pearson, r) with growth rate were identified.

Inter- and intra-component correlation analysis. A global pairwise correlation between metabolite concentrations, transcript levels, and fluxes was performed by comparing the agreement and disagreement between component profiles of restored+, novel-, overcompensation+, partially restored-, unstored+, and reinforced+. Components with matching profiles with correlation coefficients > 0.88 (Pearson, R) were correlated; with components matching profiles with correlation coefficients < –0.88 (Pearson, R) were anti-correlated. A similar global pairwise correlation between metabolite concentrations, transcript levels, and fluxes was performed (units of log2(FC vs. ref)) for comparison (data not shown). Components with a correlation coefficient > 0.88 (Spearman, r) were correlated; Components with a correlation coefficient < –0.88 (Spearman, r) were anti-correlated.

Regulation to network component correlation analysis. Significantly correlated components were compared to annotated gene-to-reaction, and metabolite-to-reaction interactions annotations in iJO1366, and to annotated transcription factor-to-gene, metabolite-to-transcription factor, metabolite-to-transcription factor-to-gene, metabolite-to-transcript, and metabolite-to-reaction regulatory interactions from the EcoCyc database. EcoCyc database identifier were mapped to iJO1366 identifiers using a combination of ChEBI, MetaNetX, EC numbers, InChI strings, and manual curation. The mode of component interactions were encoded as either positive for reactant-reaction, activating, or stabilizing interactions, or negative for product-reaction, inhibiting, or de-stabilizing interactions. The sign and magnitude of the correlation coefficient (Pearson, r) of matching categories was compared to the mode of interaction to determine agreement (correlation coefficient > 0.88 and positive mode, or correlation coefficient < –0.88 and negative mode). The inverse was used to determine disagreement.

The classification of global regulators follows the definition given by Martinez-Antonio et al. Global transcription factors are defined to include CRP, IHF, FNR, FIS, ArcA, Lrp, and Hns. A secondary level of regulators are defined to include NarL, NarX, MicF, CapR, RobR, PurR, PhoB, CrpX, and SoxH. The secondary level and lower level regulators (e.g., local transcription factors) were further broken into classes for local and general stresses.

Regulator activation categorization. A profile for the activation status of each regulator for each knockout evolution was determined. The analysis was first limited to regulated entities that had only a single annotated regulator. The analysis was then expanded to include all regulators and regulated entities. A category weight for each regulated entity for each endpoint was calculated as follows: weight, fi = abs(correlation, ij) / (1 + number of regulators, k) where i = endpoint, j = category, k = regulators, n = number of endpoints per knockout evolution, correlation = correlation coefficient, nRegulators = number of regulators per regulated entity. A confidence score for each regulator for each knockout was calculated as follows: confidence = sum(weight, i,j,k) where i = knockout, j = endpoint, and k = regulated gene. A higher confidence score indicates a consistently higher correlation to the category across all regulated entities that are regulated by the regulator.

Code availability. Published software used in this study are noted in the Methods. Custom software used for the analyses presented in this study are deposited on Github (https://github.com/dmcloskey).

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

References:


Acknowledgements
We thank José Urría for helpful discussion and guidance when implementing the knockouts in the pre-evolved strain. We thank Jamey Young for helpful discussions throughout the MFA analysis. We thank Laurence Yang for helpful discussions regarding optimization and statistical analysis. This work was funded by the Novo Nordisk Foundation Grant Number NNFI00CC1016517.

Author contributions
D.M. designed the experiments; generated the strains; conducted all aspects of the metabolomics, fluxomics, phenomics, transcriptomics, and genomics experiments; performed all multi-omics statistical, graph, and modeling analyses; and wrote the manuscript. T.E.S. ran the ALE experiments. E.B. assisted with structural analysis. R.S. processed the DNA and RNA samples. S.X. assisted with metabolomics and fluxomics data collection, sample processing, and peak integration. Y.H. assisted with fluxomics data collection and sample processing. A.M.F designed and supervised the evolution experiments, and contributed to the data analysis and the manuscript. B.O.P conceived and outlined the study, supervised the data analysis, and co-wrote the manuscript.

Additional information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41467-018-06219-9.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2018