ARTICLES
https://doi.org/10.1038/s41589-020-0593-y
nature chemical biology

A quantitative method for proteome reallocation using minimal regulatory interventions

Gustavo Lastiri-Pancardo[1], Jonathan S. Mercado-Hernández[1], Juhyun Kim[2], José I. Jiménez[2] and José Utrilla[1]✉

Engineering resource allocation in biological systems is an ongoing challenge. Organisms allocate resources for ensuring survival, reducing the productivity of synthetic biology functions. Here we present a new approach for engineering the resource allocation of *Escherichia coli* by rationally modifying its transcriptional regulatory network. Our method (ReProMin) identifies the minimal set of genetic interventions that maximizes the savings in cell resources. To this end, we categorized transcription factors according to the essentiality of its targets and we used proteomic data to rank them. We designed the combinatorial removal of transcription factors that maximize the release of resources. Our resulting strain containing only three mutations, theoretically releasing 0.5% of its proteome, had higher proteome budget, increased production of an engineered metabolic pathway and showed that the regulatory interventions are highly specific. This approach shows that combining proteomic and regulatory data is an effective way of optimizing strains using conventional molecular methods.

The removal of accessory nonessential functions is one of the strategies used to engineer microbial phenotypes. This approach relies on the assumption that cellular resources for gene expression are limited and, therefore, by removing genes unneeded in a certain environment, the cell is capable of allocating resources to other functions (for example, expression of recombinant genes). These minimization approaches are mostly done by reducing genome size and gene number including performing random deletions[1,2], however, the precise way in which the resource allocation takes place after the genetic intervention is not considered.

Organisms respond to the environment by cellular signaling pathways encoded in regulatory networks[3]. The intricacy of the lifestyle of an organism is generally translated into signaling complexity[4]. Biological regulatory networks are robust[5] and evolvable[6] to cope with environmental and lifestyle perturbations, however, this robustness involves intrinsic trade-offs, such as resource allocation strategies. It has been shown that cellular states are naturally 'primed' for typical upcoming changes. Bacteria anticipate to fluctuations in the environment[7,8] draining resources from functions that are mostly performed in relatively stable conditions. The expression of anticipation functions, also called hedging functions, is encoded in the regulatory network and has a proteomic cost[9]. Genome-scale models along with experimental datasets enable the calculation of the minimal theoretical proteome needed to sustain growth in a defined condition[10]. Therefore, comparing minimal theoretical proteomes with measured proteomes reveals the costs of the hedging proteome allocation. Proteome econometric approaches can facilitate the engineering of cellular states or phenotypes aimed at displaying an engineered function. Recent studies have focused on the host–construct interactions for increasing predictability of synthetic constructs[11–13]. In addition to these approaches, the rational design of the host used for expression following econometric models can be adopted to improve the performance of synthetic constructs, including production phenotypes for molecules of added value. Among other benefits, streamlined organisms obtained this way are less likely to develop undesired emerging behaviors[11].

We foresee the use of regulatory mechanism as a control layer that will aid in the design of cellular phenotypes. Our ability to engineer biological systems depends on understanding how cells sense and respond to their environment at a system level. Few studies have tackled this issue and none of them in a rational way.

In this work we developed a new top-down cell engineering strategy for *E. coli* using the transcriptional regulatory network (TRN) as a control layer for proteome allocation. By combining high-throughput proteomic information, regulatory network interactions and gene essentiality observations, we implemented a method capable of finding the minimal set of genetic interventions required to divert the resources invested in superfluous hedging into increased biosynthetic potential. The resulting strains exhibited an increased availability of cellular resources to express engineered functions.

Results

Identification of dispensable TFs for proteome release. The genome-scale model of metabolism and gene expression (ME-model) computes the minimal theoretical proteome and allows calculating the cost of expressing hedging functions. It can be used to simulate different scenarios of expression of the hedging proteome (as unused protein fraction coefficient, see Methods)[14]. These simulations allowed us to calculate the costs and the benefits of different interventions, for example by modulating the expression of the hedging proteome, expressed in terms of growth, the size of both essential and recombinant proteome sectors (Extended Data Fig. 1).

We built on ME-models to design strains containing the minimal genetic interventions that reduced the greatest amount of proteomic resources not required to grow in a specific condition. Our method used transcription factors (TFs) as the key dials controlling the allocation of the hedging proteome in a predefined specific environment. We began by establishing batch growth in minimal medium (M9) supplemented with glucose as the sole carbon source as the defined environment for the first case of this study. Then, by compiling experimental and genome-scale model generated essential gene sets, we generated a list of essential genes for growth in this specific

[1]Systems and Synthetic Biology Program, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico. [2]Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK. ✉e-mail: utrilla@ccg.unam.mx

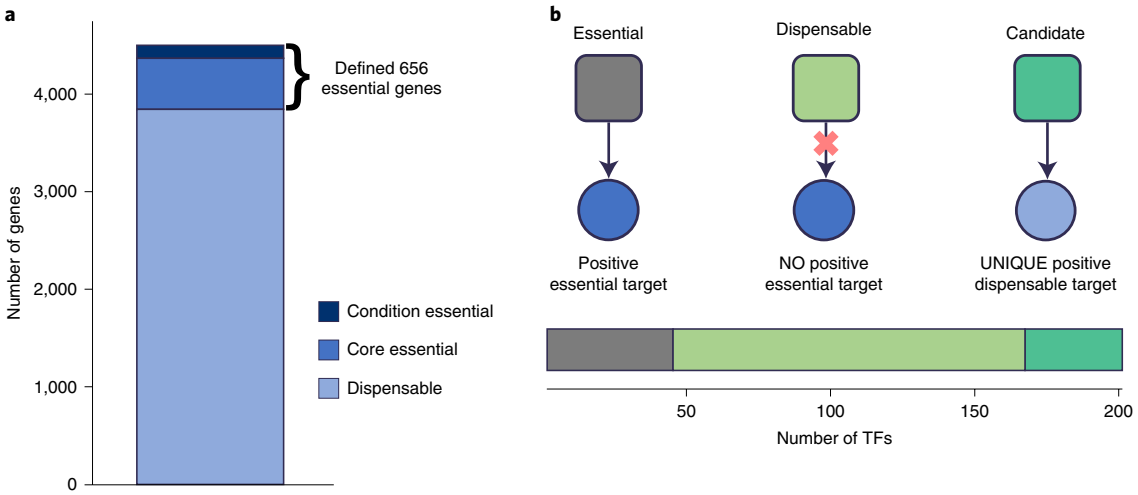


Fig. 1 | Gene essentiality and TRN analysis in the predefined condition. a, Essentiality profile of the *E. coli* genes (~4,600) under selected growth condition. The essential genes considered for this analysis are divided into core essential (523 genes, always needed) and conditional (133 genes, M9-glucose needed). **b**, Graphical representation of the subnetwork of TF–gene interactions considered for the classification of the TFs; gray squares represent essential TFs, light green squares dispensable TFs, dark green squares candidate TFs, dark blue circles essential genes, light blue circles dispensable genes and arrows positive interactions. The lower bar shows the number of TFs in each classification from the total considered (203 TFs).

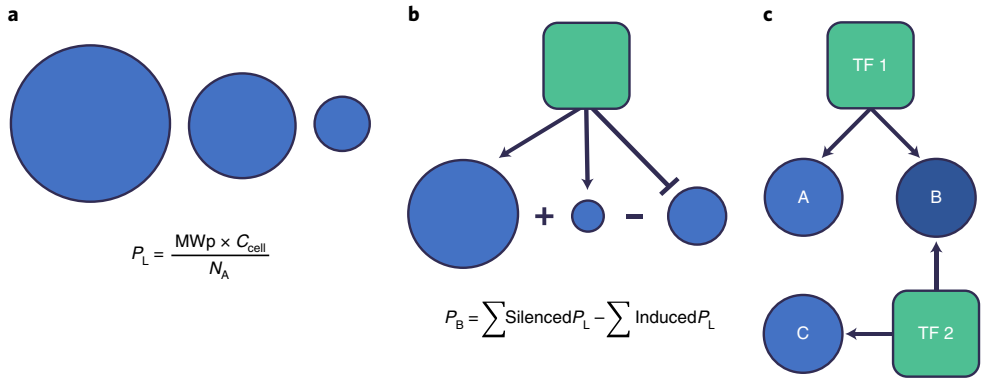


Fig. 2 | Emerging properties from proteomics data integration. a, The P_L of a gene is defined as the molecular weight of the protein (MW_P) multiplied by the number of copies per cell (C_{cell}) divided by Avogadro's number (N_A) (6.022×10^8 fg equivalent). The more expressed the gene is, the more P_L it generates. **b**, The P_B of a TF defined as the sum of the P_L of the silenced genes minus the sum of the P_L of the induced genes. **c**, Schematic of a simple case of shared regulation in which removing both TFs silences all target genes but this is not the case when the TFs are silenced individually.

environment (Fig. 1a and Supplementary Table 1, see Methods). Once the case-specific gene essentiality was defined, we analyzed the TF–gene interactions compiled in RegulonDB[15]. After determining gene essentiality and TF–gene regulatory interactions, we analyzed the subnetwork of interactions of each TF looking for dispensable TFs, defined as those that do not activate the expression of any essential gene. According to our analysis, 156 of the 200 TFs contained in the regulatory network can be eliminated (Fig. 1b). Since our goal was to reduce the hedging proteome, out of the 156 dispensable TFs we selected as candidates for nonessential function reduction the 34 TFs with at least one unique (meaning it is not activated by any other TF) positive regulated gene (Supplementary Table 2) (see Methods); this gives the certainty of silencing at least one gene.

Integration of proteomic and regulatory network data. We determined the proteome associated to each nonessential TF in our network integrating a quantitative proteomic dataset[16] that provides protein copy number per cell under 22 different growth conditions with 95% of proteome coverage (by mass). Here we defined two emerging properties derived from the quantitative proteomics data integration: the proteomic load of a gene (P_L) in femtograms (fg) of protein per cell (Fig. 2a) and the proteomic balance (P_B) of a TF resulting from the summation of the P_L of the genes that would result silenced or activated by the elimination of a TF (Fig. 2b). P_B is conceptually important to rank the TFs according to the size of the proteome they control, since it considers the net addition of protein mass (in fg of protein per cell) liberated when removing a TF.

Computational search of minimal TF eliminations. Many TFs have shared target genes (Fig. 2c), in fact, many of them are part of a simplified version of a dense overlapping regulon networks motif[17], meaning that a particular combination of TFs is needed to 'fully' silence these targets, generating a full landscape of potential proteome liberation composed by all the different combinatorial TFs deletions.

To assist with the design of the mutant strains we developed a computational method. We called our method ReProMin (regulation-based proteome minimization); it integrates the TF–gene

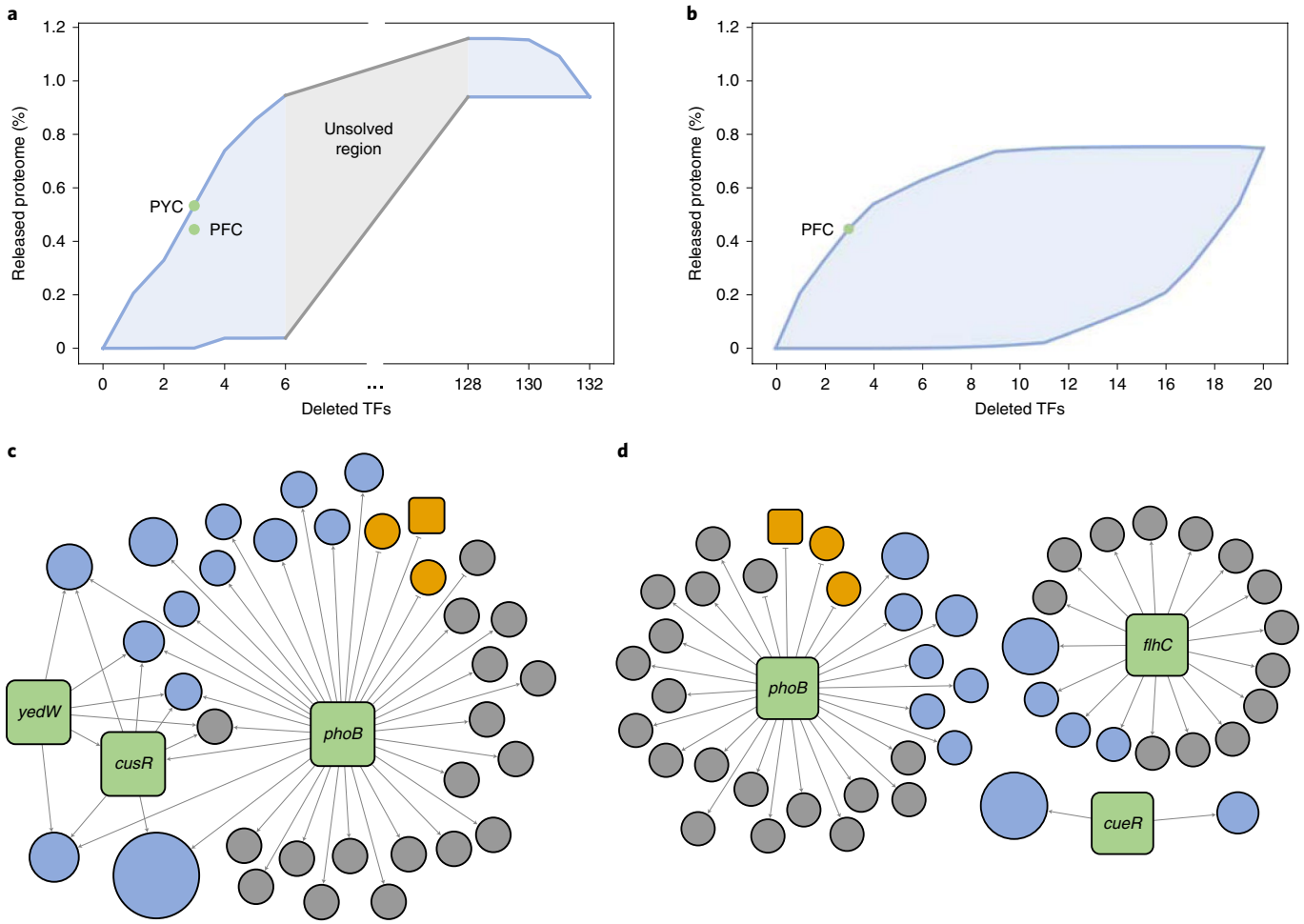


Fig. 3 | Proteome liberation calculations using ReProMin. a, Potential proteome liberation landscape corresponding to the glucose shared-target case; the solved region is shown in blue while the unsolved region in gray. **b**, Solved proteome liberation landscape for the unique-target case in glucose. The locations in the landscapes of the generated mutants are shown with a green circle. Subnetwork of predicted regulated targets of the best three KO **c**,**d**, Shared (**c**) and unique (**d**) target cases. In both cases, green squares represent deleted TFs, blue circles predicted silenced targets, yellow circles predicted induced targets and gray circles genes with no proteomic coverage; the size of the circles is proportional to the P_L of the target. High-resolution labeled versions of the subnetworks are available in Extended Data Fig. 3.

interaction network and quantitative proteomic data to generate and solve the proteome reduction landscape for a particular TF list, finally returning the *n*-size combination of mutations that silences the higher proteomic load in a particular growth condition (see Methods). Using ReProMin we computed some parts of a proteome liberation landscape depending on the number of combinations (Fig. 3a). For a particular TF list, the proteome liberation landscape boundaries are defined by the TF combination that releases the highest and lowest P_L for all the possible TF combination sizes; meaning that the larger the TF list, the larger and complex the landscape space. The computing time required to solve a landscape is defined by the number of possible combinations that increments exponentially as the TF list becomes larger (see equation (2) in Methods)

ReProMin calculations reveal potential proteome release. We used the glucose minimal medium condition as the starting point for the analysis, since we defined the gene and TF essentiality in that condition from experimental data. We considered two cases to do the calculations. The shared target case: nonessential TFs with positive P_B (132 TFs), which takes into account some TFs with no unique regulated genes and the unique target case: considering previously defined candidate TFs with a positive P_B (20 TFs).

For the shared target case the computational tool was able to solve combinations up to six TFs ($>6\times10^9$ combinations roughly taking 95 h of computation, Supplementary Table 3), but we did not evaluate the next case due to its extremely large number of combinations (seven TFs $>11\times10^{10}$ combinations). To maximize the solved landscape space, we continued resolving other areas of the space where the number of combinations is small enough to be solved (when $n\rightarrow0$ or $n\rightarrow r$, see Methods, equation (2)).

Our calculations revealed that the elimination of all 132 nonessential TFs is not the best case of proteome liberation. However, the elimination of 130 TFs would potentially liberate up to 1.06% of the total *E. coli* proteome but up to 0.94% can be liberated by removing the top combination of six TFs, while up to 0.53% of the total proteome can be released by removing a top combination of three TFs (Fig. 3a).

For the unique target case, the elimination of the entire 20 candidate TFs would liberate up to 0.72% of the proteome, and our simulations show that there is not a significant improvement in resource release after the elimination of eight TFs. In fact, 60% of the total potential liberation can be achieved by removing just the top three TFs (Fig. 3b).

We also performed ReProMin calculations for other conditions in which proteomic data is available, such as growth on galactose,

acetate or glycerol + casamino acids (casAA) minimal medium and rich LB medium. In each case, we used the specific set of essential genes (see Methods) and for the environment specific genes we ran essentiality simulations with a genome-scale metabolic model in the corresponding growth condition (see Methods).

As a result, we obtained 164 nonessential TFs for galactose, 166 for acetate, 171 for glycerol + casAA and 165 for rich LB medium. We found that most of the identified dispensable TFs (89%) are shared among all the tested conditions. We identified 22 candidate TFs for galactose and acetate that belong to the unique target case and have a positive P_B. Proteome liberation calculations were made using these subsets of candidate TFs, predictions show that we can release 0.88 and 0.81% of the total proteome in galactose and acetate, respectively, with the deletion of all these TFs (Extended Data Fig. 2a,b). For glycerol, we found 24 candidate TF, liberating up to 2.9% of total proteome (Extended Data Fig. 2c) and finally for rich LB medium we identified 20 candidate TF, with a potential 0.5% of liberation, being the worst condition for potential proteome liberation tested (Extended Data Fig. 2d). This is in agreement with previous proteome usage analysis, where it has been showed that at higher growth rates there is less of the dispensable proteome[10,18].

Generation of combinatorial strains. Even though *E. coli* is one of the most studied organisms and its TRN has been widely investigated, only half of its genes have regulatory information (RegulonDB). We prevented detrimental effects on gene expression due to our incomplete knowledge of the regulatory network by selecting the smallest combination of TFs that liberates the greatest amount of resources.

Based on our ReProMin predictions, we generated two triple knockout (KO) strains for the best combinations of the shared target and unique target cases predicted in glucose minimal medium. In the shared target case, this corresponded to the strain PYC (Δ*phoB* – phosphate scavenging system, Δ*yedW* – unknown gene, Δ*cusR* – copper/silver export system regulator) with a P_B of 1.3 fg representing 0.53% of the total proteome in glucose growth conditions (Fig. 3a). This design was a particular case of shared regulation where most of the target genes are only silenced by the deletion of all the three TFs together (Fig. 3c). In the unique target case the resulting strain was PFC (Δ*phoB*, phosphate scavenging system; Δ*flhC*, flagella master regulator and Δ*cueR*, copper efflux system) with a P_B of 1.08 fg representing 0.44% of the total proteome in glucose (Fig. 3b). The unique target case has a higher degree of confidence in the design than the shared target case due to a simpler regulatory subnetwork, since we noticed that unique target cases mainly belong to the single input module network motif (SIM)[17] (Fig. 3d). We also generated a strain, using an intuitive approach, in which we eliminated three TFs that regulate nongrowth-related functions. To construct this strain, we randomly selected three TFs from those previously found downregulated in strains with regulatory mutations selected by adaptive laboratory evolution (ALE)[9]. The resulting strain was called FOG (Δ(*fliA*, *oxyR*, *gadE*)) the genes code for flagella sigma factor, oxidative stress master regulator and acid resistance regulator, respectively. The FOG strain was not generated by our design pipeline; therefore, the regulatory interventions affect some essential functions (nine essential genes) mainly involved in de novo synthesis of AMP and heme groups and it was used as a control for comparison with the computationally designed strains.

Finally, to test ReProMin in a different environment for which high confidence gene essentiality data is available, we also constructed the best 3KO strain in LB rich medium (Extended Data Fig. 2d). The resulting strain was PYN (Δ*phoB*, phosphate scavenging system; Δ*yqhC*, furfural reduction[19] and Δ*ntrC*, nitrogen regulation two-component system) potentially liberating 0.33% of the total proteome.

RNA-seq confirmed the specificity of introduced mutations. The predictive power of ReProMin depends on the accuracy of the interactions captured in the *E. coli* TRN. We validated the predicted transcriptional changes for the case PFC that showed the greatest proteome release. We measured PFC and the wildtype (WT) strains' transcriptome profiles obtained by RNA-seq in the same environment. This experiment aimed at determining the degree of success in gene silencing at the transcriptional level, and at assessing other possible transcriptional perturbations resulting from our regulatory modifications. Our results showed that no transcripts corresponding to the three deleted TFs were detected in PFC (Extended Data Fig. 4a). By mapping the fold change obtained in the analysis to the predictions of the computational tool, it was possible to visualize the impact at the transcriptional level of the missing regulators on their targets (Extended Data Fig. 4c). Four targets associated to *flhC*, corresponding to genes forming the flagella (*flgB*, *flgC*, *flgE* and *flgG*) were completely silenced; furthermore, all the other flagella-related genes also registered a decrease on their expression. Regarding *phoB*, two targets were successfully silenced (*phnI* and *phnL*), both genes belong to an operon that is induced under phosphate starvation and is required for use of phosphonate and phosphite as phosphorous sources[20], many other targets related to this operon also reduced their expression. On the contrary, *phnK* present in the same operon was overexpressed. We were unable to map any transcripts to six genes belonging to the previously mentioned operon, which may not be entirely expressed in the absence of phosphate starvation. Furthermore, *phoR* (part of the *phoB–phoR* two-component system) also reduced its expression. Finally, both targets of *cueR* (*copA* and *cueO*) also decreased drastically their abundance.

Regarding the accuracy of our ReProMin predictions, 30 genes of 47 predicted silenced genes were silenced at different levels, whereas nine predicted silenced genes presented higher expression values than the WT strain. Carefully reviewing discrepancies among predicted and measured transcriptional changes revealed that some evidences on which the *E. coli* TRNs are built, are computationally predicted or derived from indirect observations. These weak pieces of evidence resulted in false predictions and may be refined if those networks are improved. Finally, transcripts of eight predicted targets were not found in either strain. These observations show that in 72% (28 of 39 measured genes) of the cases the predictions of the computational tool were accurate (Extended Data Fig. 4d).

In addition to the designed transcriptional changes, we found 17 genes differentially expressed (eight downregulated genes and nine upregulated) (Extended Data Fig. 4b and Supplementary Table 4). This RNA-seq analysis shows that besides the intended transcriptional changes, few off-target effects were identified at the transcriptomic level in the PFC strain growing on glucose M9 medium.

To theoretically quantify the actual mass of the liberated proteome fraction, we did an estimation of the translation rates of the gene targets associated to the deleted TFs plus the previously mentioned differentially expressed genes (Fig. 4a) (see Methods).

The calculations showed that the mass of the proteome fraction associated to the removed TFs targets decreased from a 0.48% of the total proteome in the WT to a 0.25% in PFC, CueR being the TF that contributes most to the liberation (Fig. 4b). Initially, ReProMin calculations estimated a liberation of 0.44% of the total proteome in PFC, according to these calculations (with the uncertainty that they may have) this would correspond to a 57% of success in the estimated resource liberation. Liberated resources are presumably redistributed for upregulated genes, which show a larger proteome fraction in the mutant (Fig. 4c), and could become available for the expression of engineered functions.

UT designed strains show increased cellular budget. Our three experimentally generated mutants (shared target case, unique target case and control) were evaluated in rich (LB) as well as in minimal

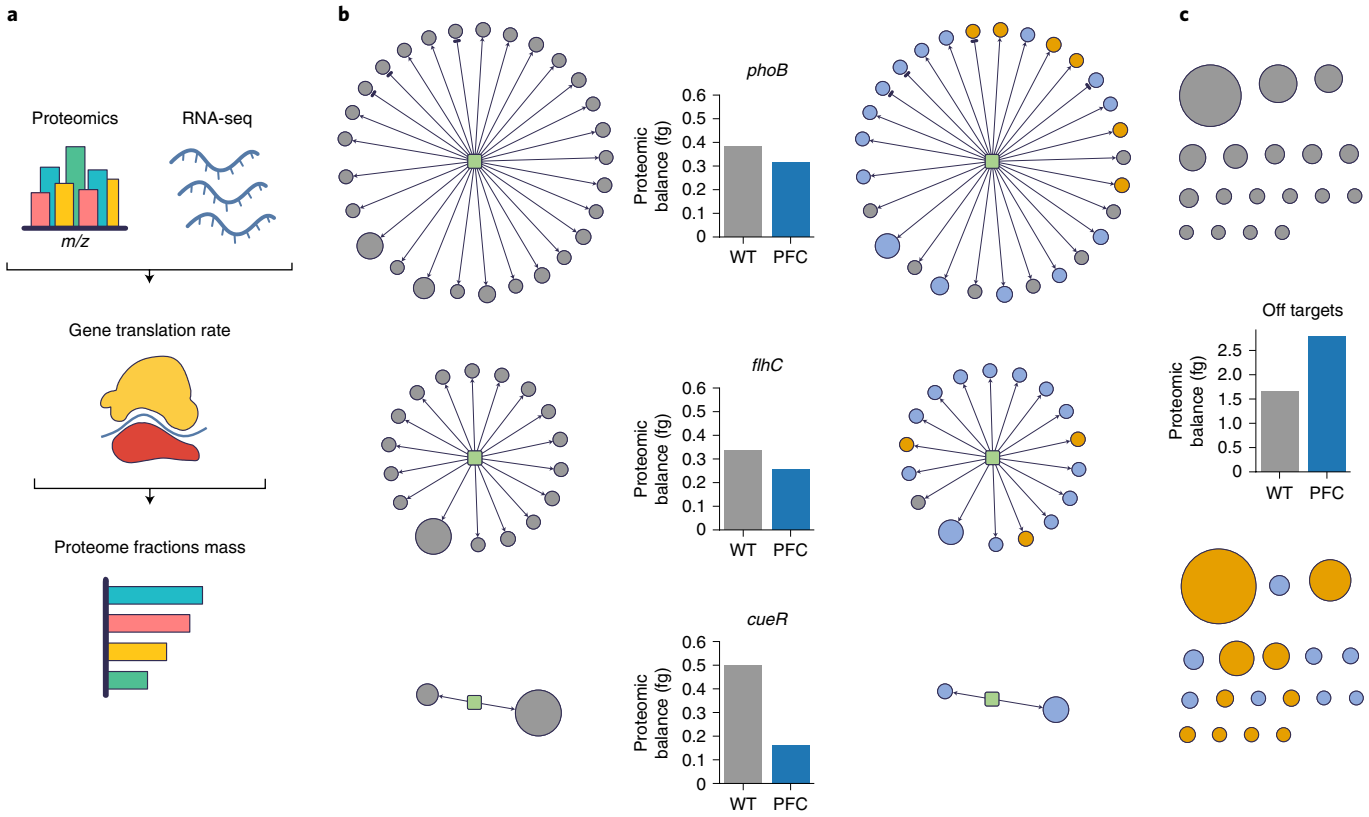


Fig. 4 | Deleted TF proteome fraction mass estimation. a, Pipeline used to estimate the proteome changes from RNA-seq data. **b**, Proteome fraction mass representation and comparison for the three deleted TFs associated targets for WT (left) and PFC (right). **c**, Proteome fraction mass representation and comparison of the identified off-targets for WT (top) and PFC (bottom). Green squares represent deleted TFs; in the case of PFC, blue circles signify targets that reduced their mass compared to the WT, yellow circles targets that increased their mass and gray circles targets not found expressed thus not contributing to the mass; the size of the circles is proportional to the P_L of the target. In all cases, graph bars represent the sum of the mass of all genes in each group.

medium containing three different carbon sources (acetate, galactose, glucose). The ReProMin designed strains (PYC and PFC) showed neither growth defects nor increase in the growth rate or biomass yield in any of the four conditions tested. On the contrary, the control strain (FOG) showed growth defects in all growth conditions tested (Extended Data Fig. 5). This phenotypic defect shown by FOG strain may be the result of the elimination of the principal acid resistance system activator (*gadE*), since we are using batch growth with no pH control. For glucose minimal medium, we also evaluated the effect of recombinant protein production using a plasmid expressing a genetic circuit with two fluorescent reporters (Fig. 5a)[21]. The burden caused by carrying a plasmid was reflected as a decrease in the growth rate in all tested strains (Extended Data Fig. 6a); this decrease is higher when the plasmid was expressing the genetic circuit; however, the burden displayed by both ReProMin designed strains was lower compared to the WT counterpart. Additionally, the PFC strain also showed a higher final biomass production (Extended Data Fig. 6b). It has been described that the expression levels of two protein reporters encoded on the same plasmid but without a regulatory connection between them is captured by a linear relationship that can be interpreted as an isocost line, a concept used in microeconomics to describe how two products can be bought with a limited budget, so the more is used on one, the less can be used on the other. These lines represent the boundary of the production budget of a given strain and growth condition (Fig. 5b)[21]. We obtained the isocost lines at balanced growth, defined as growth during the exponential phase in which a steady-state of the cellular concentration of both green (GFP) and red fluorescent protein (RFP) reporters is achieved (in other words, the cellular concentration of fluorescent protein does not change in time after roughly 5 h) determined by two different methods: mean plate reader fluorescence and mean fluorescence measured by flow cytometry. The line corresponding to PFC strain shows a parallel upward shift compared to the WT strain, which represents an increase of 9% in absolute fluorescence (Fig. 5c), 5% in normalized fluorescence per cell (Fig. 5d) and 12% in mean fluorescence per cell (Extended Data Fig. 7). This difference is increased at the stationary phase of the culture (~22 h) where higher maximal biomass is achieved and the quantity of recombinant protein is increased up to 18% (Extended Data Fig. 9a). The shared target case strain (PYC) showed no budget increase compared to the WT strain in absolute fluorescence (Fig. 5c), however it showed increased budget in normalized fluorescence (Fig. 5d). The PYC mutant show reduced maximal biomass production (Extended Data Fig. 6b), these results suggest that there is increased budget but also reduced growth capacity, the latter probably due to unexpected regulatory interactions resulting from the more complex case of the shared target case.

Our ME-model simulations showed that reducing the unmodeled protein fraction (UPF) from 0.36 to 0.25, increase the maximum recombinant proteome sector from 0.15 to 0.20 (Extended Data Fig. 1). These increases in recombinant protein production are also in agreement with the experimental observations in previous works[22], where for glucose minimal medium growth conditions the maximum observed heterologous fraction was 14% of the total

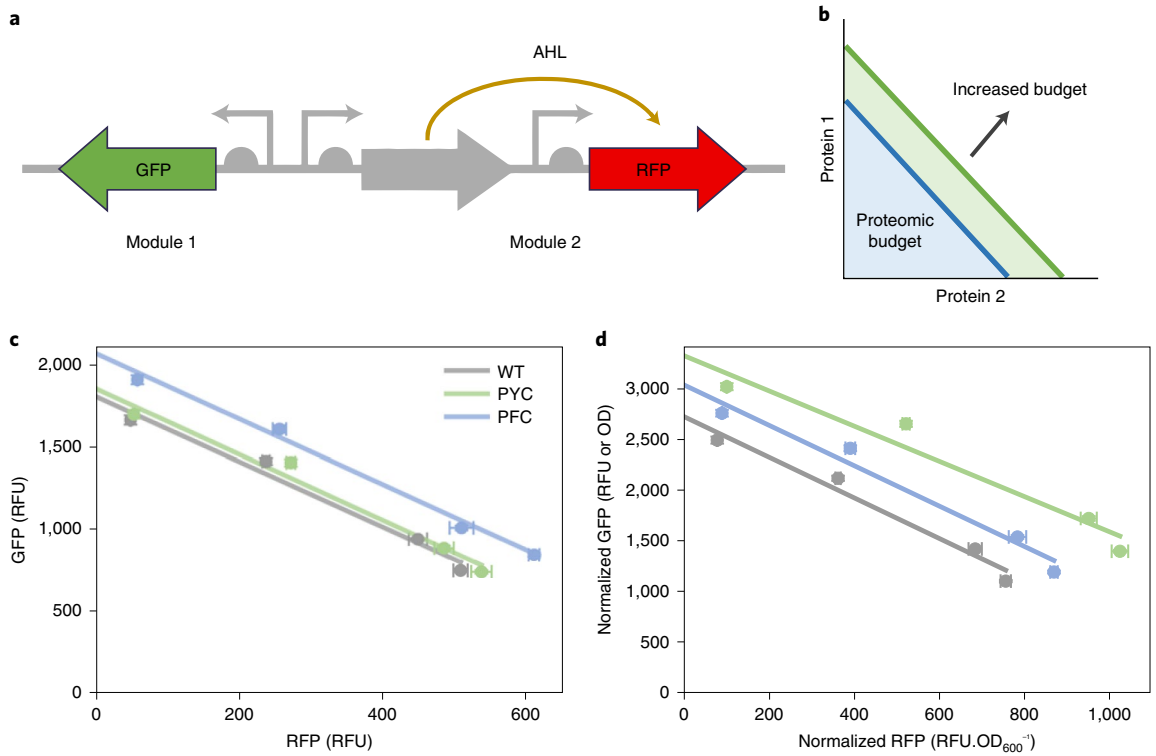


Fig. 5 | Synthetic circuit characterization. a, Schematic of the gene circuit evaluated, which encodes the fluorescent reporters GFP and RFP: the first is constitutively expressed, while the latter is under the control of a AHL-inducible promoter. **b**, When plotting the expression of one protein against the other at different levels of induction, an isocost line is obtained. The size of the area below the line represents the total proteome budget dedicated to the circuit, an upward shift in the line represents an increase in the budget. **c**,**d**, Isocost lines of the designed strains during balanced growth showing absolute fluorescence (**c**) and normalized fluorescence (**d**). In all cases, points represent the Gaussian fitted fluorescence value ±2 s.d. for $n = 9$ (see Methods) of red reporter (*x* axis) plotted against the green reporter (*y* axis) in an increasing inducer concentration (0, 2.5, 5, 20 nM AHL). A linear regression was used to fit the points to a line.

proteome, then a 0.5–1% increase in proteome availability may result in a 10% increase in the heterologous budget, which is close to what we found in our PFC strain.

We also evaluated and constructed the fourth best combination of mutants for glucose minimal medium. ReProMin predicts as PFC with the addition of *marA* as the best fourth TF to KO in glucose minimal medium, with a 0.54% of P_B. The resulting strain, PFCM (Δ*marA*, multiple antibiotic resistance), also has a higher proteomic budget than the WT, however, it is not higher than PFC at balanced growth (Extended Data Fig. 8a), confirming our hypothesis that lower number of genetic interventions will have a higher probability of success due to epistatic interactions of multiple TF deletions. Additionally, to experimentally test ReProMin predictions in a different environment we evaluated the previously described PYN mutant in rich LB medium. PYN has a higher proteome budget than the WT strain during balanced growth on LB, and also at stationary phase (Extended Data Figs. 8b and 9b) that demonstrates that our method can be used in any condition where gene essentiality and proteomics information is available.

Expression of a heterologous metabolic pathway. We tested the ability of our engineered strain for synthesizing the molecule violacein as a proof-of-concept for applications of the mutants designed by our method, with a higher proteomic budget, in metabolic engineering using a costly heterologous pathway. Violacein is a pigment from *Chromobacterium violaceum* endowed with many biological activities (antibacterial, antiviral, antiparasite) and has recently gained importance in the industrial field especially for applications in cosmetics, medicines and fabrics[23]. Violacein is synthesized in a five-step metabolic pathway using tryptophan as a precursor. Here, we used the violacein pathway plasmid reported by Darlington et al.[13], where the five genes for violacein biosynthesis are arranged in two operons, one consisting of *vioA* constitutively expressed, while the rest of the pathway encoded by the *vioBCDE* genes is under the control of an *N*-acyl-homoserine lactone- (AHL-) inducible promoter[13] (Fig. 6a). This construction follows the same principle as the previous circuit so the more of one module is produced, the less of the other is expressed due to the competition for limited resources for gene expression. However, in this case the number of genes in each module is different and code for actively metabolic enzymes with different kinetic properties, which results in differential violacein biosynthesis. This system is ideal to test the ReProMin designed strains, since their capacity to produce a metabolite from a heterologous pathway is directly dependent on the fraction of the cellular machinery that can be allocated for the expression of accessory proteome in a range of conditions (from low to high competition).

We evaluated violacein production after 24 h in the WT and PFC strains using M9 glucose medium with and without tryptophan (2.0 g l^{-1}), in both cases AHL (1.25, 2.5, 5, 10, 20 nM) was added for induction. We found that the maximum production was achieved with just a minimum amount of inducer (1.25 nM) indicating that it is crucial to have a balanced expression of the pathway with the right amount of each module to maximize the synthesis of the final product. PFC showed a mean increase in mg of violacein production of 18% ($P < 0.05$) (Fig. 6b) and 20% in normalized data (mg violacein per mg protein) (Extended Data Fig. 10a). Even though the designed regulatory interventions do not target the metabolic network of the organism, we still found a significant increase in

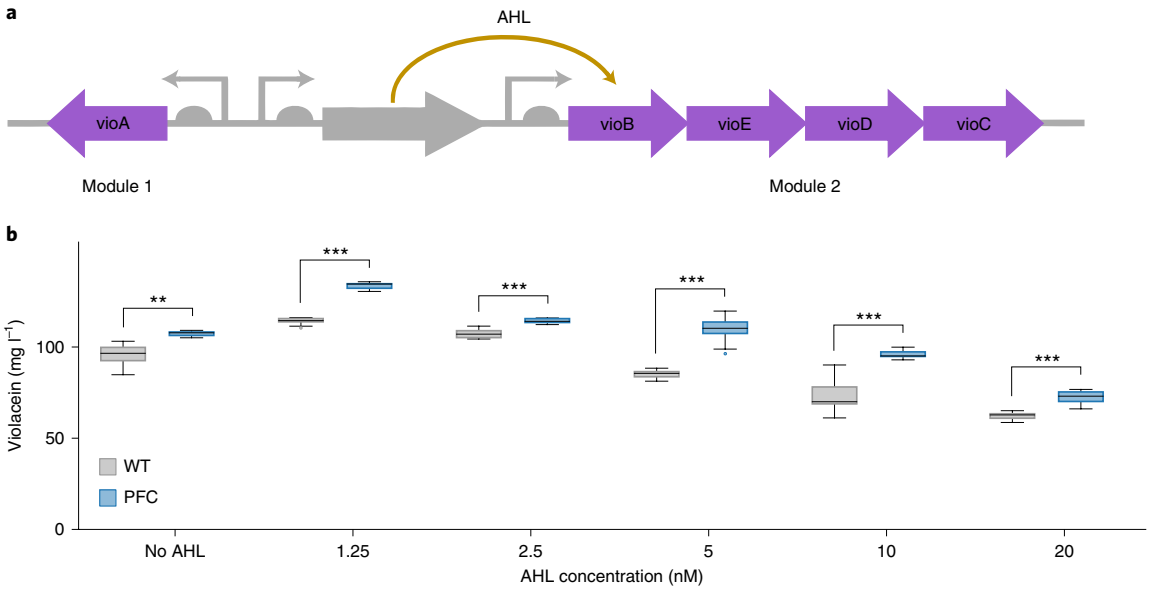


Fig. 6 | Violacein pathway evaluation. a, Schematic of the circuit expressing the pathway for violacein biosynthesis. **b**, Total violacein production using 2 g l^{-1} tryptophan after 24 h in the presence of increasing inducer (AHL) concentrations (mean ± s.d., $n = 9$). Asterisks *, ** and *** denote significant differences between WT and PFC using a two-tailed unpaired Student's *t*-test. The following *P* values were obtained for AHL concentrations: No AHL, $P = 0.0031$; 1.25 nM, $P < 0.0001$; 2.5 nM, $P < 0.0001$; 5 nM, $P < 0.0001$; 10 nM, $P < 0.0001$; 20 nM, $P < 0.0001$.

violacein production, even without tryptophan addition in a non-metabolically engineered background and without further culture condition optimization (Extended Data Fig. 10b). Previous studies have shown that proteome reallocation, such as the obtained by ALE do not require metabolic flux distribution[24]; therefore, we do not expect a relevant flux redistribution in PFC strain to be responsible for the increase in violacein production, although it would be interesting to measure it. This increase, presumably resulting from a better expression of the heterologous pathway, shows that our approach can also be harnessed to increase the production of metabolites from costly heterologous metabolic pathways.

Discussion

Gene regulatory networks are robust and can be severely rewired with interesting phenotypic outcomes[25], thus they are a perfect rational engineering target for synthetic biology applications. In this work, we have proved that the definition of an essential gene set together with regulatory network information allows the identification of TFs whose elimination leads directly to silencing proteome fractions that are not used in a particular condition. We have showed that by eliminating hedging proteome activators we can release resources and increase cellular capacity for engineered functions. The certainty of ReProMin calculations will always depend on gene essentiality data and the knowledge of gene–TF interactions. In this case, due to the gaps in our understanding of the gene–TF interactions of *E. coli*, we found that it is better to use the unique target cases where the combinations are less complex.

In agreement with the presented metabolism and gene-expression-model simulations reducing the unused protein fraction, our designed strain showed a higher proteomic budget, measured by the isocost lines, and a higher capacity to produce a metabolite from a heterologous pathway. Even though the amount of reduced proteome in the PFC strain may seem insignificant in the evaluated conditions, we have shown that a 1% reduction of unused proteome can increase the heterologous proteome fraction by 10%. The total amount of proteome available for reduction is dependent on the growth conditions and different methods of calculation may yield different numbers. For example, O'Brien et al. showed that in glucose minimal medium up to 30% of the proteome may be unused[10]. According to our classification of essential gene list (that includes several experimental datasets) and comparing metabolism and gene-expression-model proteome use predictions to actual proteomic data[16], we calculate that a 33.6% of proteome is directly available for reduction. Using our method, we were able to liberate a 1.3% in the best of the analyzed cases. Most of the genes that contribute to the dispensable proteome have no known TF or are regulated by a TF classified as essential, therefore are out of the scope of ReProMin. Furthermore, analyzing the data of a genome reduced strain of *E. coli*[26], we found that a genome reduction of 15% (743 genes) led to a proteome reduction (according to our calculations with the same proteomic dataset) of only 1.5% in the same condition (glucose minimal medium). Therefore, reducing the full 30% of dispensable proteome remains a challenge that may be tackled using a combination of approaches, such as targeting global regulators, reducing large sections of the genome or by the optimization of the core proteome[27].

By comparing our strains with an intuitive control strain, we show that inaccurate TF elimination results in detrimental effects on growth, maximal biomass and protein production (Extended Data Figs. 6 and 9). These findings indicate that the elimination of a combination of TFs is not a trivial process; it may affect essential functions and introduce phenotypic defects. Our method shows good accuracy in terms of the obtained gene expression changes measured by RNA-seq, despite our limited knowledge of the regulatory networks. In addition, the regulatory data available is condition dependent, which limits the predictive power of our method, since we need to assume that regulatory interactions are present at all times. We anticipate that developments in high-throughput technologies (such as chromatin immunoprecipitation-seq) combined with new computational approaches[28–30] will enable the fast generation of complete regulatory networks, that combined with absolute quantification of proteomes or translation rates, will enable the application of ReProMin method to even nonmodel organisms.

We presume liberated resources are redistributed mainly among the detected upregulated genes that proteome fraction is bigger in the mutant (Fig. 4c), and we expect these resources to be redirected

to the expression of engineered functions when introduced to the modified strains.

Several experimental approaches have been applied for resource allocation optimization in bacterial host engineering. Genome minimization has been mainly done by large scale genetic interventions whose outcomes are difficult to predict and do not show greater genome stability[31,32]. ALE has showed great success, especially to identify functions not related to growth[33]; however, it selects for fast growing strains that do not necessarily result in the best production phenotypes. Moreover, the underlying selection mechanisms in ALE are normally not known therefore its effects are not predictable[34]. Genome-scale models, such as the metabolism and gene-expression model, may also be used to find the proteomic cost and fitness benefit of gene expression, thus aiding in the design of proteome allocation; however, kinetic data of each protein is needed[35] and its scope focuses on growth related functions. There are only a few reports describing regulatory approaches to improve production phenotypes, such as the global transcriptional machinery engineering[36], but none of them followed a rational approach. The methodology presented in this work is a new strategy for proteome optimization with minimal genetic interventions that overcomes the serious limitations of deleting large regions of the genome; it is a flexible pipeline that can be applied to other growth and production conditions and also to different organisms where sufficient information is available. This work shows the potential of rational design of biological systems over the predominantly used trial and error approaches.

Online content

Any Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-020-0593-y.

Received: 30 August 2019; Accepted: 15 June 2020;
Published online: 13 July 2020

References

1. Hutchison, C. A. et al. Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253–aad6253 (2016).
2. Balikó, G. et al. in *Synthetic Biology: Parts, Devices and Applications* (eds Smolke, C., Lee, S. Y., Nielsen, J. & Stephanopoulos, G.) 49–80 (Wiley-VCH, 2018).
3. Ishihama, A. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. *FEMS Microbiol. Rev.* **34**, 628–645 (2010).
4. Ulrich, L. E. & Zhulin, I. B. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res.* **38**, D401–D407 (2010).
5. Kitano, H. Biological robustness. *Nat. Rev. Genet.* **5**, 826–837 (2004).
6. Isalan, M. et al. Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840–845 (2008).
7. Mitchell, A. et al. Adaptive prediction of environmental changes by microorganisms. *Nature* **460**, 220–224 (2009).
8. Tagkopoulos, I., Liu, Y.-C. & Tavazoie, S. Predictive behavior within microbial genetic networks. *Science* **320**, 1313–1317 (2008).
9. Utrilla, J. et al. Global rebalancing of cellular resources by pleiotropic point mutations illustrates a multi-scale mechanism of adaptive evolution. *Cell Syst.* **2**, 260–271 (2016).
10. O'Brien, E. J., Utrilla, J. & Palsson, B. O. Quantification and classification of E. coli proteome utilization and unused protein costs across environments. *PLoS Comput. Biol.* **12**, e1004998 (2016).
11. Nikolados, E.-M., Weiße, A. Y., Ceroni, F. & Oyarzún, D. A. Growth defects and loss-of-function in synthetic gene circuits. *ACS Synth. Biol.* **8**, 1231–1240 (2018).
12. Ceroni, F. et al. Burden-driven feedback control of gene expression. *Nat. Methods* **15**, 387–393 (2018).
13. Darlington, A. P. S., Kim, J., Jiménez, J. I. & Bates, D. G. Dynamic allocation of orthogonal ribosomes facilitates uncoupling of co-expressed genes. *Nat. Commun.* **9**, 695 (2018).
14. Lloyd, C. J. et al. COBRAme: a computational framework for genome-scale models of metabolism and gene expression. *PLoS Comput. Biol.* **14**, e1006302 (2018).
15. Santos-Zavaleta, A. et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).
16. Schmidt, A. et al. The quantitative and condition-dependent *Escherichia coli* proteome. *Nat. Biotechnol.* **34**, 104–110 (2015).
17. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002).
18. Kim, J., Darlington, A., Salvador, M., Utrilla, J. & Jiménez, J. I. Trade-offs between gene expression, growth and phenotypic diversity in microbial populations. *Curr. Opin. Biotechnol.* **62**, 29–37 (2020).
19. Turner, P. C. et al. YqhC regulates transcription of the adjacent *Escherichia coli* genes yqhD and dkgA that are involved in furfural tolerance. *J. Ind. Microbiol. Biotechnol.* **38**, 431–439 (2011).
20. Yakovleva, G. M., Kim, S. K. & Wanner, B. L. Phosphate-independent expression of the carbon-phosphorus lyase activity of *Escherichia coli*. *Appl. Microbiol. Biotechnol.* **49**, 573–578 (1998).
21. Gyorgy, A. et al. Isocost lines describe the cellular economy of genetic circuits. *Biophys. J.* **109**, 639–646 (2015).
22. Bienick, M. S., Young, K. W., Klesmith, J. R., Detwiler, E. E. & Tomek, K. J. The interrelationship between promoter strength, gene expression, and growth rate. *PLoS ONE* **9**, 109105 (2014).
23. Durán, N. et al. Advances in *Chromobacterium violaceum* and properties of violacein-Its main secondary metabolite: a review. *Biotechnol. Adv.* **34**, 1030–1045 (2016).
24. Long, C. P., Gonzalez, J. E., Feist, A. M., Palsson, B. O. & Antoniewicz, M. R. Fast growth phenotype of *E. coli* K-12 from adaptive laboratory evolution does not require intracellular flux rewiring. *Metab. Eng.* **44**, 100–107 (2017).
25. Baumstark, R. et al. The propagation of perturbations in rewired bacterial gene networks. *Nat. Commun.* **6**, 10105 (2015).
26. Posfai, G. et al. Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–1046 (2006).
27. Hidalgo, D. & Utrilla, J. in *Minimal Cells: Design, Construction, Biotechnological Applications* (eds Lara, A. & Gosset, G.) 211–230 (Springer International Publishing, 2020); https://doi.org/10.1007/978-3-030-31897-0_8
28. Fang, X. et al. Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. *Proc. Natl Acad. Sci. USA* **114**, 10286–10291 (2017).
29. Ibarra-Arellano, M. A., Campos-González, A. I., Treviño-Quintanilla, L. G., Tauch, A. & Freyre-González, J. A. Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database* **2016**, baw089 (2016).
30. Sastry, A. V. et al. The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat. Commun.* **10**, 1–14 (2019).
31. Couto, J. M., McGarrity, A., Russell, J. & Sloan, W. T. The effect of metabolic stress on genome stability of a synthetic biology chassis *Escherichia coli* K12 strain. *Microb. Cell Fact.* **17**, 8 (2018).
32. Choe, D. et al. Adaptive laboratory evolution of a genome-reduced *Escherichia coli*. *Nat. Commun.* **10**, 935 (2019).
33. Yang, L. et al. Principles of proteome allocation are revealed using proteomic data and genome-scale models. *Sci. Rep.* **6**, 36734 (2016).
34. McCloskey, D. et al. Evolution of gene knockout strains of *E. coli* reveal regulatory architectures governed by metabolism. *Nat. Commun.* **9**, 3796 (2018).
35. Nilsson, A., Nielsen, J. & Palsson, B. O. Metabolic models of protein allocation call for the kinetome. *Cell Syst.* **5**, 538–541 (2017).
36. Klein-Marcuschamer, D., Santos, C. N. S., Yu, H. & Stephanopoulos, G. Mutagenesis of the bacterial RNA polymerase alpha subunit for improvement of complex phenotypes. *Appl. Environ. Microbiol.* **75**, 2705–2711 (2009).

Methods

Metabolism and gene-expression-model simulations. All simulations were done using the model iJL1678b-ME[14]. The corresponding transcription and translation reactions for recombinant protein (GFP) production were manually added to the model using standard methods. Unused protein fraction and flux through the recombinant protein production are changeable variables in the metabolism and gene-expression model that affect predicted growth rate and proteome composition, the values of these two variables were systematically changed in the metabolism and gene-expression model to assess their effect on growth rate (UPF = 0.36, 0.30, 0.25; flux = 0, 0.001, 0.002, 0.0025, 0.0030, 0.0035, 0.0040), all other model parameters were set as default. Proteome sectors were classified according to O'Brien et al.[10].

Definition of the essential gene list. To compile the essential gene list in the glucose minimal medium condition we combined five different datasets from different sources. Three of them were experimentally generated using different methods of gene disruption: (1) random transposon mutagenesis using M9 with glucose as growth condition (Tn-seq)[37], (2) removing large fragments of the chromosome using a homologous recombination system in rich medium (LB)[38] and (3) the updated list of the mutants of the Keio collection that are lethal, the collection was generated using rich medium[39,40]. Two gene lists were generated in silico using simulations of genome-scale metabolic and expression models capable of predicting gene expression needs in a particular condition: (4) genes that are essential for growth in M9 with glucose using iOL1554-ME model[41] and (5) genes that are essential for growth in the metabolic model iJO1366 and also experimentally in M9 with glucose[42]. Within the compiled list, genes exclusively belonging to the Tn-seq and the glucose minimal medium metabolism and gene-expression-model simulation gene lists were considered conditionally essential, as these gene lists were originally generated using M9 with glucose as the growth condition, while the rest of the genes were classified as core essential for our purposes. For the cases of galactose, acetate and glycerol + casAA minimal medium conditions, we performed gene essentially analysis with the iML1515 metabolic model[43] in COBRApy[44].

For the rich LB medium case, we used the Keio mutants list.

Identification of candidate regulators and combinatorial analysis. We sorted the TF–gene interactions from RegulonDB (v.8, regulondb.ccg.unam.mx), discarding all the sigma factor–gene interactions. Next, we classified as essential all TFs that activate at least one essential gene (from the condition-specific essential gene list) and as nonessential all TFs that do not activate any essential genes. Then we analyzed the subnetwork of interactions of each nonessential TF by numerically analyzing the output level of each TF (TF_{OUT}), which is classified into positive and negative output (TF_{OUT+} and TF_{OUT-}) representing positive and negative regulated genes, respectively, and the degree of entry of each regulated gene ($GENE_{IN}$) in turn also divided into positive and negative ($GENE_{IN+}$ and $GENE_{IN-}$). We defined as candidate for proteome reduction all those TFs that activate at least one unique gene, which numerically meets the following condition:

$$TF_{OUT+} \geq 1 \wedge GENE_{n_{IN+}} = 1 \wedge GENE_{n_{IN-}} \geq 0 \quad (1)$$

where $GENE_n$ represents any gene activated by TF.

On the other hand, the proteomic dataset previously described[17] was used to calculate the P_L of each gene and P_B of each TF according to the equations in Fig. 2.

ReProMin combinatorial analysis was achieved as follows, given a list of TFs, the total number of combinations was calculated with this formula:

$$P(n, r) = \frac{n!}{r!(n-r)!} \quad (2)$$

where n represents the number of candidate TFs, and r the size of the combination ($r \leq n$)

Next, the total number of silenced and induced genes for each combination was determined following the next criteria: for every gene involved in the combination, we subtracted one from the value of $GENE_{IN+}$ for each TF that regulates the target gene positively and one to the value of $GENE_{IN-}$ for each TF that regulates the target gene negatively. At the end of this process a gene was considered silenced if:

$$GENE_{IN} = GENE_{IN+} = 0 \quad (3)$$

or induced if:

$$GENE_{IN} = GENE_{IN-} = 0 \quad (4)$$

Finally, the P_B of each combination tested was calculated and ranked.

The full computational set of tools coded in Python and datasets used in the analysis are available in the following repository (https://github.com/utrillalab/repromin). Cytoscape software v.3.7 (ref. [45]) was used to plot the network representation of the data.

Generation of combinatorial knockout strains. The combinatorial mutants were generated by sequential P1 phage transduction from the individual knockout strains of the Keio collection according to the protocol described by Miller[46]. The removal of the kanamycin resistance cassette before each transduction was done using the pE-FLP plasmid (Addgene plasmid no. 45978), pE-FLP was a gift from D. Endy and K. Shearwinand[47]. Each knockout strain was confirmed by PCR using primers flanking each gene. In all experiments *E. coli* BW25113 was used as the WT background. The characteristics of the strains, plasmids and primers used in this study are described in the supplementary material (Supplementary Table 5).

RNA sample extraction and sequencing. Strains were grown in 50 ml of M9 medium with glucose (4 g l^{-1}) M9 medium in 250 ml Erlenmeyer flasks cultures in an orbital incubator at 37 °C (250 r.p.m.). Cells were collected in mid-log phase using the Qiagen's RNAprotect bacteria reagent according to the manufacturer's specifications. Cell pellets were incubated with lysozyme, SuperaseIn and protease K for 10 min at 37 °C. Total RNA was isolated and purified using Zymo Research's Quick-RNA kit according to the manufacturer's specifications. All samples' quality was inspected in a bioanalyzer RNA chip (Agilent). Starting with 10 µg of total RNA of each sample, the removal of ribosomal RNA was done with the Ribominus kit by Invitrogen. For the construction of the libraries, the TruSeq Stranded mRNA HT Sample Prep Kit by Illumina was used. For sequencing a NextSeq 500 v.2 was used, with a configuration of 2 × 75 paired-end read and 10 million reads per sample.

Reads were mapped to reference genome *E. coli* MG1655 (RefSeq, NC_000913.3) using aligner Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2). Final differential analysis was made using the Cufflinks library (http://cole-trapnell-lab.github.io/cufflinks). Genes with a $\log_2$ fold change that was ≥1 were considered upregulated and ≤−1 were considered downregulated, considering $P \leq 0.01$ and $n = 2$.

Estimation of the theoretical proteome from RNA-seq data. To do an estimation of the translation rates from proteomics data and from transcript abundances, we assumed that the transcripts in our RNA-seq data yields the proteome reported by Schmidt et al. in glucose, so we calculated each gene translation efficiency rate (si) using this equation:

$$si = \frac{C_{cell}i}{ri}$$

where ri represents the raw fragments per kilobase of transcript per million mapped reads value of each gene and $C_{cell}i$ the number of protein copies reported. For all the genes with no proteomic data we assumed a fixed rate corresponding to the mean rate of the reported genes. Then, we used these rates to estimate the number of protein copies per cell (Pi) for our mutant PFC according to:

$$Pi = ri \times si$$

Finally, the gene load was calculated as previously described.

Growth phenotype characterization. For the evaluation of growth in different carbon sources, the following conditions were used: glucose M9 medium (4 g l^{-1}), galactose M9 medium (3.2 g l^{-1}), acetate M9 medium (2.5 g l^{-1}) and LB rich medium. Cells were cultured overnight in the corresponding medium. The next day the strains were diluted to an optical density (OD_{600}) of 0.05 in fresh medium and 150 µl of the fresh culture was transferred to a transparent 96-well plate (Corning) and incubated at 37 °C with fast linear shaking in a micro-plate reader (Synergy 2.0, BioTek) for 24 h, taking measurements for OD_{600} every 20 min. The characterization of the growth kinetics was conducted using the algorithm Fitderiv (v.1.0) developed by Swain et al.[48] with the default parameters. The algorithm performs a Gaussian fit of the raw data and in all cases the resulted fitted value ±2 s.d. (ensuring a confidence of at least 95% ($P \leq 0.05$)) was used when comparing the mutant strains against the WT.

Isocost circuit evaluation. Strains were inoculated into glucose M9 medium with gentamicin (20 µg ml^{-1}), and grown overnight. Next day, strains were diluted to an OD_{600} of 0.05 in fresh glucose M9 medium containing AHL (Sigma-Aldrich, final concentrations of 1.25, 2.5, 5, 10, 20 nM), then 150 µl of the fresh culture was transferred to a 96-well black plate with transparent bottom (Corning) and incubated as described previously, taking measurements for OD_{600}, GFP (excitation, 485 nm and emission, 528 nm) and RFP (excitation, 590 nm and emission, 645 nm). The characterization of the production kinetics of GFP and RFP was also done using the algorithm described above.

Flow cytometry measurements. For flow cytometry measurements, cell cultures were prepared as described before, but later grown in 24-well plates using 1 ml of medium. Every hour, 50 µl aliquots were taken from each well and mixed with 150 µl of PBS, the volume of the wells was kept constant by adding fresh medium. Cell suspension was loaded into an Attune NxT Flow Cytometer (ThermoFisher) and analyzed for GFP (excitation,. 488 nm and emission, 525/50 nm) and RFP (excitation, 561 nm and emission, 620/15 nm). For each sample 20,000 events were analyzed and population means were estimated using the default software of the instrument. The characterization of the production kinetics of GFP and RFP was also done using the algorithm described above.

Characterization of violacein-producing strains. The strains were inoculated into glucose M9 medium with gentamicin (20 μg ml^{-1}), and grown overnight. Next day, strains were diluted to an OD_{600} of 0.05 in fresh glucose M9 medium containing AHL (1.25, 2.5, 5, 10, 20 nM) and tryptophan (2.0 g l^{-1}), then 150 μl of the fresh culture was transferred to a 96-well plate and incubated as described before. After 24 h the plate was centrifuged (13,000*g*, 10 min), and the supernatant of each well was discarded. Violacein was extracted by suspending the pellet in each well in 150 μl absolute ethanol and incubating the plate at 95 °C for 10 min followed by pelleting cell debris (13,000*g*, 10 min). Violacein present in the extract was determined spectrophotometrically at 575 nm in a micro-plate reader (Synergy 2.0, BioTek) and quantification was made using a curve with a purchased violacein standard (Sigma-Aldrich).

Quantification of total protein. The Biuret method was used for the quantification of total protein: 1 ml of culture was taken and centrifuged (13,000*g*, 10 min). The supernatant was collected and the pellet was washed with 0.2 ml of water, resuspended and centrifuged again, the water was discarded. Pellets were resuspended in 0.1 ml of 6 M NaOH and incubated at 95 °C for 10 min to break the cells and hydrolyze the proteins. To perform the Biuret reaction, 0.1 ml of 3.2% $CuSO_4$ was added to the samples and incubated under vigorous agitation for 5 min. Next, samples were centrifuged for 2 min and 150 μl of supernatant was placed in a 96-well plate. The absorbance at 555 nm was measured in a plate reader (Synergy 2.0, BioTek).

Statistics. Samples sizes are defined in each figure legend. In all trials, three replicates were included and the experiment was repeated independently on three different days. For growth kinetics and fluorescence, the value of the replicates is presented as the Gaussian fitted value ±2 s.d. For violacein, results are presented as mean ± s.d. and statistical significance between conditions was calculated using Student's *t*-test (two-tailed). All statistical calculations and numerical analyses were performed using Python 3 packages.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

RNA-seq data from this study have been deposited in NCBI's Gene Expression Omnibus (GSE134335). Additional data is available from the corresponding author upon reasonable request.

Code availability

The code and data to run ReProMin can be found at: https://github.com/utrillalab/repromin

References

37. Yang, L. et al. Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proc. Natl Acad. Sci. USA* **112**, 10810–10815 (2015).
38. Kato, J.-I. & Hashimoto, M. Construction of consecutive deletions of the *Escherichia coli* chromosome. *Mol. Syst. Biol.* **3**, 1–7 (2007).
39. 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).
40. Yamamoto, N. et al. Update on the Keio collection of *Escherichia coli* single-gene deletion mutants. *Mol. Syst. Biol.* **5**, 335 (2009).
41. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693 (2013).
42. Orth, J. D. & Palsson, B. Gap-filling analysis of the iJO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions. *BMC Syst. Biol.* **6**, 30 (2012).
43. Monk, J. M. et al. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* **35**, 904–908 (2017).
44. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: constraints-based reconstruction and analysis for python. *BMC Syst. Biol.* **7**, 74 (2013).
45. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
46. Miller, J. H. *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for* Escherichia coli *and Related Bacteria* (Cold Spring Harbor Laboratory Press, 1992).
47. St-Pierre, F. et al. One-step cloning and chromosomal integration of DNA. *ACS Synth. Biol.* **2**, 537–541 (2013).
48. Swain, P. S. et al. Inferring time derivatives including cell growth rates using Gaussian processes. *Nat. Commun.* **7**, 13766 (2016).

Acknowledgements

We thank E. Marquez-Zavala and C. Lloyd for metabolism and gene-expression model simulations support. Y. Castillo-Franco and C. F. Mendez-Cruz for computational support and G. Hernandez-Chavez, H. King, M. Hughes and A. Sicilia for technical support. We acknowledge the funding provided by UNAM–DGAPA-PAPIIT projects IA200716 and IA201518. Newton advanced Fellowship Project NA 160328. J.J. and J.K. acknowledge the support received from the Biology and Biotechnology Research Council (grant nos. BB/M009769/1 and BB/T011289/1) and European Union's Horizon 2020 research and innovation program for the project P4SB (grant agreement no. 633962). G.L.P. acknowledges the Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), and the PhD scholarship 434655 from CONACyT.

Author contributions

J.U. and G.L.P. designed ReProMin. G.L.P. developed computational methods and performed data analysis. G.L.P. and J.S.M.H. carried out experiments. G.L.P., J.K. and J.I.J. analyzed flow cytometry experiments, isocost lines and violacein production. J.U. supervised the study. J.U., G.L.P. and J.I.J. wrote the manuscript.

Competing interests

J.U. and G.L.P. are inventors in a MX patent application filled by UNAM.

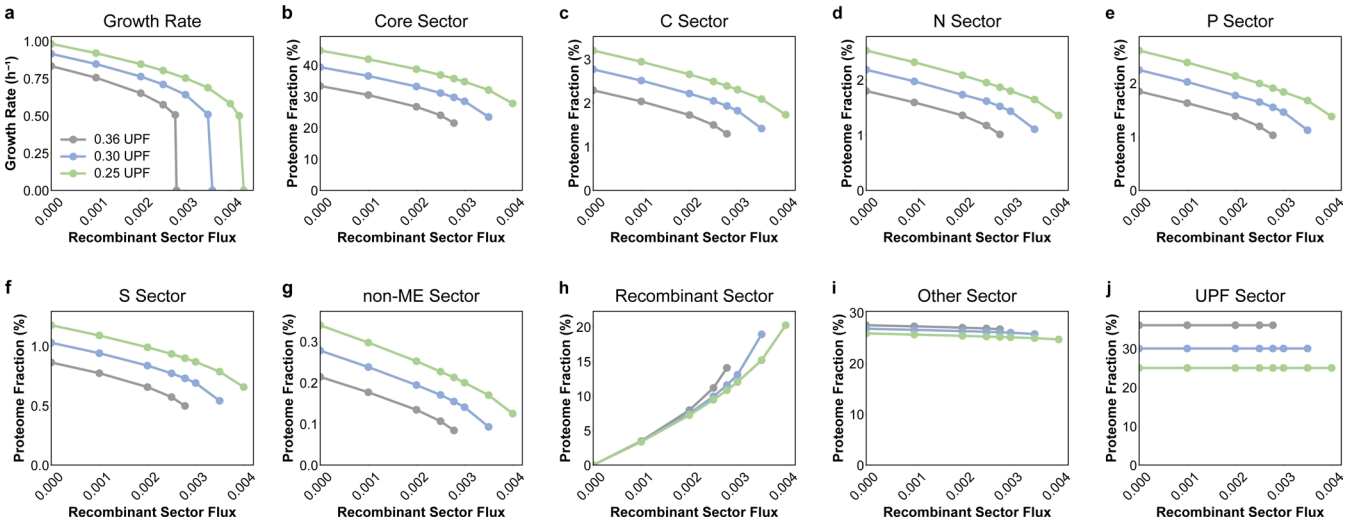
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41589-020-0593-y.

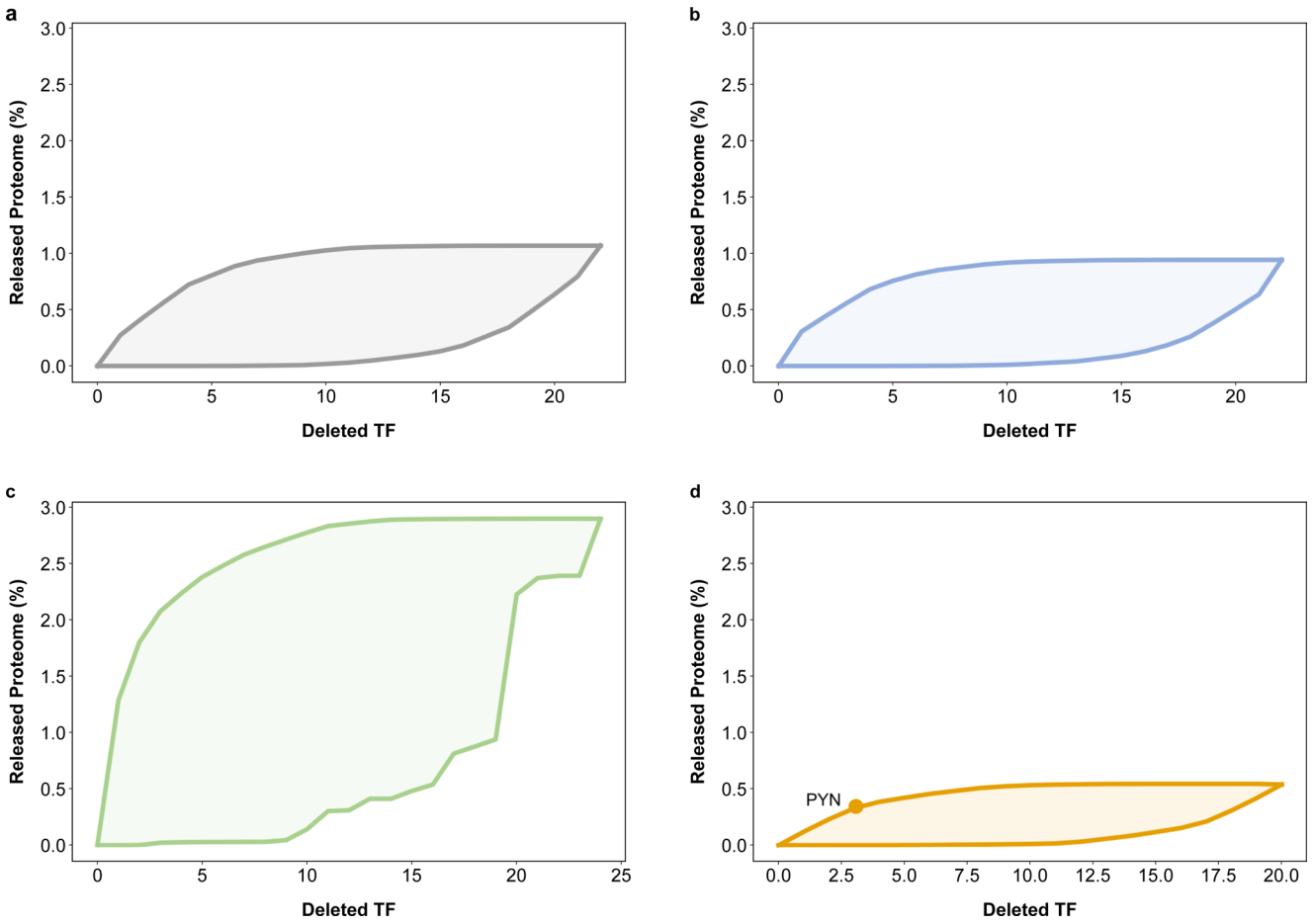
Supplementary information is available for this paper at https://doi.org/10.1038/s41589-020-0593-y.

Correspondence and requests for materials should be addressed to J.U.

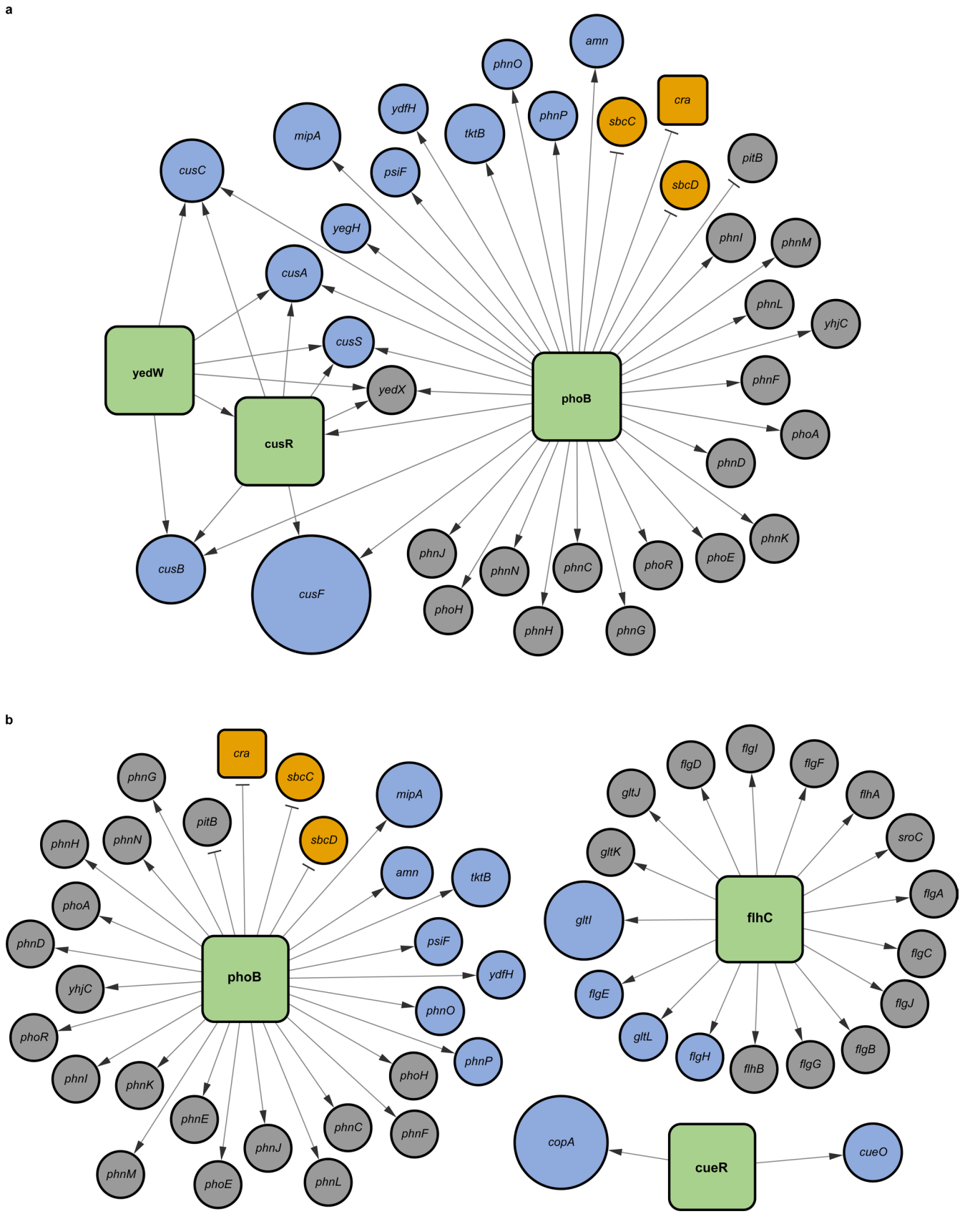
Reprints and permissions information is available at www.nature.com/reprints.



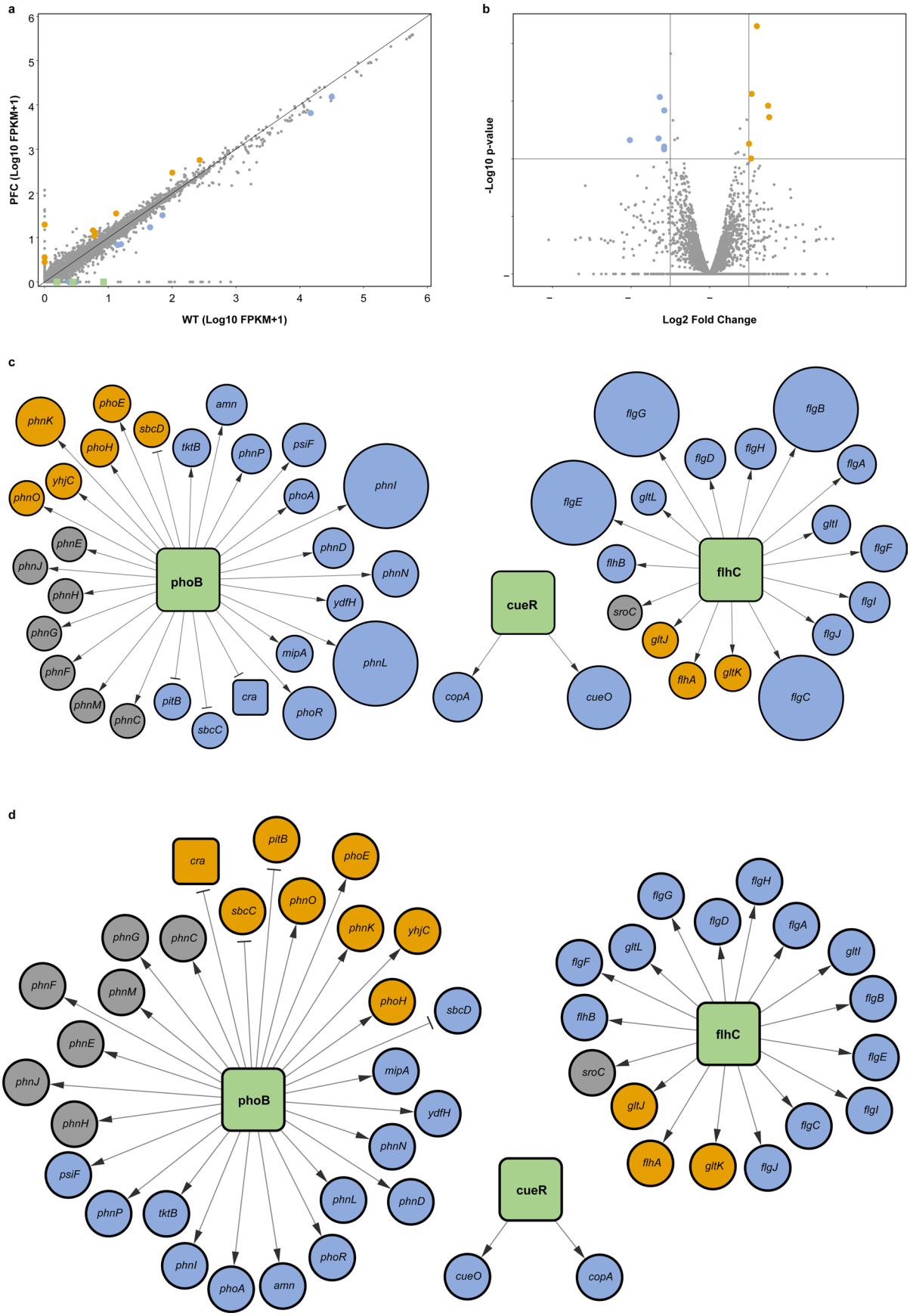
Extended Data Fig. 1 | ME-model simulations and proteome sector response to reducing the unmodeled protein fraction (UPF). The ME-model iJL1678b-ME was used to simulate the effect of the reduction of the UPF and different expression levels of an unused recombinant model protein (GFP) (see methods). Similar to the maintenance energy coefficient, the hedging proteome and other non-growth related (thus not modeled) functions are accounted for in ME-models as a part of the UPF. Each panel shows **a**, growth rate and the corresponding fraction of each proteome sector **b**, core sector and the alternative element dependent sector: **c**, the carbon sector **d**, the nitrogen sector **e**, the phosphate sector **f**, the sulphur sector **g**, the non-ME sector **h**, the recombinant sector, comprised by the maximum attained GFP expression, **i**, the other sector (non-classified) and **j**, the UPF sector. The simulation shows an increased availability of cellular resources for recombinant protein production by reducing the UPF.



Extended Data Fig. 2 | ReProMin proteome liberation landscapes corresponding to the UT case. Potential proteome liberation landscape corresponding to **a**, Galactose, **b**, Acetate, **c**, Glycerol + casAA and **d**, Rich Medium (LB).

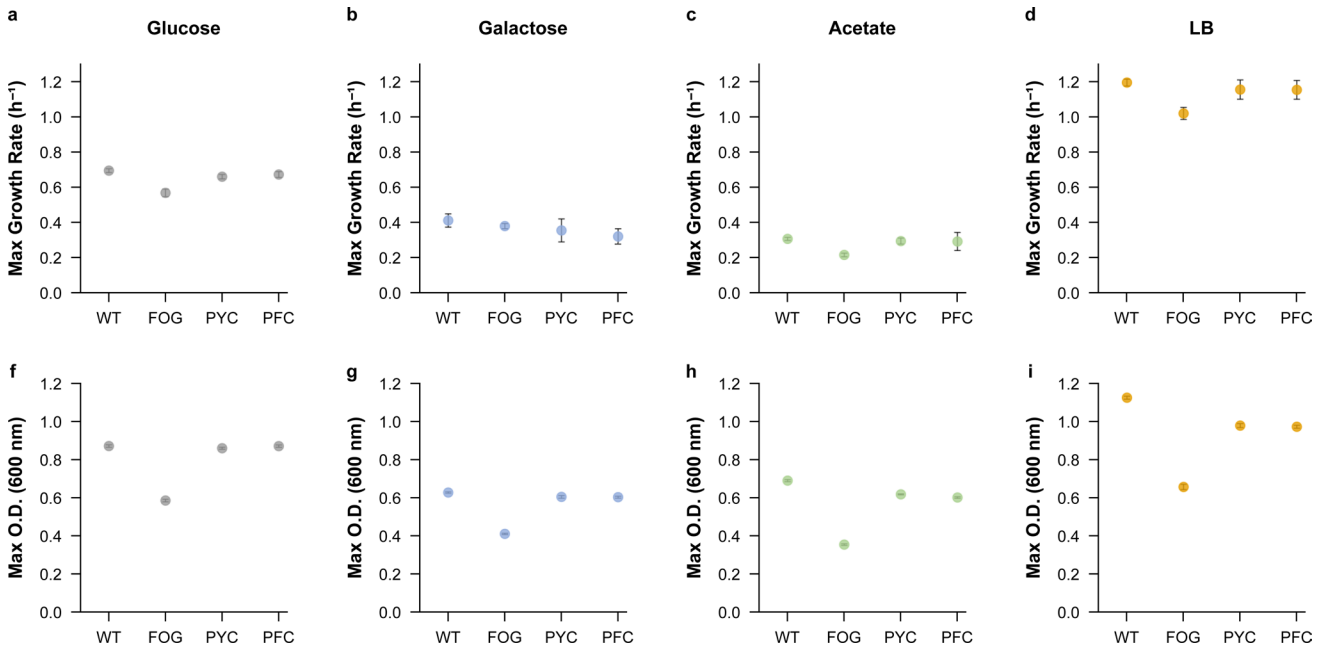


Extended Data Fig. 3 | Regulatory subnetwork of ReProMin predicted gene targets. Subnetwork corresponding to **a**, ST case PYC mutant and **b**, UT case PFC mutant; blue circles represent predicted silenced targets, yellow circles predicted induced targets and gray circles genes with no proteomic coverage; size of the circles is proportional to the P_L of the target.

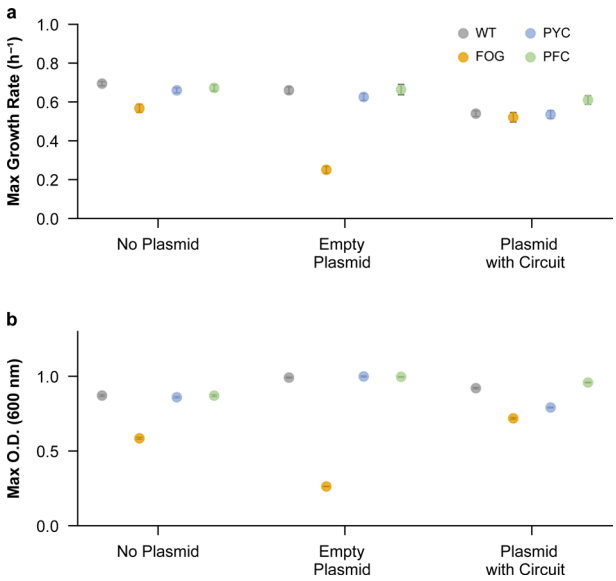


Extended Data Fig. 4 | See next page for caption.

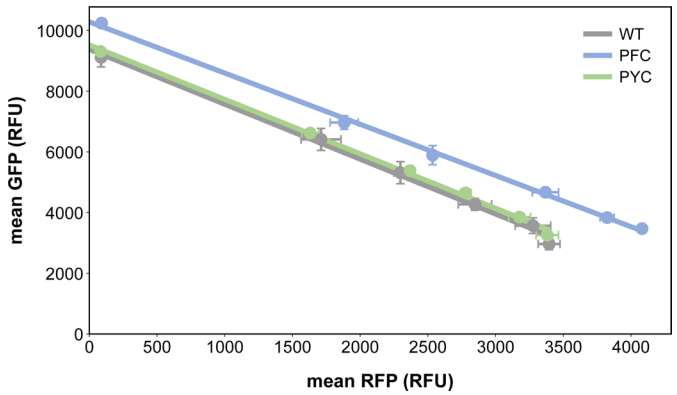
Extended Data Fig. 4 | Transcriptomic analysis of the UT case designed strain. a, Correlation plot for PFC and WT strains transcripts. Green squares represent the three deleted TFs. **b**, Volcano plot showing differential gene expression. In both cases, statistically significant genes are highlighted (blue – downregulated, yellow—upregulated) (log2 Fold Change ≥ 1 or ≤ -1, $P \leq 0.01$, $n = 2$). **c**, Integration of transcriptomics with computational tool predictions. The size of the circle corresponds to the fold change of each target (the largest circles represent fully silenced genes), in all cases blue circles represent targets releasing resources (down regulated), yellow circles represent targets generating burden (upregulated) and grey circles targets that were not found expressed. **d**, Accuracy of computational tool predictions based on RNAseq data. Yellow circles represent wrong predictions, blue circles represent accurate predictions and grey circles represent unmapped predictions (expression was not detected).



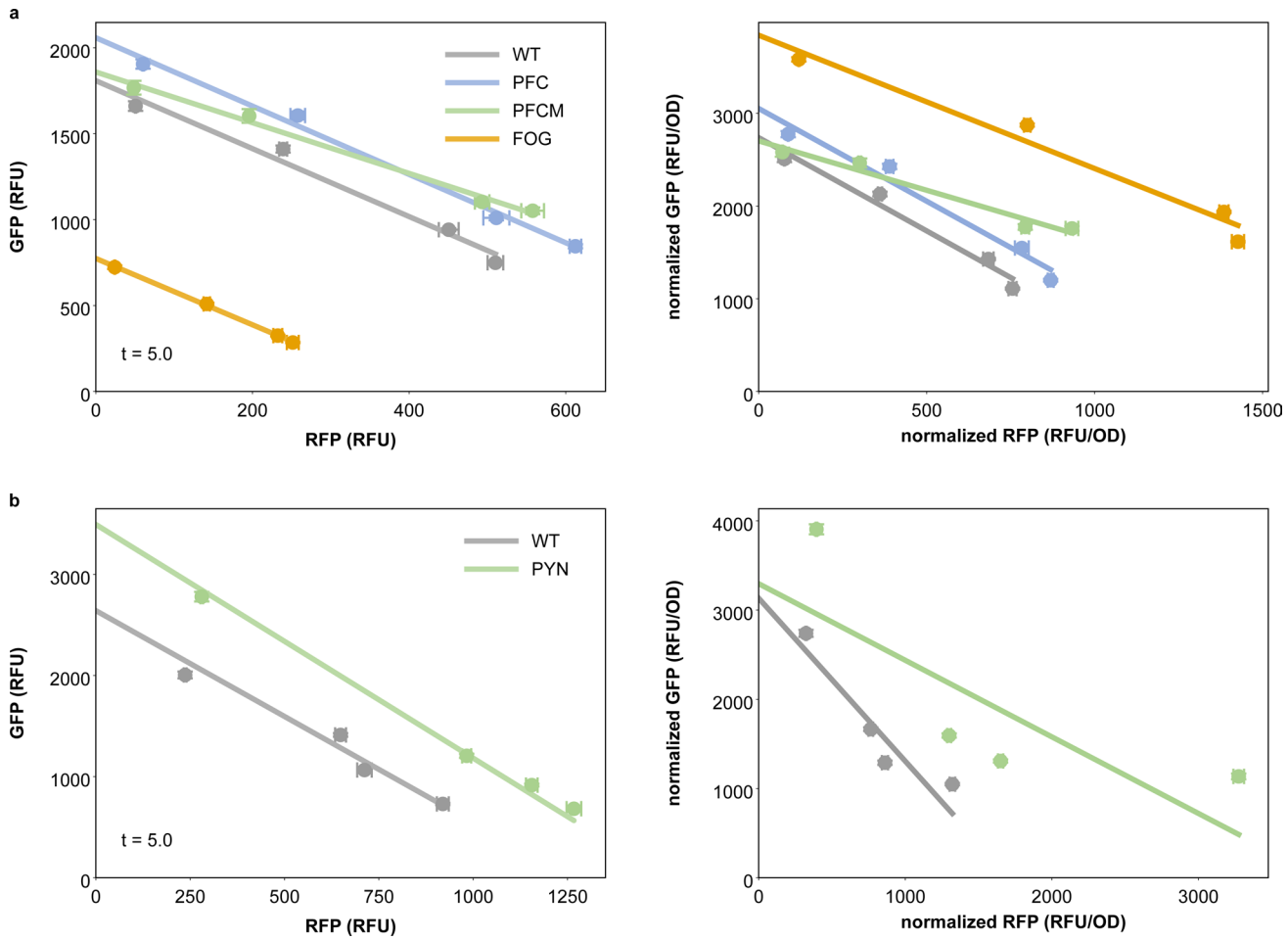
Extended Data Fig. 5 | Phenotypic evaluation generated strains based on glucose ReProMin predictions (UT and ST cases) and control. Growth on different carbon source supplemented M9 medium and rich medium (LB). **a-d**, shows max growth rate and **f-i**, shows max O.D. Points represent the Gaussian fitted value ± 2 s.d. for n = 9.



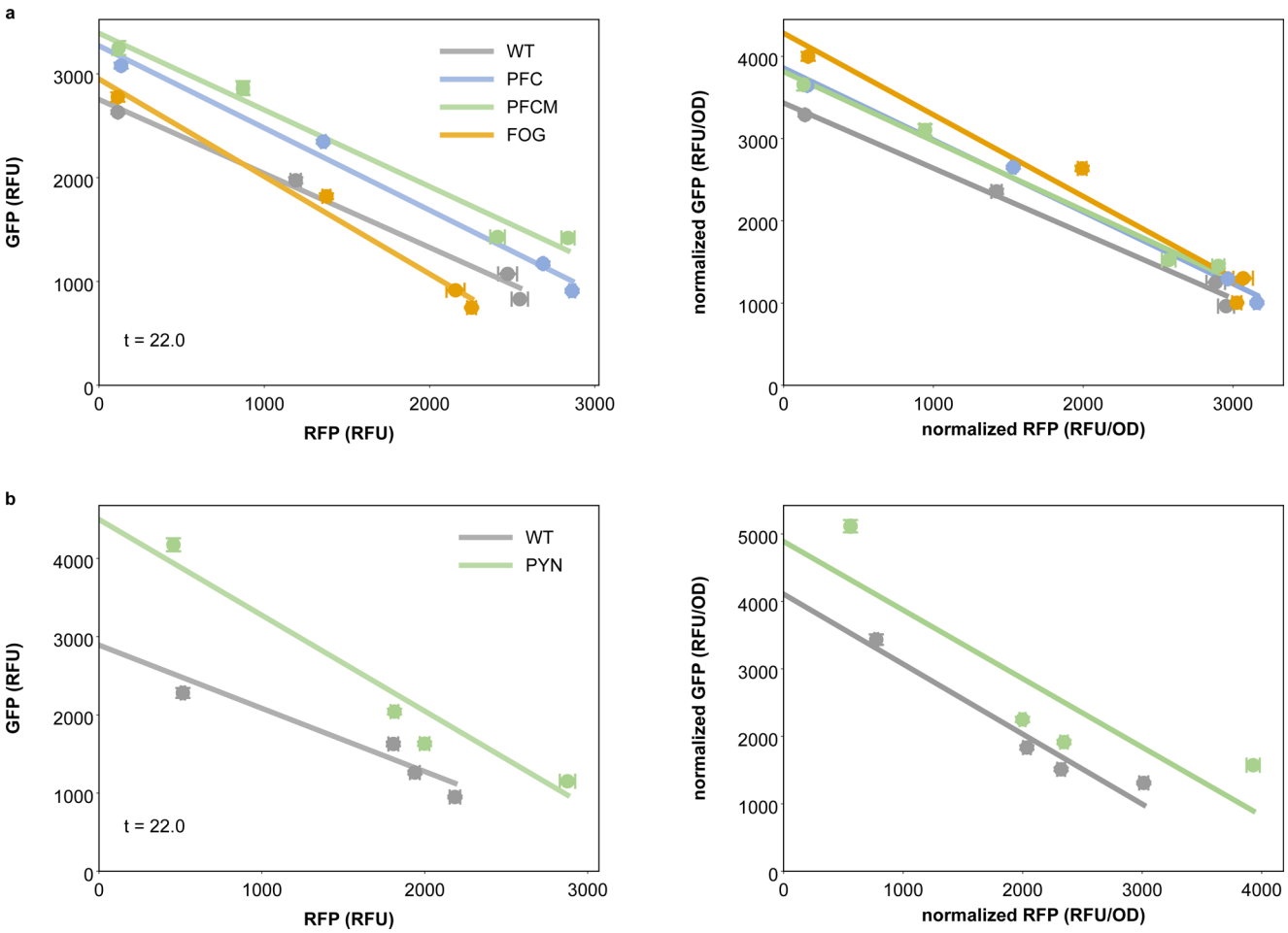
Extended Data Fig. 6 | Metabolic burden evaluation of strains based on glucose ReProMin predictions (UT and ST cases) and control. Metabolic burden while carrying empty, circuit plasmid and induced circuit plasmid, **a**, shows max growth rate and **b**, shows max O.D. Points represent the Gaussian fitted value ± 2 s.d. for n = 9.



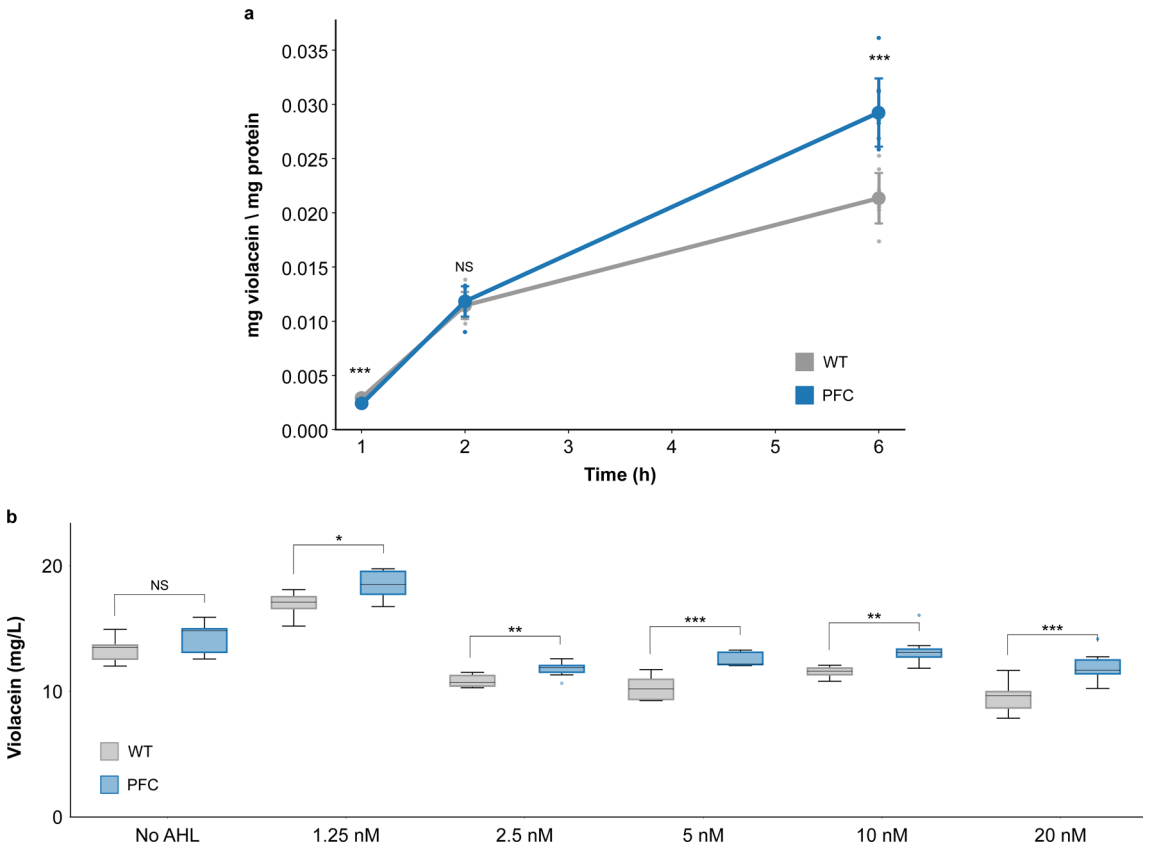
Extended Data Fig. 7 | Synthetic circuit characterization in glucose M9 medium. Isocost lines showing mean fluorescence per cell measured by flow cytometry during balanced growth (~5 h). Points represent the Gaussian fitted fluorescence value ± 2 s.d. for n = 9 of red reporter (x axis) plotted against the green reporter (y axis) in an increasing inducer concentration (0, 2.5, 5, 20 nM AHL). A linear regression was used to fit the points to a line.



Extended Data Fig. 8 | Isocost lines during balanced growth (~5 h). Isocost lines of the generated mutant strains for two growth conditions: **a**, Glucose M9 medium and **b**, Rich medium. Left: absolute fluorescence, Right: normalized fluorescence. Points represent the Gaussian fitted fluorescence value ± 2 s.d. for n = 9 of red reporter (x axis) plotted against the green reporter (y axis) in an increasing inducer concentration (0, 2.5, 5, 20 nM AHL). A linear regression was used to fit the points to a line.



Extended Data Fig. 9 | Isocost lines during stationary phase (~22 hrs). Isocost lines of the generated mutant strains for two growth conditions: **a**, Glucose M9 medium and **b**, Rich medium. Left: absolute fluorescence, Right: normalized fluorescence Points represent the Gaussian fitted fluorescence value ± 2 s.d. for n = 9 of red reporter (x axis) plotted against the green reporter (y axis) in an increasing inducer concentration (0, 2.5, 5, 20 nM AHL). A linear regression was used to fit the points to a line.



Extended Data Fig. 10 | Violacein production characterization. a, Protein normalized violacein production using 2 g/L tryptophan in the presence of AHL (20 nM) (mean ± s.d., n = 9). **b**, Total violacein production without adding tryptophan after 24 h in the presence of increasing inducer (AHL) concentrations (mean ± s.d., n = 9). Asterisks *, ** and *** denote significant differences between WT and PFC using a two-tailed unpaired Student's t-test. The following P values were obtained for normalized violacein production: 1 h, P = 0.0003; 2 h, P = 0.5647; 6 h, P < 0.0001. The following P values were obtained for violacein production with different AHL concentrations: No AHL, P = 0.0599; 1.25 nM, P = 0.0146; 2.5 nM, P = 0.0021; 5 nM, P < 0.0001; 10 nM, P = 0.0014; 20 nM, P = 0.0005.

nature research

Corresponding author(s): Jose Utrilla

Last updated by author(s): May 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Growth, fluorescence, total protein and total violacein data was collected using Biotek Gen5 Software. Flow cytometry data was collected using Attune NxT Software (ver. 4.2). RNA-seq data was collected using NextSeq 500 v2.
Data analysis	ME-Model simulations were performed using model iJL1678b and Python packages Cobrapy (ver. 0.18) and Ecolime (ver. 0.4). Gene essentiality, regulatory network analysis and ReProMin combinatorial analysis were done using custom Python 3 code. Regulatory networks representations were done using Cytoscape software (ver. 3.7). The analysis of growth and fluorescence kinetics was conducted using the algorithm Fitderiv (ver. 1.0) described in ref. 48. All graphs and associated statistical calculations and numerical analyses were performed using the latest Python 3 packages (matplotlib, pandas, numpy and scipy).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-seq data from this study have been deposited in NCBI's Gene Expression Omnibus (GSE 134335).
The code to run ReProMin can be found at: https://github.com/utrillalab/repromin

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to determine sample size for experimentation given the minimal experimental variation in assays based on our previous experience.
Data exclusions	No data was excluded from the analysis.
Replication	All experiments were carried out at least 3 times to verify its reproducibility. All the attempts of replication were succesful.
Randomization	We are always comparing mutant strains vs WT strains so no randomization was performed.
Blinding	No blinding was performed. Blinding was unnecessary as all data collection and analysis is quantitative and not qualitative in nature.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Grown in 24-well plates using 1 ml of medium. Every hour 50µL aliquots were taken from each well and mixed with 150µL of PBS, the volume of the wells was kept constant by adding fresh medium.
Instrument	Attune NxT Flow Cytometer (ThermoFisher, Waltham, MA, USA).
Software	Attune NxT Software (ver. 4.2).
Cell population abundance	For each sample 20,000 events were analysed and population means were estimated.
Gating strategy	Bacterial cells showing green (BL1, excitation 561 nm; emission 620/15 nm) and red fluorescence signals (YL2, excitation 561 nm; emission 620/15 nm) were determined with the following settings: FSC 460 V; SSC 360 V; BL1 260 V; YL2 280 V. All events detected were recorded and we applied an additional gate, identical for all samples, excluding the events in the upper

5% of both FSC and SSC, which were likely cell aggregates. The remaining 95% events (20,000 in total in this fraction) were analysed

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.