Genome-wide phenotypic analysis of growth, cell morphogenesis and cell cycle events in *Escherichia coli*.

Manuel Campos[1,2,3,¶], Genevieve Dobihal[1,3,§], Christine Jacobs-Wagner[1,2,3,4,*]

1 Microbial Sciences Institute, Yale University.
2 Department of Molecular, Cellular and Developmental Biology, Yale University.
3 Howard Hughes Medical Institute, Yale University.
4 Department of Microbial Pathogenesis, Yale Medical School.

*** Contact information: christine.jacobs-wagner@yale.edu, tel: +1-203-737-7219, fax: +1-203-737-6715.**

¶ Present address: Laboratoire de Microbiologie et de Génétique Moléculaires (LMGM), Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, Université Paul Sabatier, F-31062 Toulouse, France

§ Present address: Department of Microbiology and Immunology, Harvard Medical School, Boston, MA 02115, USA.

Abstract

Cell size, cell growth and the cell cycle are necessarily intertwined to achieve robust bacterial replication. However, a comprehensive and integrated view of these fundamental processes is lacking. Here, we describe an image-based quantitative screen over the single-gene knockout collection of *Escherichia coli*, which led to the identification of many new genes involved in cell morphogenesis, population growth, nucleoid (bulk chromosome) dynamics and cell division. Functional analyses, together with high-dimensional classification, unveil new associations of morphological and cell cycle phenotypes with specific functions and pathways. Additionally, correlation analysis across ~4,000 genetic perturbations demonstrate that growth rate is not a determinant of cell size. Cell width and length are also uncorrelated, suggesting that cells do not control their size by monitoring surface area or volume; instead cells appear to regulate width and length independently. Furthermore, our analysis identifies scaling relationships between cell size and nucleoid size and between nucleoid size and the relative timings of nucleoid separation and cell division, linking cell morphogenesis to the cell cycle via the global architecture of the chromosome.

Introduction

Cells must integrate a large variety of processes to achieve robust multiplication. Bacteria, in particular, are remarkable at proliferating, which has been key to their ecological success. During their fast-paced replication, bacterial cells must perform a multitude of tasks. They have to uptake and process nutrients, generate energy, build cellular components, duplicate and segregate their genetic material, couple growth and division, maintain their shape and size, while sensing their environment, repairing cellular damages and performing other important functions. These tasks must be integrated to ensure successful cellular replication. Decades of work have garnered extensive knowledge on specific processes, genes and pathways. However, we still lack a comprehensive view of the genetic determinants affecting cell morphogenesis and the cell cycle. It is also unclear how cellular activities are integrated to ensure that each division produces two viable daughter cells.

Systematic genome-wide screens, rendered possible by the creation of arrayed single-gene knock-out collections, have been successfully used to gain a more comprehensive perspective on cell morphogenesis and the cell cycle in yeast [23, 30, 45]. Here, we present a high-content, quantitative study that uses the Keio collection of *Escherichia coli* gene deletion strains [4] and combines microscopy with advanced statistical and image analysis procedures to examine the impact of each non-essential *E. coli* gene on cell morphology, growth, nucleoid (bulk chromosome) dynamics and cell constriction.

Results

High-throughput imaging and growth measurements of the *E. coli* Keio collection

To gain an understanding of the molecular relationship between growth, cell size, cell shape and specific cell cycle events, we imaged 4,227 strains of the Keio collection. This set represents 98% of the non-essential genome (87% of the complete genome) of *E. coli* K12. The strains were grown in 96-well plates in M9 medium supplemented with 0.1% casamino acids and 0.2% glucose at 30°C. Cells were stained with the DNA dye DAPI, and spotted on large custom-made agarose pads (48 strains per pad) prior to imaging by phase contrast and epifluorescence microscopy (Fig 1A). On average, about 360 cells were imaged for each strain. To provide a reference, 240 replicates of the parental strain (BW25113, here referred to as WT) were also grown and imaged under the same conditions as the mutants. In parallel, using a microplate reader, we recorded the growth curves of all the strains (Fig 1A), to which we fitted the Gompertz function to estimate two population-growth features: the maximal growth rate (α_{max}) and the saturating density (OD_{max}) of each culture (Appendix Fig S1A).

High-throughput dataset curation using support vector machine

Cells and their contours were detected in an automated fashion using the open-source software Oufti [46]. The large size ($> 1,500,000$ cells detected) of the dataset precluded the validation of each cell contour by visual inspection. Therefore, we implemented an automated classification method based on support vector machine (SVM) [20] to identify and discard incorrectly detected cells (Fig 1B). To generate a training dataset for the SVM model, we visually scored (positive or negative) 43,774 cell contours from the parental strain and the 419 mutants displaying the greatest deviations in cellular dimensions before data curation. This inclusion of the most aberrant mutants in the training dataset allowed us to build a versatile model that performed well on the wide range of cell sizes and shapes present in the Keio collection. The quality of the fit of the SVM model to the training dataset was evaluated by a 10-fold cross-validation [25], which gave a misclassification error rate under 10%. The model was further validated on an independent dataset of 102,137 visually scored cell contours taken from the same group of WT and mutant strains. We found that our SVM model performed very well on this validation set, as shown by the high AUROC (area under the 'receiver operating characteristic' curve) value of 0.94 (Appendix Fig S1B). By comparing the model classification with visual scoring (Fig 1C), we found that only about 3% of cell contours in the validation set were incorrectly

identified as positive (false positives) by the SVM model. Importantly, these misclassified cells introduced no biases in the measurement of morphological features (Appendix Fig S1C), even when considering the 419 most aberrant strains (Appendix Fig S1D). This validated SVM model was used to curate the entire dataset, retaining about 1,300,000 identified cells (∼300 cells/strain). In addition, we verified the reproducibility of our experimental approach by separately imaging two independent replicates of 192 strains that include 2 copies of the parental (WT) strain and 190 mutants with severe morphological defects. Even for cell width, the smallest dimension measured, we observed a Pearson correlation (ρ) of 0.92 (Fig 1D), indicating high reproducibility.

Quantification of cell morphological features across the genome

We obtained a wealth of quantitative information from image analysis of these strains using Oufti [46]. From phase-contrast images, we measured cellular dimensions (length, width, perimeter, cross-sectional area, aspect ratio and circularity) and their variability by calculating the coefficient of variation (CV, mean divided by the standard deviation). We also extracted the mean and CV of other morphological parameters (surface area, volume and surface-to-volume ratio). For constricted cells, we determined the relative position of division along the cell length (division ratio). Note that since the identity of the cell poles (old versus new) was unknown, randomization of cell pole identity automatically produced a mean division ratio of 0.5, even for an off-center division. Therefore, measurements of mean division ratio were meaningless and not included in our analysis. However, the CV of the division ratio was included since a high CV indicates either an asymmetric division or an imprecise division site selection. In total, each strain was characterized by 19 morphological features. The name and abbreviation for all the features can be found in Table S1.

After taking into consideration experimental variability (see Materials and methods, Appendix Fig S2 qnd S3), we calculated a normalized score (s) for each feature and each strain (see Materials and methods). Even with a conservative threshold of 3 standard deviations ($s \leq -3$ or ≥ 3, or absolute score $|s| \geq 3$) away from the WT , a large number (725) of single gene deletion strains were associated with one or more morphological defects and qualified as morphological hits (Fig 2). This result indicates that a large fraction (∼16%) of the non-essential genome directly or indirectly affects cell size and shape. Similar genomic commitment to cell size/shape was also observed for budding yeast [30].

Quantification of growth and cell cycle features across the genome

From the images, we also calculated the degree of constriction for each cell, and inferred the fraction of constricting cells in the population for each strain (see Materials and methods). The latter reflects the timing of initiation of cell constriction relative to the cell cycle. In addition, the analysis of the DAPI staining with the objectDetection module in Oufti [46] provided additional parameters, such as the number of nucleoids per cell and the fraction of cells with one versus two nucleoids. From the latter measurement, we estimated the relative timing of nucleoid separation. We also measured the degree of nucleoid constriction in each cell for each strain, and compared it to the degree of cell constriction to obtain the Pearson correlation between these two parameters, as well as the average degree of nucleoid separation at the onset of cell constriction (Appendix Fig S1E). As a result, each strain was associated with 5 cell cycle features, in addition to the 19 morphological features and 2 growth features mentioned above (see Table S1).

While the cell cycle features examined seemed to be less sensitive to gene deletion than cell morphology, there were still a high number (147) of gene deletions that were associated with one or more dramatically ($|s| \geq 3$) altered cell cycle features (Fig 2). Similarly, we identified over 169 mutants with severe ($|s| \geq 3$) growth defects (Fig 2) despite the growth medium being supplemented with amino acids.

Severe defects in growth, cell morphology or the cell cycle associated with a wide variety of cellular functions

For each feature, the genes deleted in mutant strains with a $|s| \geq 3$ encompassed a wide range of cellular functions based on a COG (Clusters of Orthologous Groups) distribution analysis (Fig 3, Appendix Fig S4). This diversity highlights the high degree of integration of cell morphology and the cell cycle in overall cellular physiology.

Certain COG families were statistically enriched for some phenotypes (Fig 3). We recovered expected associations, such as category D (cell cycle control, cell division and chromosome partitioning) with high mean length ($<L>$) and high length variability (CV_L), category L (DNA replication, recombination and repair) with high CV_L, and category M (cell wall/membrane/cell wall biogenesis) with high mean width ($<W>$) (Fig 3A). Indeed, defects in DNA partitioning and repair can lead to a cell division block [43], and impairment in cell envelope biogenesis has been reported to cause cell widening [6, 36]. COG categories associated with translation or some aspect of metabolism were, unsurprisingly, enriched in mutants with growth defects (Fig 3B).

Often, these COG enrichments were carried over to features (area, volume, perimeter, circularity, etc.) that directly relate to width and length (Fig 3). However, we also observed differential COG enrichments even for highly related features, highlighting the importance of considering features beyond mean and CV of length and width. For example, category U (intracellular trafficking, secretion and vesicular transport) was enriched among mutant strains with high mean area ($<A>$) and volume ($<V>$), but normal $<L>$ or $<W>$ (Fig 3A), suggesting that small deviations in length and width can combine to produce significant differences in area and volume. On the other hand, deletions in category C genes (energy production and conversion) were normally represented for most phenotypes, but were conspicuously underrepresented among mutants with high mean shape factors, to the point that it was barely associated with a high mean aspect ratio ($<Ar>$) and not at all with a high mean circularity ($<C>$) (Fig 3A, Appendix Fig S4A). Thus, deletion of genes involved in energy and conversion can increase or decrease the size of the cell without affecting its shape (aspect ratio and circularity), implying that defects in length and width are often compensatory for this category of mutants.

High-dimensional classification of the morphological mutants

While the gene deletion annotation of the Keio library is not perfect, our large dataset provided a powerful platform to examine global trends and to identify gene function enrichments in phenotypic classes of mutants with $|s| \geq 3$. First, we considered morphological phenotypes. Instead of ranking strains on a feature-by-feature basis, we sought to classify strains based on their combination of features, or 'phenoprints', to better capture the phenotypic complexity of morphology. Each strain in our dataset is characterized by scores for 19 morphological features. We added two growth-related features (OD_{max} and α_{max}) to this morphological phenoprint because growth rate is often implicitly assumed to control cell size. This assumption derives from the early observation that bacterial cell size (mean cell mass) scales with growth rate when the latter is modulated by varying the composition of the culture medium [47]. This scaling relationship is often referred to as the 'growth law'.

The combination of these 21 scores was used to classify a dataset composed of 240 wild-type replicates (controls) and the 797 mutant strains with a $|s| \geq 3$ for at least one morphological or growth feature. To transform our 21-dimension clustering problem into a simpler two-dimensional (2D) similarity map (see Materials and methods), we used the machine learning "t-distributed stochastic neighbor embedding" (tSNE) algorithm [58]. The principle of tSNE is to minimize distances between phenoprints with high mutual information. Taking advantage of the stochastic nature of tSNE, we generated 100 maps to identify stable clusters, or island using the density-based clustering algorithm dbscan [19]. This combined tSNE-dbscan approach identified multiple isolated islands formed by the same strains in each map (Fig 4A). In fact, more

than 90% of the strains were reproducibly (> 90% of the time) found within the same island of the "morpho archipelago" (Fig EV1). The wild-type replicates clustered together to form the 'WT' island while the mutant strains consistently separated in 17 islands (Fig 4A and Fig EV1). With 21 features, we might have expected a continuum of phenoprints representing the vast number of possible combinations of different phenotypes, which would have resulted in the absence of separated dense areas on the tSNE maps. Instead, the presence of natural boundaries between islands supports the idea that some feature combinations are favored.

Each island was characterized by an average phenoprint (Fig 4B), with a given feature often segregating in different islands. For example, slowly growing mutants were found in both islands 15 and 16, but mutants in island 15 were, on average, short with a comparatively normal width whereas mutants in island 16 were wide with a normal length (Fig 4B). Thus, island 16 illustrates a group of strains that departs from the growth law, as they produce cells that are larger than WT despite growing slower. Another departure from the growth law is illustrated by island 5, which includes strains with small cells but normal growth rate (Fig 4B).

Genes, functions and pathways associated with cell size and shape

Our tSNE classification identified many new genes associated with specific phenotypes, even for extreme ones. For example, island 17 grouped strains characterized by cells that were very long and highly variable in length (and consequently in area, volume, surface area and perimeter), but had a normal width (Fig 4B). Such a cell filamentation phenotype has been well studied, and our classification recovers expected gene deletions such as Δ*minC*, Δ*envC*, Δ*tatC* and Δ*dedD* (Fig 4C and Fig EV2A). But island 17 also includes 4 gene deletions (Δ*rdgB*, Δ*uup*, Δ*croE* and Δ*ydaS*) that were unknown for their cell filamentation phenotype, suggesting new or unappreciated functions connected to cell division. For example, Uup is a DNA-related protein known to prevent the precise excision of transposons [28]. The working model postulates that Uup interacts with the replisome to prevent replication forks stalling at the repeated sequences flanking transposons, a step required for the formation of a Holliday junction and excision [44]. Replisomes also frequently stop at other chromosomal regions during replication, which can cause DNA lesions [14]. If this DNA damages are left uncorrected, they lead to inhibition of cell division. The cell filamentation phenotype associated with the deletion of *uup* may suggest that Uup plays a fundamental role in limiting replisome stalling under normal growth conditions, possibly at structured DNA sites such as inverted repeats.

RdgB is an enzyme that reduces the levels of non-canonical purines deoxyinosine (dITP) and deoxyxanthosine (dXTP) to prevent DNA damage associated with their incorporation into the chromosome; rdgB becomes essential for viability in a *recA*- background [11, 37]. The high frequency of cell filamentation among Δ*rdgB* cells, despite the presence of a fully functional recombination machinery, underscores the importance of a tight control of dITP and dXTP levels in the cell.

The two remaining genes in island 17 were cryptic prophage genes *croE* and *ydaS* (Fig 4C and Figure EV2). They illustrate how this screen can identify functions for genes that are normally not expressed under normal growth conditions. Genes in the Keio collection were deleted by an in-frame replacement of a kanamycin resistance cassette that has a constitutive promoter and no transcriptional terminator, to ensure expression of downstream genes in operons [4]. However, for repressed or poorly expressed operons, the kanamycin cassette promoter can lead to unregulated expression of downstream genes in operons. This was the case for the *croE* and *ydaS* deletion strains, as cells became normal in cell length when the kanamycin cassette was excised (Fig EV2B and C). These results, together with the absence of phenotype associated with the deletions of the downstream genes, suggest that it was not the loss of *croE* and *ydaS* but rather the expression of the prophage genes located directly downstream (*ymfL* and *ydaT*, respectively) that was responsible for the observed cell filamentation phenotype. Consistent with our hypothesis, it has been postulated that *ymfL* is involved in cell division [41, 62]. While *ymfL* probably encodes a cell division inhibitor, the prophage gene *ydaT* likely inhibits cell division indirectly by acting on DNA replication or segregation, given the absence of well-segregated DAPI-stained nucleoids in filamentous Δ*ydaS* cells still carrying the kanamycin cassette (Fig EV2C).

Note that each island represented a continuum of phenotypes dominated by the features that lead to their clustering in one common island. For instance, island 2 contained deletion strains displaying the dominant phenotype of long, but not filamentous, cells (s for $<L>$ of 3.5, compared to 5.2 and 10.8 for islands 10 and 17, respectively). Beyond the global segmentation of the morpho-space, each island displayed some internal structure. This is illustrated in Fig 4D, which shows the gradient of the dominating ($<L>$) and secondary (CV_L) features within island 2.

This fine internal organization reflects the objective function of the tSNE algorithm, which seeks to minimize distances between similar phenoprints. This property provided us with an excellent layout to consider tSNE maps as networks (e.g., Fig 4C), from which we could perform local functional enrichment analyses based on gene ontology (GO) term enrichment. This approach enabled the functional annotation of the tSNE networks while taking into account the map topology, without explicit clustering (see Materials and methods). This functional analysis highlighted both expected and surprising functional associations with specific morphological phenoprints (Fig 4E). For example, the phenoprint dominated by slow growth and small cell size, which is a hallmark of starved cells, was, not surprisingly, associated with an enrichment of strains deleted for genes involved in sulfur assimilation and metabolism (Fig 4E). We also found that cell division and DNA recombination genes segregated into distinct islands (2 and 4, respectively), which reflects how these two groups of genes affect morphological features in different ways. Loss of cell division genes resulted in cell length increase across the cell population, causing a greater $<L>$ and CV_L (s = 4.6 and 4.9, respectively) whereas deletion of DNA recombination genes only affected cell division in the subset of cells that presumably encountered DNA lesions, increasing primarily CV_L (s = 4.9), but not significantly $<L>$ (s = 0.5) (Fig 4E).

In addition, we identified an enrichment for genes in the Enterobacterial Common Antigen (ECA) biosynthesis pathway (Fig 4E) among gene deletions that dramatically affected cell width control (island 16). The ECA mutants were wider, often lost their rod shape and formed rounder cells, as shown by their high aspect ratio score (Fig EV3A). This phenotype is reminiscent to the cell shape defects caused by drugs (e.g., fosfomycin) that inhibit peptidoglycan synthesis [33, 40]. Island 16 included other cell envelope mutants with a similar phenotype, such as gene deletions related to colonic acid biosynthesis or lipopolysaccharide (LPS) modification. These results are consistent with recent studies showing that cell shape deregulation can be caused by a competition between the ECA, LPS, CA and peptidoglycan precursor pathways for the same undecaprenyl phosphate lipid carrier [31, 32]. The phenotype of other gene deletions in island 16 could be rationalized with a similar competition argument, as several of them are related to central metabolism. The metabolic genes may be essential for the production of key metabolites important for the synthesis of cell envelope precursors. The $\Delta rapZ$ strain, which had severe cell width phenotypes (Fig EV3B), may be an example. RapZ post-transcriptionally regulates the amount of GlmS [22], which catalyzes the first committed step away from the upper glycolysis pathway and toward the synthesis of a central precursor (UDP-N-acetyl-α-D-glucosamine) for the biogenesis of peptidoglycan, LPS and ECA.

We also identified pathways associated with phenotypes that were not easy to rationalize. Deletion of genes encoding the high-affinity phosphate transporter (*pstACS*) resulted in a reduction in cell width ($<s>$ = -3.8, Fig EV3C), without significantly slowing down growth ($<s>$ for α_{max} = -0.4) (Fig 4E). Interestingly, deletion of genes encoding subunits of ATP synthase, which results in a metabolic switch to fermentation, lead to a decrease in average cell width (s = -4.3, Fig EV3C) with no change in average cell length (s = 0.2) or growth rate (s = 1.8) (Fig 4E). Since the cells were imaged during exponential phase, this phenotype could not be linked to their inability to grow to high cell density. This result suggests that the ATP synthase itself or differences in metabolism alter cell shape and size independently of growth rate.

Identification of genes affecting nucleoid separation and cell constriction dynamics

We applied the same tSNE analysis to the 7 cell cycle and growth features of the 264 strains displaying a severe defect ($|s| \geq 3$) for at least one cell cycle or growth feature. The 240 independent wild-type replicates were included in the analysis as controls. We robustly identified a WT island and 12 distinct mutant islands in this cell cycle space (Fig 5A). Each island was characterized by an average phenoprint (Fig 5B). Islands 11 and 12 were phenotypically close to WT. Islands 2 and 6 grouped mutants with growth defects and little to no cell cycle phenotypes (Fig 5B and C). The neighboring islands (3, 5, 8 and 9) were dominated by cell growth features with some combination of nucleoid separation and cell constriction defects. Four islands (1, 4, 7 and 10) grouped interesting gene deletion strains with altered cell cycle progression, but without significant growth defects (Fig 5B and D).

Functional analysis on all strains identified GO term enrichments with phenoprints that show strong growth defect (Fig 5E). We did not find any GO term enrichment associated with cell cycle defects independently of growth. Furthermore, the proportion of genes of unknown functions was particularly high for cell cycle-specific islands (Fig 5F), reaching proportions above 40% for islands 4 and 7. These observations highlight the limited extent of our knowledge about the genetic basis of nucleoid and cell constriction dynamics.

Our analysis of nucleoid separation and cell constriction provided a genome-wide perspective on the processes affecting DNA segregation and cell division. While each event has been investigated for years at the molecular level, we know little about their coordination. We found that nucleoid separation is tightly correlated with the initiation of cell constriction across the ∼4,000 deletion strains (Pearson $\rho = 0.65$, Fig 6A) and at the single-cell level (Appendix Fig S1E). A well-known genetic factor involved in this coordination is MatP [42]. This DNA-binding protein organizes and connects the chromosomal terminal macrodomain (*ter*) to the division machinery [18]. Consistent with this function, we observed that the Δ*matP* mutant, which segregated into island 4, failed to coordinate nucleoid separation with cell constriction, as evidenced by the separation between the curves in Fig 6B. Interestingly, the curves also showed that the Δ*matP* mutant separates its nucleoid early while dividing at about the same time as WT (Fig 6A and B). This surprising result suggests that MatP delays nucleoid separation.

The remaining 16 genes from island 4, which also displayed an early nucleoid separation phenotype, had either an uncharacterized function (e.g., *ypfH*) or a function unrelated to nucleoid dynamics such as *polA* and *pldB*, which encode DNA repair protein Pol I and lysophospholipase L2, respectively (Fig 6C).

The 30 mutants grouped in island 7 were primarily characterized by an early initiation of cell constriction (Fig 5B and D), to the point that the timing of cell constriction and nucleoid separation virtually collapsed. Fig 6D shows two such examples with Δ*ybaN* and Δ*hlsU*. YbaN is a protein of unknown function. HslU has two functions in the cell, one as a subunit in a protease complex with HslV, and the other as a chaperone [50, 52]. Since we did not observe any significant defect in cell constriction timing for the Δ*hslV* mutant, the Δ*hlsU* phenotype is more likely linked to the chaperone activity.

Identification of cell size control mutants

How cells achieve size homeostasis has been a longstanding question in biology. While the control mechanism at play remains under debate [1, 12, 24, 27, 29, 54–56, 61], we and others have recently shown that under the growth conditions considered in this study, *E. coli* follows an adder principle in which cells grow a constant length (ΔL) before dividing [12, 55]. We sought to use this screen to survey the role of genes in cell length control. We first explored the relationship between $<L>$ and CV_L among mutants. Globally, the degree of correlation between these two variables displayed two regimes, with no correlation for 'short' mutants and a strong positive correlation for 'long' mutants (Fig 7A).

The observation that short mutants displayed, on average, a normal CV_L (same noise as WT) indicates that they regulate their length distribution as precisely as WT. These results suggest that the adder principle, and therefore the timing of cell division, is just as precise in short mutants as in WT cells. This result is interesting because short mutants have traditionally received a lot of attention in cell size control studies. A well-known short mutant in *E. coli* is the *ftsA** strain, which is thought to misregulate size control by triggering division prematurely [21, 26]. However, we found that, similar to the trend shown by short mutants in our screen, *ftsA** cells constrict at the same cell age as WT (Fig 7B). In hindsight, this result makes sense since the WT and *ftsA** strains have the same doubling time [21] and therefore take the same amount of time to divide. Perhaps a better way to consider short mutants with normal CV_L is not as mutants that have a premature division, but as small-adder mutants that add an abnormally small cell length increment ΔL between divisions.

Long mutants, on the other hand, tended to lose their ability to maintain a narrow size distribution, as CV_L increased with $<L>$ (Fig 7A). The origin for an increase in CV_L may signify a loss of precision in the timing of division, but it may alternatively originate from an aberrant positioning of the division site (or both). The $\Delta minC$ mutant is an example of aberrantly large CV_L (Fig 7C) due to the mispositioning of the division site and not to a defective adder [12]. This class of mutants can easily be identified in our dataset by their large variability in division ratios (CV_{DR}). Conversely, a high CV_L associated with a normal variability in division ratios points to a mutant that has a more variable ΔL between divisions.

We suspected that interesting cell size control mutants might be missed by only considering CV_L. The distribution of cell lengths in a population is a convolution of cell length distributions at specific cell cycle periods. Since there is significant overlap in length distributions between cell cycle periods, a substantial change in CV_L at a specific cell cycle period (e.g., cell constriction) does not necessarily translate into obvious changes in CV_L of the whole population, as shown in simulations (Fig EV4). Our screen allowed us to identify constricting cells and hence to determine the length variability for the cell constriction period. This cell cycle period-specific analysis identified $\Delta mraZ$ as a potential gain-of-function cell size homeostasis mutant (Fig 7C). For this mutant, division (CV_{DR}) and growth rate [17] were normal, but the length distribution of its constricted cells (CV_L= 0.05) was remarkably narrower than that of WT constricted cells (CV_L = 0.12). MraZ is a highly conserved transcriptional regulator that downregulates the expression of the *dcw* cluster [17], which includes cell wall synthesis and cell division genes [3]. Our data suggests that MraZ and the regulation of the *dcw* cluster affect the balance between cell growth and division.

Dependencies between cellular dimensions and cell cycle progression

A fundamental question in biology is how cells integrate cellular processes. A common approach to address this question is to look at co-variation between processes or phenotypes following a perturbation (e.g., mutation, drug treatment). However, using a single type of perturbation can lead to misinterpretation, as the perturbation may affect the co-varying phenotypes independently. Increasing the number of independent perturbations alleviates the interpretation problem by averaging out the specific effect associated with each perturbation. Therefore, the large number and variety of mutants in our study provided an opportunity to identify global effects and dependencies between morphological, cell cycle and growth phenotypes through correlation analysis.

To build an interaction network, we used the well-established, information-theoretic algorithm ARACNE [39]. This method considers all pairwise correlations between features at the same time and identifies the most relevant connections by removing those that are weak or that can be explained via more correlated paths. In this analysis, we only considered quantitative non-collinear features that describe morphology, nucleoid shape, growth, nucleoid separation and cell constriction (see Materials and methods). The resulting network recovered obvious connections, such as the relation of area with length and width. It also showed the absence of a connection between growth rate (α_{max}) and size features ($<A>$, $<L>$ and $<W>$) (Fig 8A), again underscoring the independence of cell size from growth rate under a given growth condition (Fig EV5). In

fact, growth rate features displayed little connectivity to morphological or cell cycle features (Fig 8A), as shown by their close-to-zero Kendall correlations τ (Fig EV5, note that Kendall ranked correlation was selected over Pearson correlation because of the heavy asymmetric left tail in the distribution of α_{max}).

Another interesting lack of connection was between $<L>$ and $<W>$ (Fig 8A), as these two variables were largely uncorrelated ($\rho = 0.11$, Fig 8B). This result is significant from a cell size control standpoint. If cells were controlling their size by monitoring how much volume or surface area they add during growth, we would expect a global anti-correlation between length and width such that an increase in cell length would be, on average, compensated by a decrease in width, and vice versa. The lack of correlation argues that cell length and width are controlled independently.

The overall structure of the network (Fig 8A) revealed that the cell cycle features (yellow nodes) are connected to morphological features (blue nodes) through the dimensions of the nucleoid (grey nodes). The mean cell area and mean nucleoid area (considering the sum of nucleoids in the cell) were highly positively correlated ($\rho = 0.83$), in a growth rate-independent manner (Fig 8C).We previously showed by time-lapse imaging of single cells that the nucleoid size linearly increases with cell size throughout the cell cycle [46]. Here, we found that nucleoid size remarkably scales with cell size across ~4,000 mutants despite the wide range of cellular dimensions present in the Keio collection: small mutants had a small nucleoid size, and big mutants had a big nucleoid size (Fig 8C). This linear relationship held true regardless of the number of nucleoid per cell (Fig 8D). In addition to its strong positive correlation with the average cell size, the average nucleoid size was negatively correlated with the relative timing of nucleoid separation (ρ = -0.49, Fig 8E). These connections suggest a dependency between cell size and nucleoid separation: the bigger the cell, the bigger the nucleoid is and the earlier nucleoid separation occurs in relative cell cycle unit (Fig 8F). The relative timing of cell constriction was also negatively correlated (although to a lesser degree) with the average nucleoid size (ρ = -0.25) and the average cell size (ρ = -0.19), thus causing the gap between nucleoid separation and cell constriction to increase with cell size.

Discussion

In this study, we used a multi-parametric approach to quantitatively survey the role of all non-essential *E. coli* genes on cell shape, cell size, cell growth and the late cell cycle stages, nucleoid separation and cell constriction. The results provide a valuable resource of phenotypic references for both characterized and uncharacterized genes, as well as a rich dataset to explore the correlation structure between morphological, growth and cell cycle features at the system level.

The large proportion of genes and the wide variety of functions impacting cell size and shape and the progression of late cell cycle stages (Fig 2 and 3, Appendix Fig S4) underscore the degree of integration of cell morphogenesis and cell cycle progression in all aspects of *E. coli* cell physiology. It also implies that most morphological and cell cycle phenotypes cannot easily be imputed to a specific pathway or cluster of genes. In fact, genes involved in the same cellular process can have very different, and even sometimes opposing, effects. This is illustrated by genes associated with translation. Deletion of ribosomal subunit genes leads to a diversity of morphological phenotypes, such as thin ($\Delta rpsY$), wide ($\Delta rpsO$), short ($\Delta rplY$), and short and thin ($\Delta rpsT$). This diversity of phenotypes is also observable for deletions of genes encoding enzymes that modify ribosome RNAs or tRNAs (e.g., $\Delta rsmD$ and $\Delta mnmC$ strains are long, whereas $\Delta rluD$ and $\Delta truA$ strains are wide). The latter suggests an unexpected role for RNA modifications in cell morphogenesis.

Overall, this study greatly expands the number of genes associated with cell morphogenesis ($\sim$ 800) and the cell cycle ($\sim$ 150). Notably, it provides a phenotype for 480 genes of uncharacterized function (out of 1250 so-called 'y-genes'). The proportion of mutant strains in this category is substantially higher than the proportion over the whole genome (38% versus 29%), suggesting that the phenotypes that we quantified and the growth conditions we used are favorable to explore the function of these genes and learn new biology.

This study also revealed new phenotypes for previously characterized gene deletions. We mentioned above the unexpected filamentation phenotype of the Δ*uup* strain (Fig 4 and Fig EV2A) and proposed a tentative connection between the known function in precise transposon excision and DNA damage through replisome stalling. We also identified unanticipated links. For example, the requirement for lysophospholipase L2 (PldB) in the coupling of nucleoid separation and cell constriction (Fig 6C) suggests a connection between phospholipid metabolism and the coordination of late cell cycle stages.

We adopted an original approach combining the tSNE and dbscan algorithms to effectively cluster strains with similar phenoprints into islands (Fig 4 and 5). This granular representation of the phenotypic space allowed us to expand on well-studied archetypal phenotypes such as 'filamentous' and 'fat' (islands 17 and 16 of the morpho archipelago, respectively, see Fig 4). This classification also allowed us to populate less well-studied phenotypes, from which we can learn new insight into cell morphogenesis and the cell cycle. For example, the substantial number of thin mutants reported here may prove as valuable as fat mutants to study cell morphogenesis from a different angle. The clustering results also revealed entirely new classes of mutants (e.g., island 1 in the morpho archipelago and islands 1, 4 and 7 in the cell cycle archipelago). In our view, the cell cycle islands 1, 4 and 7 are particularly interesting because they offer a genetic toolkit to explore nucleoid and cell constriction dynamics, which have remained poorly understood despite their essential role in cellular replication.

The phenoprints reported in this study are necessarily tied to the specific experimental conditions of the screen. Differences in growth conditions lead to different metabolic requirements and growth limitations. For instance, none of the mutant strains auxotrophic for nucleotides were able to grow in our synthetic medium, which lacks nucleotide precursors. We note that growth in 96-well plates likely corresponds to micro-aerophilic conditions. Accordingly, we identified morphological deviations for strains deleted for genes known to be only expressed under micro-aerophilic or anaerobic conditions, revealing new metabolic connections to cell morphogenesis. For example, deletion of *ybcF*, which is predicted to encode an enzyme involved in anaerobic purine degradation [53], results in a fat cell phenotype (Supplementary file 2).

In this study, each gene deletion can be seen as a perturbation. The sheer number of perturbations (∼4,000) guarantees a large number of independent perturbations and offers a unique opportunity to infer the underlying structure of the correlations between the different phenotypes. Such relationships, or lack thereof, can be very informative. For instance, we found that cell length and width are largely uncorrelated (Fig 8B), suggesting that cells do not control their size by monitoring their surface area or volume, but rather control their length and width independently. We also found that growth rate is not predictive of cell size. When growth rate is varied with growth media of different chemical composition, cell size scales with growth rate [47]. This "growth law" is often interpreted as growth rate dictating the average size of the cell, although it does not explain why temperature can alter growth rate without any significant effect on mean cell size [47]. The lack of substantial correlation between cell size and growth rate across 4,000 genetic perturbations that affect various cellular functions (Fig EV5) shows unambiguously that growth rate itself does not set cell size. Our results support the idea that the original scaling observation with different growth media likely stems from differences in cellular metabolism and that growth rate and cell size are metabolically co-regulated [57]. Indeed, metabolically-starved mutants (which are common in island 15 of the morpho archipelago) displayed both slow growth and small size. Growth rate also correlated poorly with the relative timing of nucleoid separation and cell constriction (Fig EV5). The absence of correlation between growth rate and the timing of these cell cycle events was also observed for the wild-type strain when the growth rate was varied by changing the composition of the growth medium [16]. Collectively, our findings show that the cell can accommodate a large range of sizes and relative timings of nucleoid segregation and cell division with no effect on growth rate, and vice versa. This flexibility may offer greater evolvability of cellular dimensions and cell cycle progression.

The complexity of cellular systems can sometimes be reduced to simple quantitative relationships, or 'biological laws', which have been very useful in identifying the governing principles by which cells integrate various processes [48]. Our correlation analysis identified a 'nucleoid law' that describes the linear relationship

between nucleoid size and cell size. This remarkable scaling property is independent of growth rate and holds across the wide range of cellular perturbations present in the ∼4,000 deletion strains tested in this study (Fig 8C). The nucleoid law draws a striking parallel with the 100-year-old observation that nucleus size scales with cell size in eukaryotes [13], an empirical relationship that has been reported for many eukaryotic cell types since [60]. This suggests a universal size relationship between DNA-containing organelles and the cell across taxonomic kingdoms, even for organisms that lack a nuclear envelope.

Our information-theoretic Bayesian network analysis (Fig 8) enabled us to go beyond pairwise correlations by integrating the complex set of interdependences between morphology, growth and cell cycle events. This analysis unveiled an unexpected connection between cell size and the relative timing of nucleoid separation and cell constriction through nucleoid size across thousands of genetic perturbations (Fig 8E and F). This finding suggests that the size of the nucleoid and, by extension, the overall structure of the chromosome are important elements of the coordination mechanism between cell morphogenesis and the cell cycle.

Materials and Methods

Bacterial growth conditions

The Keio collection contains 3,787 annotated single-gene in-frame deletion strains, 412 strains (also known as JW strains) with kanamycin cassette inserted at unknown locations, and the remainder (28) were repeats [4]. All strains, including *E. coli* K12 BW25113 [15] and derivatives (strains from Keio collection), as well as *E. coli* K12 MG1655 and the isogenic *ftsA** [21] were grown in LB medium (10 g/L NaCl, 5 g/L yeast extract, 10 g/L tryptone) or M9 medium (6 g/L $Na_2HPO_4{\cdot}7H_2O$, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g NH_4Cl, 2 mM $MgSO_4$, 1 μg/L thiamine) with 0.2% glucose as the carbon source and supplemented or not with 0.1% casamino acids as specified in the text and figure legends.

Screening set-up and microscopy

All *E. coli* strains were grown overnight at 30°C in 96-well plates in M9 supplemented with 0.1% casamino acids, 0.2% glucose and kanamycin (30 μg/mL). Cultures were diluted 1:300 in 150 μL of fresh M9 medium supplemented with 0.1% casamino acids and 0.2% glucose, and grown in 96-well plates at 30°C with continuous shaking in a BioTek plate reader. DAPI was added to the cultures to a final concentration of 1 μg/mL 15 to 20 min prior imaging. All (parent and mutant) strains were sampled within a very narrow range of OD_{600nm} (0.2 ± 0.1; min = 0.108 ; max = 0.350) corresponding to the exponential growth phase. We did not detect any trend between morphological/cell cycle features and the OD_{600nm} at which each culture was sampled. Cells were deposited (0.5 μL per strain) on a large, 0.75-μm thick, M9-supplemented agarose pads with a multichannel pipet. The pads were made by pouring warm agarose containing supplemented M9 medium between a (10.16 x 12.7 x 0.12 cm) glass slide and a (9.53 x 11.43 cm) n° 2 coverglass (Brain Research Laboratories, Newton, MA, USA).

Microscopy was performed on an Eclipse Ti-E microscope (Nikon, Tokyo, Japan) equipped with Perfect Focus System (Nikon, Tokyo, Japan) and an Orca-R2 camera (Hamamatsu Photonics, Hamamatsu City, Japan) and a phase-contrast objective Plan Apochromat 100x/1.45 numerical aperture (Nikon, Tokyo, Japan). The initial field of view for each strain was chosen manually and 9 images were taken automatically over a 3x3 square lattice with 200 nm step, using 80 ms exposure for phase contrast and 600 ms exposure for the DAPI channel using Nikon Elements (Nikon, Tokyo, Japan).

Image processing

Cell outlines were detected using Oufti software [46] available at `http://oufti.org/`. All data processing was then performed using MATLAB (The MathWorks Inc., Natick, MA, 2000). Custom-built codes were used to automate the aggregation of data from the cell outlines of all the strains.

For cell and nucleoid detections, the same parameters in the Oufti's cellDetection and objectDetection were consistently used. In order to avoid unnecessary bias in the cell outlines, the parameters defining the initial guess for the cell contour fit were set to intermediate values, while the parameters constraining the fit of the final outline were set to negligible values. For example, we increased the *fsmooth* parameter value to 100 in order to capture both short and long cells, and we set the width spring constant parameter *wspringconst* to 0 so as to avoid biasing the cell width estimate toward the initial guess value. The edges in the DAPI fluorescence signal were detected with a Laplacian of Gaussian filtering method that takes into account the dispersion of the point spread function (PSF) of our microscopy setup at a wavelength of 460 nm (input parameter σ_{PSF} set to 1.62 pixels).

Data analysis

Dataset curation – Support Vector Machine model. Due to the size of the dataset ($> 1,500,000$ cells detected globally), we adopted an automated approach to identify poorly (or wrongly) detected cells across the entire dataset. We developed an SVM model based on 16 normalized features: cell length, cell width, cell area, cell volume, cell perimeter, cell constriction degree, division ratio, integrated phase signal, integrated DAPI fluorescence signal, mean cell contour intensity in phase contrast, variability of cell width along the cell, nucleoid area, single cell nucleoid variability, circularity ($2 \times \pi \times cellarea/(cellperimeter)^2$), nucleoid intensity and number of nucleoids. We trained a binary classifier (positive or negative) over wild-type strain replicates as well as 419 mutants with the most severe morphological defects prior to data curation. We visually scored 145,911 cells and used 30% of them (43,774) to train the model. The model was evaluated using a k-fold cross-validation approach, leading to a generalized misclassification rate of 10%. We used the remaining 70% of the data set (102,137 cells) to validate the model. This SVM classifier achieves a balanced classification rate of 84% and features an AUROC of 0.94 (Appendix Fig S1B). Furthermore, the resulting group of false negatives was not significantly different from the true positives (Appendix Fig S1C and D), indicating that the classification did not introduce a bias by excluding a specific class of 'good' cells from the analysis.

Data processing. For each feature, we checked and corrected for any bias associated with plate-to-plate variability, differences in position on the 96-well plates, timing of imaging and optical density of the culture (Appendix Fig S2 and S3). For each plate, we set the median values of each feature, F, to the median feature value of the parental strain. The F values were transformed into normalized scores by a transformation akin to a z-score transformation but more robust to outliers.

$$s = 1.35 \times (F_i - median(F_i^{WT}))/iqr(F_i^{WT})$$

where F_i is the corrected value for the mutant strains for feature i, F_i^{WT} is the value for the wild-type strain for feature i, and iqr stands for interquartile range. As the interquartile range of normally distributed data is equal to 1.35 times their standard deviation, we scaled the score by this factor so as to express the scores in terms of standard deviations away from the median.

The temporal biases for the fraction of cells committed (or not) to division and the fractions of cells with 1, 2 or more nucleoids were corrected using a Dirichlet regression to maintain the relative proportions between classes (Appendix Fig S3) [38].

Data exploration, dimensionality reduction and clustering. A similarity measure between strains was needed to identify and separate different phenoprints. This measure was then used as an input for a dimensionality reduction algorithm to group strains together. Pearson correlations or Euclidean distances classically provide such similarity measures, and Principle Component Analysis (PCA) and/or hierarchical or k-means clustering are often used. However, PCA tends to explode datasets and Pearson correlations do not always reflect the desired type of similarity. As an extreme case, consider two strains with two phenoprints that are proportional, one with values within a very small score range, such as [-1 1], while the other with score values spanning the [-10 10] range. These two strains will get a maximal similarity measure through

a correlation analysis, despite the fact that the first strain is wild-type-like while the other is an outlier. Instead we chose to use a recently described algorithm, called t-distributed Stochastic Neighbor Embedding, or t-SNE [58], to project our multidimensional datasets in 2 dimensions and generate, at the same time, similarity measures between strains. t-SNE estimates low-dimensional space distances between points based on their similarity, as opposed to dissimilarity as in the case of PCA, thereby highlighting local similarities rather than global disparities.

We used the stochastic nature of the t-SNE algorithm to evaluate the robustness of the resulting projection by repeating the procedure multiple times (n = 100 for each tSNE map). We coupled this dimensional reduction procedure with a density-based clustering algorithm, dbscan [19]. The two input parameters of the dbscan algorithm, ϵ and minPoints, were optimized so as to generate a maximum number of islands without separating the bulk of WT strains in two or more islands. We identified as robust clusters the groups of strains falling together in the same clusters more than 90% of the time.

Map exploration. Each t-SNE map is a similarity map, and can therefore be treated as a network where the nodes represent strains and the edges the Euclidean distance between strains in the tSNE map. Building up on recent quantitative network analysis tools [5], we calculated the local enrichment in the maps of different strain-associated attributes, such as COG and GO terms. Briefly, the sum of the attributes in a local area (within a radius around each point, defined as a percentile of the distribution of all the distances between points) was compared to a background score (defined as the average score obtained over 1000 identical maps with randomly permutated attributes) with a hypergeometric test. The significant local enrichments were considered at a threshold of 0.05 after a false discovery rate correction that used the Benjamini-Hochberg-Yekutieli algorithm, taking into account dependencies between tests [9].

Cluster of orthologous gene enrichment analysis. We associated *E. coli* BW25113 genes with COGs using the web server [59]. The enrichment analyses were performed using a custom-built algorithm in MATLAB based on a two-tailed hypergeometric test to compute p-values, which were subsequently adjusted with the Benjamini-Hochberg False Discovery Rate procedure [8]. Because the COG categories are largely independent, we did not consider any correction for the dependence between tests.

Gene ontology analysis. We used ontologies from the Gene Ontology website (`http://www.geneontology.org/ontology/gene_ontology.obo`, version 2016-05-27) [2], and annotations were obtained from EcoCyc for *E. coli* strain MG1655 [35]. Analysis was performed using a MATLAB custom-built algorithm that includes a hypergeometric test to compute p-values that were subsequently adjusted with the Benjamini-Hochberg-Yekutieli False Discovery Rate procedure [9].

Bayesian network. The Bayesian network presented in Fig 8 was generated in R with the bnlearn package [49], using the ARACNE algorithm as described in [39]. The network was bootstrapped 200 times, and all the edges were identified in more than 70% of the networks. We assessed the strength and the origin of collinearity among features using Belsley diagnostic method [7], with the in-built collintest.m function in MATLAB. We excluded features associated with a 'condition number' above the classical threshold of 30.

Data representation. All graphs were generated using MATLAB, except for the networks in Fig 4C and Fig 8A panels, which were created using Cytoscape v3.2 [51] and the Rgraphviz package in R [34], respectively. For Fig 4C, we used the edge-weighted, spring embedded algorithm in-built in Cytoscape. We considered the pairwise Euclidean distances between the 8 strains of island 17 as the weights of the edges connecting the nodes (or strains).

The density scales in scatter plots represent the number of points around each point in a radius equal to the 0.03 percentile of the pairwise distances distribution.

The WT isocontours representing the 0.5, 0.75 and 0.95 probability envelopes for the 240 WT replicates were calculated using a 2D kernel density estimation function over a 128-by-128 lattice covering the entire set of points (Supplementary file 3). The bandwidth of the kernel was internally determined [10].

The piecewise linear model where both lines intersect at the regime change

$$y = \frac{(1 - sign(x - d))}{2} \times (ax + b) + \frac{(1 + sign(x + d))}{2} \times (cx + d(a - c) + b)$$

was fitted to the binned data ($<L>$ versus CV_L for all strains) in Fig 7A using MATLAB built-in non-linear least-squares algorithm. The resulting parameters values (with 95% confidence bounds) were: a=0.007 [-0.017;0.030], b=0.195 [0.124;0.265], c=0.195 [0.138;0.251], d=3.392 [3.29;3.495] .

Simulations of cell length distributions. Cell length distributions at any given cell age were assumed to be log-normally distributed with different dispersion values. The CV of the distribution for the WT strain (CV = 0.11) was previously experimentally determined [12]. The cell length distributions at 100 different ages equidistantly distributed between 0 (birth) and 1 (division) were convolved with the cell age distribution, assuming an exponentially growing culture, $Pr(age) = 2^{-age}$.

Acknowledgements

We are grateful to the Yale E. coli Genetic Stock Center for providing a large number of strains. We also thank Pr. William Margolin for the kind gift of the *E. coli* MG1655 strain and the *ftsA** derivative. This work was partly supported by the National Institutes of Health (R01 GM065835 to C.J.-W.). We also thank the Jacobs-Wagner laboratory for fruitful discussions and for critical reading of the manuscript. C.J.-W. is an investigator of the Howard Hughes Medical Institute.

Author's contributions

C.J.-W, and M.C. designed experiments. G.S.D and M.C performed experiments. M.C. performed high-throughput imaging and statistical analyses. C.J.-W. supervised the project. C.J.-W. and M.C. wrote the manuscript.

References

1. A. Amir. Cell size regulation in bacteria. *Phys Rev Lett*, 112(20), 2014.

2. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000.

3. JA Ayala, T Garrido, MA De Pedro, and M Vicente. New comprehensive biochemistry, vol 27: Bacterial cell wall, 1994. ISBN 9780444880949, eBook ISBN: 9780080860879.

4. T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of *Escherichia coli* k-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol*, 2:2006 0008, 2006.

5. A. Baryshnikova. Systematic functional annotation and visualization of biological networks. *Cell Syst*, 2(6):412–21, 2016.

6. G. J. Bean, S. T. Flickinger, W. M. Westler, M. E. McCully, D. Sept, D. B. Weibel, and K. J. Amann. A22 disrupts the bacterial actin cytoskeleton by directly binding and inducing a low-affinity state in mreb. *Biochemistry*, 48(22):4852–7, 2009.

7. David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression diagnostics : identifying influential data and sources of collinearity*. Wiley series in probability and mathematical statistics. Wiley, New York, 1980.

8. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, 57(1):289–300, 1995.

9. Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, 29(4):1165–1188, 2001.

10. Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Ann Stat*, 38(5):2916–2957, 2010.

11. B. Budke and A. Kuzminov. Production of clastogenic dna precursors by the nucleotide metabolism in *Escherichia coli*. *Mol Microbiol*, 75(1):230–45, 2010.

12. M. Campos, I. V. Surovtsev, S. Kato, A. Paintdakhi, B. Beltran, S. E. Ebmeier, and C. Jacobs-Wagner. A constant size extension drives bacterial cell size homeostasis. *Cell*, 159(6):1433–46, 2014.

13. E. G. Conklin. Cell size and nuclear size. *J Exp Zool*, 12(1):1–98, 1912.

14. M. M. Cox, M. F. Goodman, K. N. Kreuzer, D. J. Sherratt, S. J. Sandler, and K. J. Marians. The importance of repairing stalled replication forks. *Nature*, 404(6773):37–41, 2000.

15. K. A. Datsenko and B. L. Wanner. One-step inactivation of chromosomal genes in *Escherichia coli* k-12 using pcr products. *Proc Natl Acad Sci U S A*, 97(12):6640–5, 2000.

16. T. Den Blaauwen, N. Buddelmeijer, M. E. Aarsman, C. M. Hameete, and N. Nanninga. Timing of ftsz assembly in *Escherichia coli*. *J Bacteriol*, 181(17):5167–75, 1999.

17. J. M. Eraso, L. M. Markillie, H. D. Mitchell, R. C. Taylor, G. Orr, and W. Margolin. The highly conserved mraz protein is a transcriptional regulator in *Escherichia coli*. *J Bacteriol*, 196(11):2053–66, 2014.

18. O. Espeli, R. Borne, P. Dupaigne, A. Thiel, E. Gigant, R. Mercier, and F. Boccard. A matp-divisome interaction coordinates chromosome segregation with cell division in *E. coli*. *EMBO J*, 31(14):3198–211, 2012.

19. M. Ester, H-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

20. R. E. Fan, P. H. Chen, and C. J. Lin. Working set selection using second order information for training support vector machines. *J Mach Learn Res*, 6:1889–1918, 2005.

21. B. Geissler, D. Shiomi, and W. Margolin. The ftsa* gain-of-function allele of *Escherichia coli* and its effects on the stability and dynamics of the z ring. *Microbiology*, 153(Pt 3):814–25, 2007.

22. Y. Gopel, K. Papenfort, B. Reichenbach, J. Vogel, and B. Gorke. Targeted decay of a regulatory small rna by an adaptor protein for rnase e and counteraction by an anti-adaptor rna. *Genes Dev*, 27(5):552–64, 2013.

23. V. Graml, X. Studera, J. L. Lawson, A. Chessel, M. Geymonat, M. Bortfeld-Miller, T. Walter, L. Wagstaff, E. Piddini, and R. E. Carazo-Salas. A genomic multiprocess survey of machineries that control and link cell shape, microtubule organization, and cell-cycle progression. *Dev Cell*, 31(2):227–39, 2014.

24. L. K. Harris and J. A. Theriot. Relative rates of surface and volume synthesis set bacterial cell size. *Cell*, 165(6):1479–92, 2016.

25. T. Hastie, R. Tibshirani, and J. Freidman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, 2009.

26. N. S. Hill, R. Kadoya, D. K. Chattoraj, and P. A. Levin. Cell size and the initiation of dna replication in bacteria. *PLoS Genet*, 8(3):e1002549, 2012.

27. P. Y. Ho and A. Amir. Simultaneous regulation of cell size and chromosome replication in bacteria. *Front Microbiol*, 6:662, 2015.

28. J. D. Hopkins, M. Clements, and M. Syvanen. New class of mutations in *Escherichia coli* (uup) that affect precise excision of insertion elements and bacteriophage mu growth. *J Bacteriol*, 153(1):384–9, 1983.

29. S. Iyer-Biswas, C. S. Wright, J. T. Henry, K. Lo, S. Burov, Y. Lin, G. E. Crooks, S. Crosson, A. R. Dinner, and N. F. Scherer. Scaling laws governing stochastic growth and division of single bacterial cells. *Proc Natl Acad Sci U S A*, 111(45):15912–7, 2014.

30. P. Jorgensen, J. L. Nishikawa, B. J. Breitkreutz, and M. Tyers. Systematic identification of pathways that couple cell growth and division in yeast. *Science*, 297(5580):395–400, 2002.

31. M. A. Jorgenson, S. Kannan, M. E. Laubacher, and K. D. Young. Dead-end intermediates in the enterobacterial common antigen pathway induce morphological defects in *Escherichia coli* by competing for undecaprenyl phosphate. *Mol Microbiol*, 100(1):1–14, 2016.

32. M. A. Jorgenson and K. D. Young. Interrupting biosynthesis of o antigen or the lipopolysaccharide core produces morphological defects in *Escherichia coli* by sequestering undecaprenyl phosphate. *J Bacteriol*, 198(22):3070–3079, 2016.

33. F. M. Kahan, J. S. Kahan, P. J. Cassidy, and H. Kropp. The mechanism of action of fosfomycin (phosphonomycin). *Ann N Y Acad Sci*, 235(0):364–86, 1974.

34. D.H. Kasper, J. Gentry, L. Long, R. Gentleman, S. Falcon, F. Hahne, and D. Sarkar. Rgraphviz: Provides plotting capabilities for R graph objects. http://bioconductor.org/packages/release/bioc/html/rgraphviz.html.

35. I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martinez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muniz-Rascado, Q. Ong, S. Paley, I. Schroder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen, and P. D. Karp. Ecocyc: fusing model organism databases with systems biology. *Nucleic Acids Res*, 41(Database issue):D605–12, 2013.

36. T. K. Lee, C. Tropini, J. Hsin, S. M. Desmarais, T. S. Ursell, E. Gong, Z. Gitai, R. D. Monds, and K. C. Huang. A dynamically assembled cell wall synthesis machinery buffers cell growth. *Proc Natl Acad Sci U S A*, 111(12):4554–9, 2014.

37. L. Lukas and A. Kuzminov. Chromosomal fragmentation is the major consequence of the rdgb defect in *Escherichia coli*. *Genetics*, 172(2):1359–62, 2006.

38. M.J. Maier. Dirichletreg: Dirichlet regression for compositional data in r, January http://epub.wu.ac.at/4077/1/Report125.pdf 2014.

39. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform*, 7 Suppl 1:S7, 2006.

40. J. L. Marquardt, E. D. Brown, W. S. Lane, T. M. Haley, Y. Ichikawa, C. H. Wong, and C. T. Walsh. Kinetics, stoichiometry, and identification of the reactive thiolate in the inactivation of udp-glcnac enolpyruvoyl transferase by the antibiotic fosfomycin. *Biochemistry*, 33(35):10646–51, 1994.

41. P. Mehta, S. Casjens, and S. Krishnaswamy. Analysis of the lambdoid prophage element e14 in the *E. coli* k-12 genome. *BMC Microbiol*, 4:4, 2004.

42. R. Mercier, M. A. Petit, S. Schbath, S. Robin, M. El Karoui, F. Boccard, and O. Espeli. The matp/mats site-specific system organizes the terminus region of the *E. coli* chromosome into a macrodomain. *Cell*, 135(3):475–85, 2008.

43. E. Mulder and C. L. Woldringh. Actively replicating nucleoids influence positioning of division sites in *Escherichia coli* filaments forming cells lacking dna. *J Bacteriol*, 171(8):4303–14, 1989.

44. D. Murat, P. Bance, I. Callebaut, and E. Dassa. Atp hydrolysis is essential for the function of the uup atp-binding cassette atpase in precise excision of transposons. *J Biol Chem*, 281(10):6850–9, 2006.

45. Y. Ohya, J. Sese, M. Yukawa, F. Sano, Y. Nakatani, T. L. Saito, A. Saka, T. Fukuda, S. Ishihara, S. Oka, G. Suzuki, M. Watanabe, A. Hirata, M. Ohtani, H. Sawai, N. Fraysse, J. P. Latge, J. M. Francois, M. Aebi, S. Tanaka, S. Muramatsu, H. Araki, K. Sonoike, S. Nogami, and S. Morishita. High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci U S A*, 102(52):19015–20, 2005.

46. A. Paintdakhi, B. Parry, M. Campos, I. Irnov, J. Elf, I. Surovtsev, and C. Jacobs-Wagner. Oufti: an integrated software package for high-accuracy, high-throughput quantitative microscopy analysis. *Mol Microbiol*, 99(4):767–77, 2016.

47. M. Schaechter, O. Maaloe, and N. O. Kjeldgaard. Dependency on medium and temperature of cell size and chemical composition during balanced grown of salmonella typhimurium. *J Gen Microbiol*, 19(3):592–606, 1958.

48. M. Scott and T. Hwa. Bacterial growth laws and their applications. *Curr Opin Biotechnol*, 22(4):559–65, 2011.

49. M. Scutari. Learning bayesian networks with the bnlearn r package. *J Stat Softw*, 35(3):1–22, 2010.

50. I. S. Seong, J. Y. Oh, J. W. Lee, K. Tanaka, and C. H. Chung. The hslu atpase acts as a molecular chaperone in prevention of aggregation of sula, an inhibitor of cell division in *Escherichia coli*. *FEBS Lett*, 477(3):224–9, 2000.

51. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

52. M. Slominska, A. Wahl, G. Wegrzyn, and K. Skarstad. Degradation of mutant initiator protein dnaa204 by proteases clpp, clpq and lon is prevented when dna is seqa-free. *Biochem J*, 370(Pt 3):867–71, 2003.

53. A. A. Smith, E. Belda, A. Viari, C. Medigue, and D. Vallenet. The canoe strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS Comput Biol*, 8(5):e1002540, 2012.

54. S. Taheri-Araghi. Self-consistent examination of donachie's constant initiation size at the single-cell level. *Front Microbiol*, 6:1349, 2015.

55. S. Taheri-Araghi, S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun. Cell-size control and homeostasis in bacteria. *Curr Biol*, 25(3):385–91, 2015.

56. Y. Tanouchi, A. Pai, H. Park, S. Huang, R. Stamatov, N. E. Buchler, and L. You. A noisy linear map underlies oscillations in cell size and gene expression in bacteria. *Nature*, 523(7560):357–60, 2015.

57. S. Vadia and P. A. Levin. Growth rate and cell size: a re-examination of the growth law. *Curr Opin Microbiol*, 24:96–103, 2015.

58. L. van der Maaten and G. Hinton. Visualizing data using t-sne. *J Mach Learn Res*, 9:2579–2605, 2008.

59. G. H. Van Domselaar, P. Stothard, S. Shrivastava, J. A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner, and D. S. Wishart. Basys: a web server for automated bacterial genome annotation. *Nucleic Acids Res*, 33(Web Server issue):W455–9, 2005.

60. L. D. Vukovic, P. Jevtic, L. J. Edens, and D. L. Levy. New insights into mechanisms and functions of nuclear size regulation. *Int Rev Cell Mol Biol*, 322:1–59, 2016.

61. M. Wallden, D. Fange, E. G. Lundius, O. Baltekin, and J. Elf. The synchronization of replication and division cycles in individual *E. coli* cells. *Cell*, 166(3):729–39, 2016.

62. X. Wang, Y. Kim, Q. Ma, S. H. Hong, K. Pokusaeva, J. M. Sturino, and T. K. Wood. Cryptic prophages help bacteria cope with adverse environments. *Nat Commun*, 1:147, 2010.

Figure legends

Figure 1. Experimental approach and reproducibility. **(A)** Experimental workflow. Single gene knock-out strains from the Keio collection were grown in M9 supplemented medium at 30°C in 96-well plates. DNA was stained with DAPI prior to imaging and 9 images were taken in both phase contrast and DAPI channels. The images were then processed with Oufti to identify the cell and nucleoid contours. In parallel, we recorded the growth curve of each imaged strain in order to extract growth parameters. **(B)** A SVM model was trained via visual scoring of 43,774 cells. **(C)** Confusion matrix of the SVM model based on a large validation dataset (102,137 cells), illustrating the distribution of the SVM classifier output in comparison to the visual classification. **(D)** Comparison of the average width of 192 strains obtained from two independent 96-well cultures of the 190 most remarkable Keio strains and 2 WT replicates.

Figure 2: Distribution of morphological, cell cycle and growth phenotypes in the *E. coli* Keio strain collection. Bubble graphs representing, for each feature, the number of strains with a score value, s, beyond 3, 4, 5 or 6 times the interquartile range away from the median of the WT distribution (240 replicates). The size of the circles (bubbles) reflects the number of strains with a score beyond a specific, color-coded threshold for s, as indicated. The 'reference' bubble graph illustrates the expectations from a dataset of the same size (4,227 strains), assuming a standardized normal distribution of scores (with a mean of 0 and a standard deviation of 1).

Figure 3: Feature-based COG enrichment analysis. Pie charts representing, on a feature-by-feature basis, the relative distribution of COG categories among the gene deletion strains associated with a severe phenotype as specified: **(A)** $s \geq 3$, **(B)** $s \leq -3$. The enriched COG categories are labeled and highlighted with an exploded pie sector. The under-represented COG categories are further highlighted by an asterisk. Enrichments and under-representations with an associated (FDR-corrected) q-value < 0.05 were considered significant. Only morphological and growth features with at least one enriched or under-represented COG category are represented.

Figure 4: The morpho archipelago. **(A)** Average 2D tSNE map of the 797 strains with at least one morphological feature with a $|s| > 3$, plus the 240 independent WT replicates used as controls. Color-coded islands resulting from the dbscan algorithm ($\epsilon = 7$, minPts $= 3$) were defined by groups of strains clustering together with the same dbscan parameters in more than 90% of the generated tSNE maps. **(B)** Heatmap showing, for each island, the average score of each morphological and growth feature used for the construction of the map. **(C)** Network representation of island 17 grouping filamentous mutants. The weights were directly derived from the average distances between corresponding mutant strains in the 2D tSNE maps. **(D)** Internal structure of the mean and CV of length in island 2. The average gradients of phenotypes over the area of the island 2 are represented with arrows. **(E)** Phenoprints associated with each enriched GO term represented as a clustergram. Both rows and columns were ordered using a hierarchical clustering algorithm based on Euclidean distances. Enriched GO terms associated with an FDR-corrected q-value below 0.05 are shown.

Figure 5: The cell cycle archipelago. **(A)** Stable islands in the cell cycle archipelago. The cell cycle and growth phenoprints were used to map the 264 mutant strains with at least one cell cycle or growth feature with a $|s| > 3$, as well as the 240 independent WT replicates, in 2D using tSNE. As for the morpho archipelago, the data were clustered using the dbscan algorithm. The groups of strains clustering together in more than 90% of the maps defined an island. **(B)** Heatmap showing, for each island, the average score of each cell cycle and growth feature used for the construction of the map. **(C)** The islands were colored according to the average score for two growth features and two cell cycle features. **(D)** Violin plots illustrating the distribution of scores for all the strains included in islands WT, 1, 4, 7 and 10 for the two growth features and three cell cycle features. The black dot and bar show the mean and standard deviation, respectively. **(E)** Phenoprints associated with each enriched GO term represented as a clustergram. Both rows and columns were ordered using a hierarchical clustering based on Euclidean distances. Enrichments associated with an FDR-corrected p-value below 0.05 were considered significant. **(F)** The islands in the cell cycle archipelago

were colored according to the proportion of 'y-genes' (genes of unknown function). The color map scales from white (0%) to black (60%).

Figure 6: Nucleoid separation and cell constriction dynamics. **(A)** Scatter plot of the relative timing of cell constriction versus the relative timing of nucleoid separation. The gray scale indicates the density of dots in a given area of the chart. The dotted contours represent the 0.5, 0.75 and 0.95 probability contours of the 240 WT replicates. The Pearson correlation is $\rho = 0.65$. The black dotted diagonal represents the line where a dot should be if both nucleoid separation and cell constriction had happened at the same time. Red dots highlight strains shown in panels **(B)**, **(C)** and **(D)**. **(B)** Average dynamics of nucleoid separation and cell constriction for WT and the $\Delta matP$ mutant strain. The cumulative distributions of the fraction of cells with two nucleoids (blue) and of the fraction of cells with a constriction degree above 0.15 (red) were plotted against the cell length percentile. **(C)** Same plots as in **(B)** for three strains clustering in island 4 with the $\Delta matP$ strain. The WT curves were plotted in gray for comparison. **(D)** Same plots as in **(B)** for the $\Delta hslV$ strain and two island 7 strains, $\Delta ybaN$ and $\Delta hslU$.

Figure 7: Cell length regulation mutants. **(A)** Scatter plot of the mean cell length versus the CV of the length for all the strains. The gray color levels indicate the density of points in the vicinity of each strain. The orange dots and error bars represent the mean and standard error of the mean per bin. The black line is a piecewise linear fit with a single cross-over point to the binned data (orange) to highlight the two global regimes in the relationship between mean cell length and CV of length. We performed the fit with a (bi-square weighing) linear least square algorithm. **(B)** Cumulative distribution of the proportion of constricting cells for the *ftsA** mutant and its parent. **(C)** Scatter plot of the CV of the cell length for the whole population versus the CV of the cell length for constricting cells only. The contour lines represent the 0.5, 075 and 0.95 probability envelope of the 240 independent WT replicates. The gray color levels indicate the density of points in the vicinity of each strain. The $\Delta mraZ$ strain discussed in the text is highlighted in red.

Figure 8: Inter-dependence of cell morphology and cell cycle progression. **(A)** Network showing the functional relationship between 16 non-collinear morphological, growth and cell cycle features. The network is an undirected network highlighting the most informative connections detected by the ARACNE algorithm. The thickness of an edge represents the fraction of the networks containing this specific edge after bootstrapping the network 200 times, from 70% (thinnest) to 100% (thickest). **(B)** Scatter plot of the normalized mean cell length and mean cell width of all 4,227 Keio strains and 240 WT replicates. Each dot represents a strain, and the gray level illustrates the density of neighbors in the vicinity of each point in the graph. The dotted contours represent the 0.5, 0.75 and 0.95 probability envelopes of the 240 WT replicates. **(C)** Heatmap showing the mean growth rate value for data binned by both mean cell area and mean nucleoid area. The cell and nucleoid areas are strongly correlated ($\rho = 0.83$). The median value of α_{max} per bin is color-coded according to the color scale. **(D)** Scatter plot of the mean cell area versus the mean nucleoid area for cells with 1, 2, 3 or $\geq$4 nucleoids for each strain. The histogram in the inset illustrates the average proportions of cells with 1, 2, 3 or $\geq$4 nucleoids per strain. Although there are typically few cells in each strain with 3 or $\geq$4 nucleoids, at least one cell with $\geq$3 nucleoids was detected for 61% of the strains. **(E)** Heatmap showing the mean growth rate value for data binned by both the mean nucleoid area and the relative timing of nucleoid separation. The mean nucleoid area negatively correlates with the relative timing of nucleoid separation (ρ = -0.49). **(F)** Heatmap showing the mean cell area value for data binned by both the relative timing of cell constriction and the relative timing of nucleoid separation. The relative timing of cell constriction is strongly correlated to the relative timing of nucleoid separation ($\rho = 0.65$).

Expanded view figure legends

Figure EV1. Morpho archipelago. A typical tSNE map based on 19 morphological and 2 growth features. The same color code as in Figure 4A was used to represent the islands derived from the dbscan clustering. The light grey dots represent strains that were not consistently (less 90% of the time) associated with one of the island.

Figure EV2. Filamentous mutants. **(A)** Representative phase-contrast images of the 8 mutants forming the island 17, together with the parental strain BW25113 (WT) for comparison. **(B)** Effect of the kanamycin-resistance cassette on the phenotype of the Δ*croE* strain. The schematic at the top shows the color-mapped score of mean cell length for the deletion of each gene of the *croE* operon. Below are phase-contrast and fluorescent images of DAPI-stained cells of the Δ*croE* strain carrying the kanamycin resistance cassette (top) or after the removal of the cassette (bottom). **(C)** Effect of the kanamycin-resistance cassette on the phenotype of the Δ*ydaS* strain. The schematic at the top shows the color-mapped score of the CV of cell length for each gene of the *ydaS* operon. Below are phase-contrast and fluorescent images of DAPI-stained cells of the Δ*ydaS* strain carrying the kanamycin resistance cassette (top) or after the removal of the cassette (bottom).

Figure EV3. Specific pathways associated with impaired cell morphology. **(A)** Schematic of the ECA biosynthetic pathway in which each gene name has been colored by the severity of the mean aspect ratio ($<Ar>$) phenotype. **(B)** Scatter plot of cell width versus cell length for three independent cultures of the Δ*rapZ* strain (n = 564, 268 and 343 cells). The dotted lines represent isocontours of a 2D histogram of cell length and cell width for the parental strain (WT, n = 1,045 cells). The cell width distributions of the WT and Δ*rapZ* strains are represented on the right of the scatter plot (all three replicates for the Δ*rapZ* strain were pooled together). **(C)** Schematics of the high-affinity ABC phosphate transporter and the ATP synthase, in which the subunits have been colored according to the severity of the mean cell width ($<W>$) phenotype in the corresponding gene deletion strains.

Figure EV4. Simulation showing that the cell length variability of the entire population can mask abnormal cell length variability at a specific cell cycle period. Cell length distributions were simulated over different ranges of cell ages (see Materials and methods). The cell length distribution of constricting cells was determined by summing the cell length distributions of all cells of age > 0.8, assuming different CV of the cell length distribution (0.05, 0.11 and 0.2) at a specific age. The cell length distribution of the whole population was determined by summing the distributions at all ages, from birth to division.

Figure EV5. Growth rate correlates poorly with morphological features, cell constriction and nucleoid dynamics. **(A)** Scatter plots showing the growth rate of each Keio strains and WT replicates relative to their mean cell or nucleoid area. Each dot represents a strain, and the gray level illustrates the density of neighbors in the vicinity of each strain. The dotted contours reflect the 0.5, 0.75 and 0.95 probability envelopes of the 240 WT replicates. The corresponding τ correlation coefficients are indicated on each graph. **(B)** Heatmap showing the mean growth rate value for data binned by the relative timings of cell constriction and nucleoid separation. **(C)** Same as **(B)**, except for data binned by mean cell length and mean cell width.

Supplementary figure legends

Figure S1. Feature determination and SVM model validation. **(A)** Typical growth curve represented as the log_2 of OD_{600nm} as a function of time. The dotted line shows the linear fit to the segment of maximal growth. The last hour of growth (gray box) was used to estimate the saturation level of the culture. **(B)** AUROC curve (performance curve) related to the SVM model on the dataset that was not used to train the model. The dotted line $y = x$ represents the expectation from a random classification. **(C)** Ratios between the mean values of each predictor for the visually-curated and SVM-curated datasets, showing the lack of bias. **(D)** Cumulative distributions of mean score ratios between visually- and SVM-curated datasets of mean and CV predictor values for the 419 strains with the most extreme phenotypes. The steepness of the curves shows that the SVM model performed well, even for strains with strong phenotypic defects. **(E)** Plot showing how two cell cycle features, the correlation between the degrees of constriction of the nucleoid and of the cell (ρ CD) and the projected degree of nucleoid constriction at the onset of cell constriction (CDN_{C0}), were calculated using a WT culture as an example. The degree of constriction for both the nucleoid and the cell (considering only cells with a degree of cell constriction over 15%) were used to calculate their Pearson correlation coefficient (ρ CD). The correlation coefficient can be interpreted as the slope of the line passing through the data, and the intercept of this line with the y-axis provides the average degree of constriction of the nucleoid at the onset of cell constriction (CDN_{C0}).

Figure S2. Evaluation of positional and temporal biases related to imaging. **(A)** Plate-by-plate normalization. Each plate is color-coded according to the 96-well plate number. The black dots represent the mean feature value per plate, ± standard deviation. **(B)** Scatter plots for each feature of all strains as a function of the time elapsed from spotting cells on the pad for imaging. The red line in each graph represents the smoothing spline (calculated with a span of 10%) that was used to correct any temporal bias.

Figure S3. Temporal bias correction for proportional features. **(A)** Scatter plots of the measured relative proportions of cells committed to division or not (blue and red dots, respectively) as a function of the time spent on the pad prior to imaging. The solid black lines represent the correction factors over time, derived from a quadratic form Dirichlet regression to the data. The Dirichlet regression allows for the maintenance of the additivity of the proportion [38]. **(B)** Same as in **(A)** for the complementary proportions of cells with 1 (blue), 2 (red) or more than 2 (yellow) nucleoids. The fitted model was quadratic for the first two features (1 and 2 nucleoids) and linear for the cells with more than 2 nucleoids.

Figure S4. Feature-based COG distribution analysis. Pie charts representing, on a feature-by-feature basis, the relative distribution of COG categories among the gene deletion strains associated with a severe phenotype: **(A)** $s \geq 3$, **(B)** $s \leq -3$. All the features that were not included in Figure 3 are represented. The enriched COG categories are highlighted with an exploded pie sector. Enrichments with an associated (FDR corrected) q-value < 0.05 were considered significant.

Table S1

Morphology / size features	
Feature name	**Symbol**
Mean cell length	$<L>$
Mean cell width	$<W>$
Mean cell area	$<A>$
Mean cell volume	$<V>$
Mean cell surface area	$<SA>$
Mean cell perimeter	$<P>$
Mean cell circularity	$<C>$
Mean cell aspect ratio	$<Ar>$
Mean cell surface area to volume ratio	$<SA/V>$
Cell length variability	CV_L
Cell width variability	CV_W
Cell area variability	CV_A
Cell volume variability	CV_V
Cell surface area variability	CV_{SA}
Cell perimeter variability	CV_P
Cell circularity variability	CV_C
Cell aspect ratio variability	CV_{Ar}
Cell surface to volume ratio variability	CV_{SA}
Division ratio variability	CV_{DR}
Growth features	
Feature name	**Symbol**
Max. growth rate	α_{max}
Optical density at growth saturation	OD_{max}
Cell cycle features	
Feature name	**Symbol**
Correlation in nucleoid and cell constriction	ρCD
Nucleoid constriction degree at the initiation of cell constriction	CDN_{C0}
Relative timing of cell constriction	Rel.timing div
Relative timing of nucleoid separation	Rel.timing nuc
Fraction of cells with 2 nucleoids	$\% \geq 2N$

Table S1. Features considered in this study and their associated symbols. The aspect ratio was defined as the ratio of cell width over cell length at the single-cell level. The circularity, C, was defined as $C = 4\text{pi}A/P^2$, at the single-cell level, where P stands for perimeter and A for area. The relative timing of cell constriction and nucleoid separation were estimated as the proportions of cells without any significant constriction (constriction degree < 0.15) or with a single nucleoid, respectively. For all cells with a significant constriction degree, we calculated the Pearson correlation coefficient between the constriction degrees of the cell and of its nucleoid (ρCD. The nucleoid constriction degree at the initiation of cell constriction (CDN_{C0}) was determined as the intercept of a line with a slope determined by the correlation coefficient that best fitted the single-cell data used to calculate ρCD (see Appendix Fig S1E).

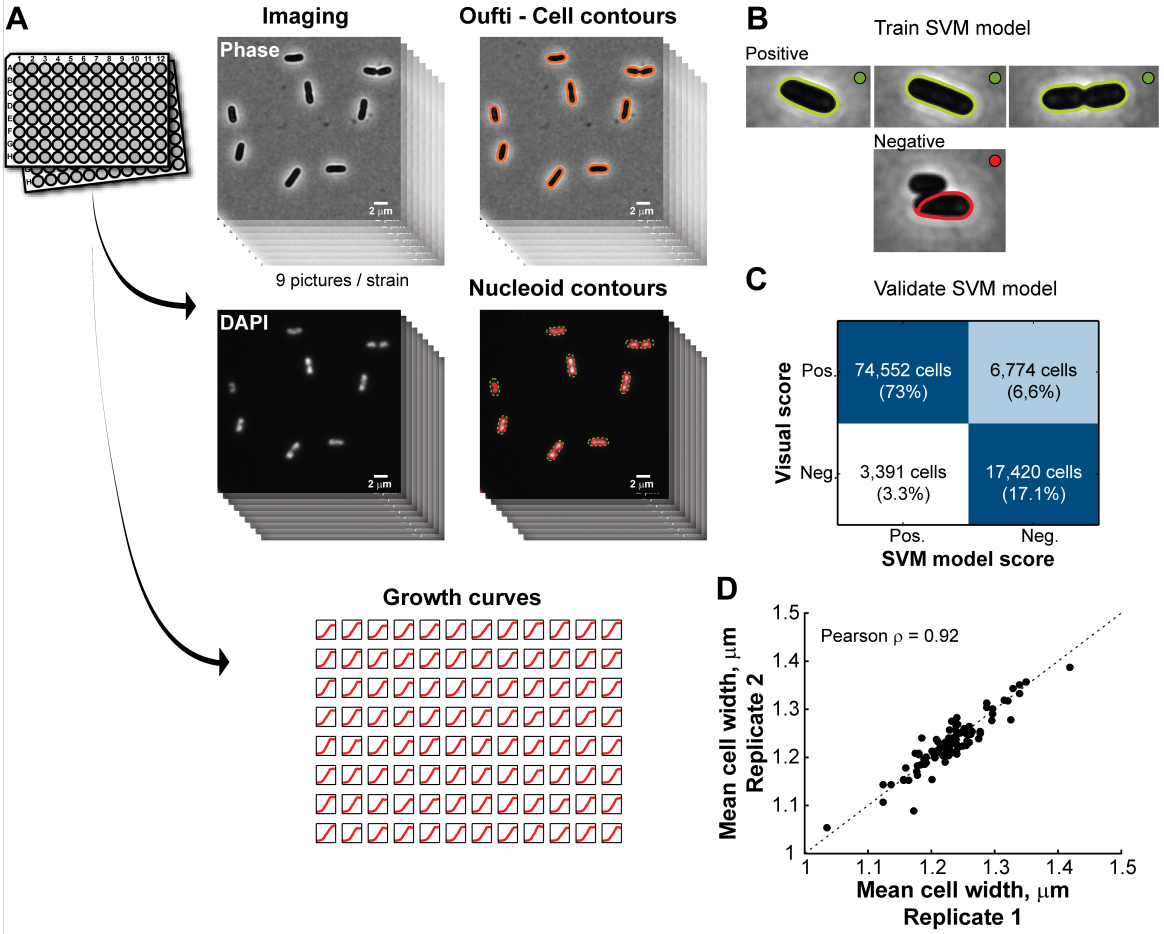

Figure 1

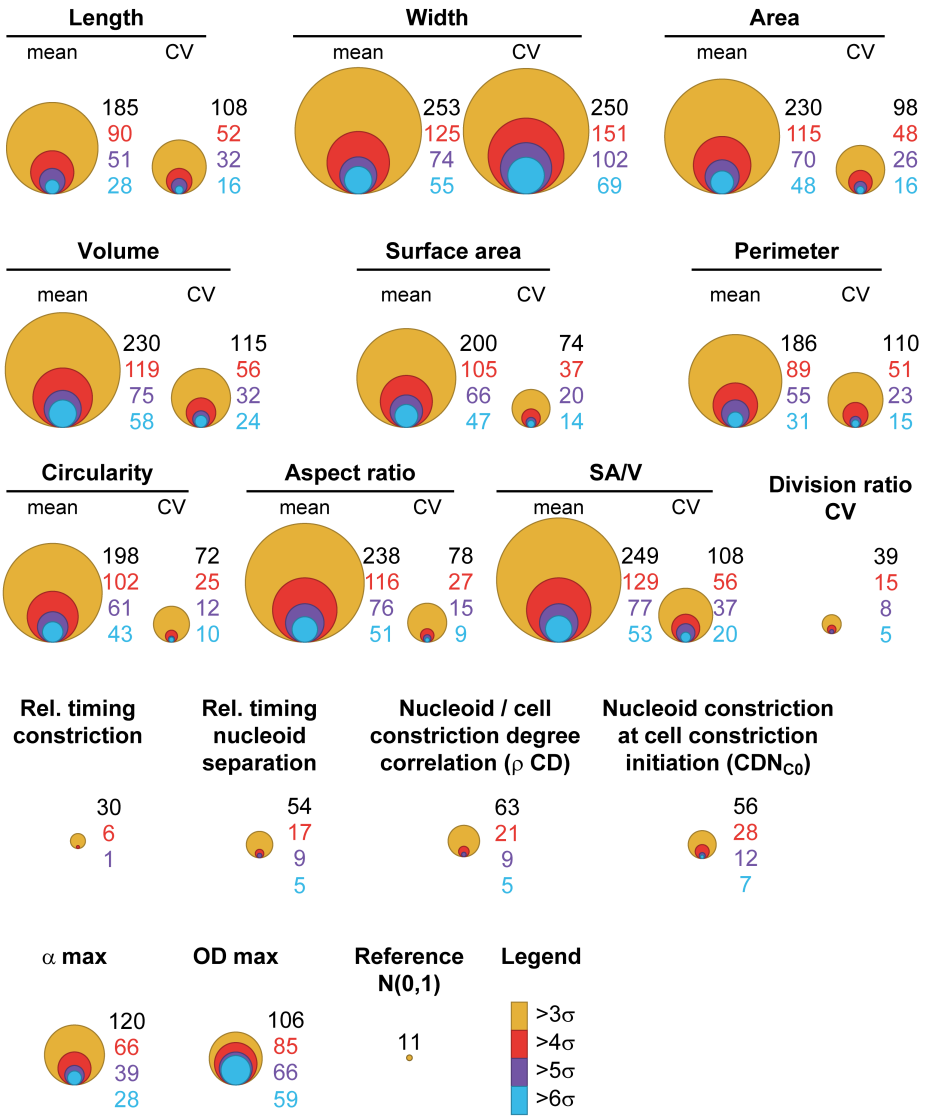

Figure 2

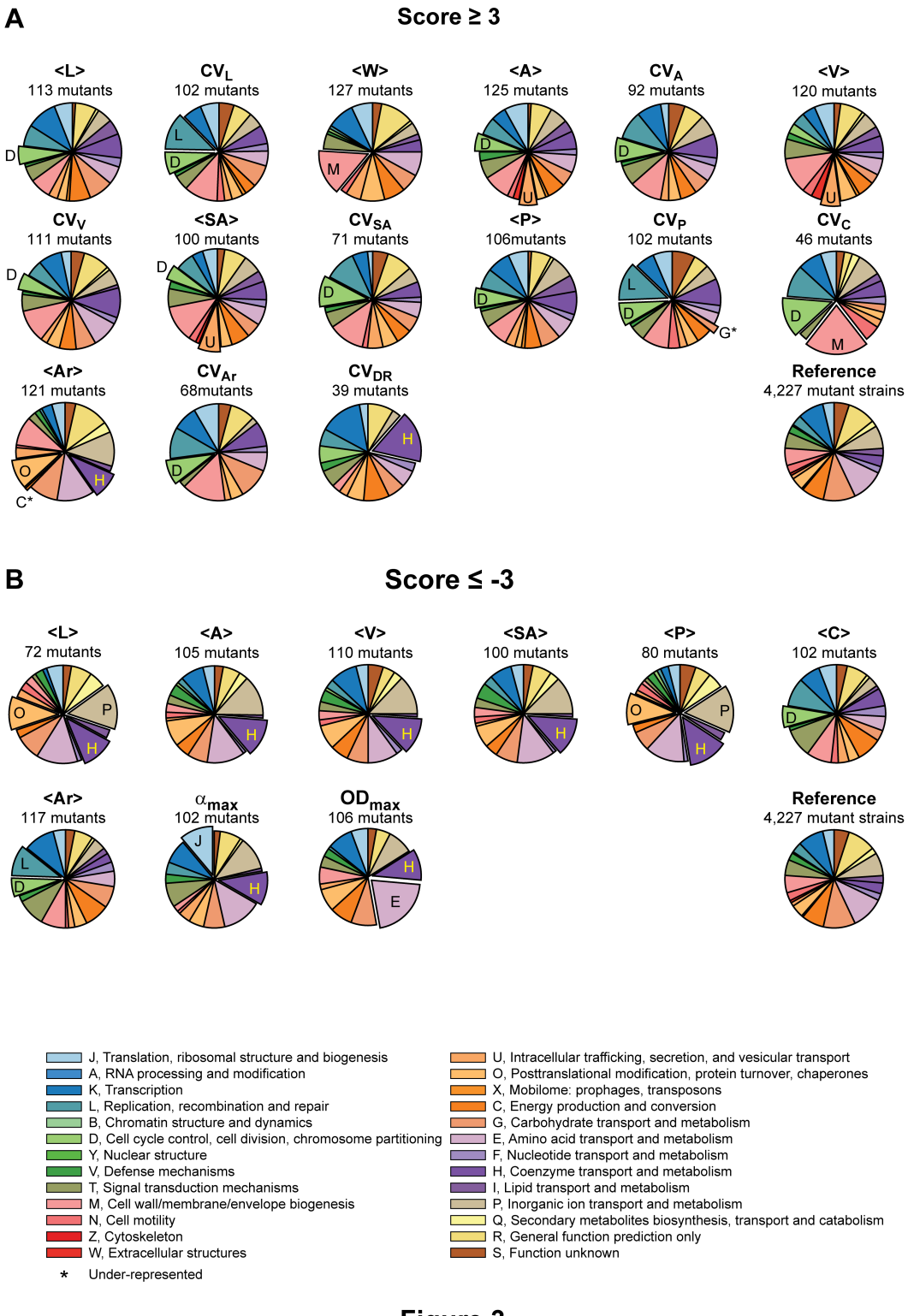

Figure 3

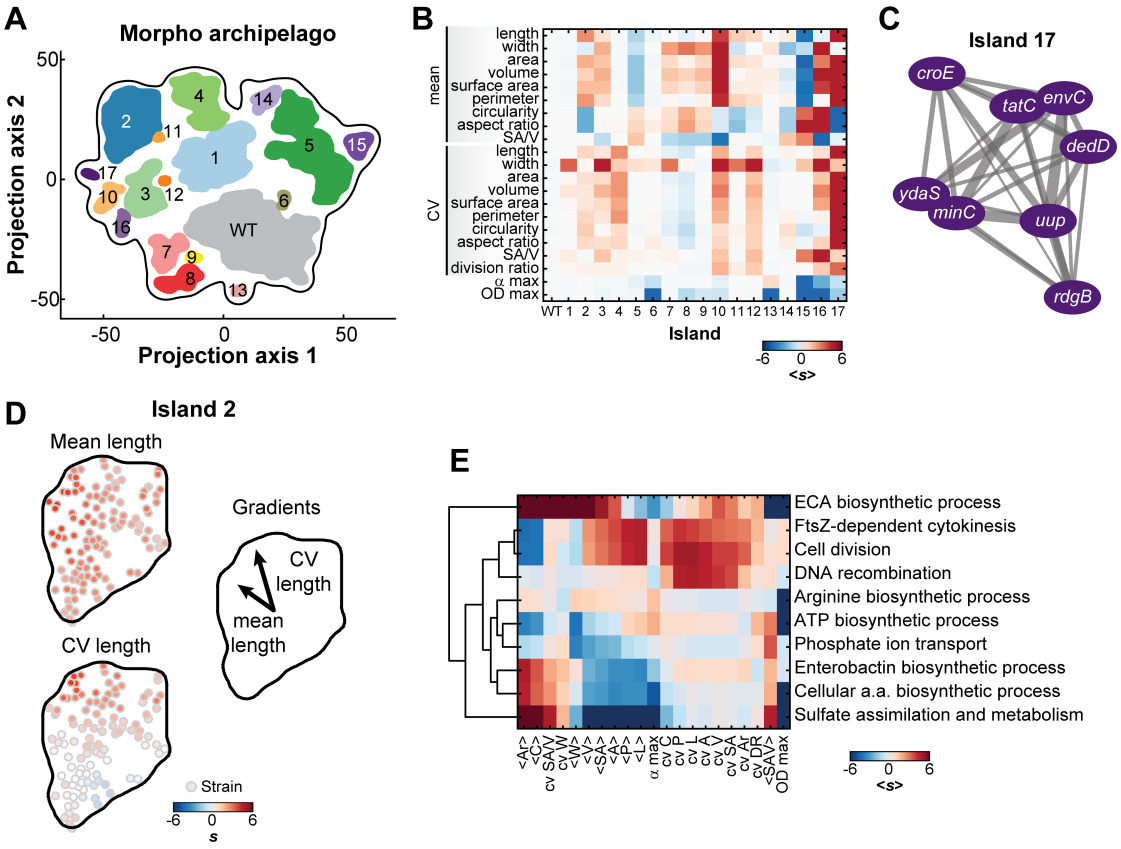

Figure 4

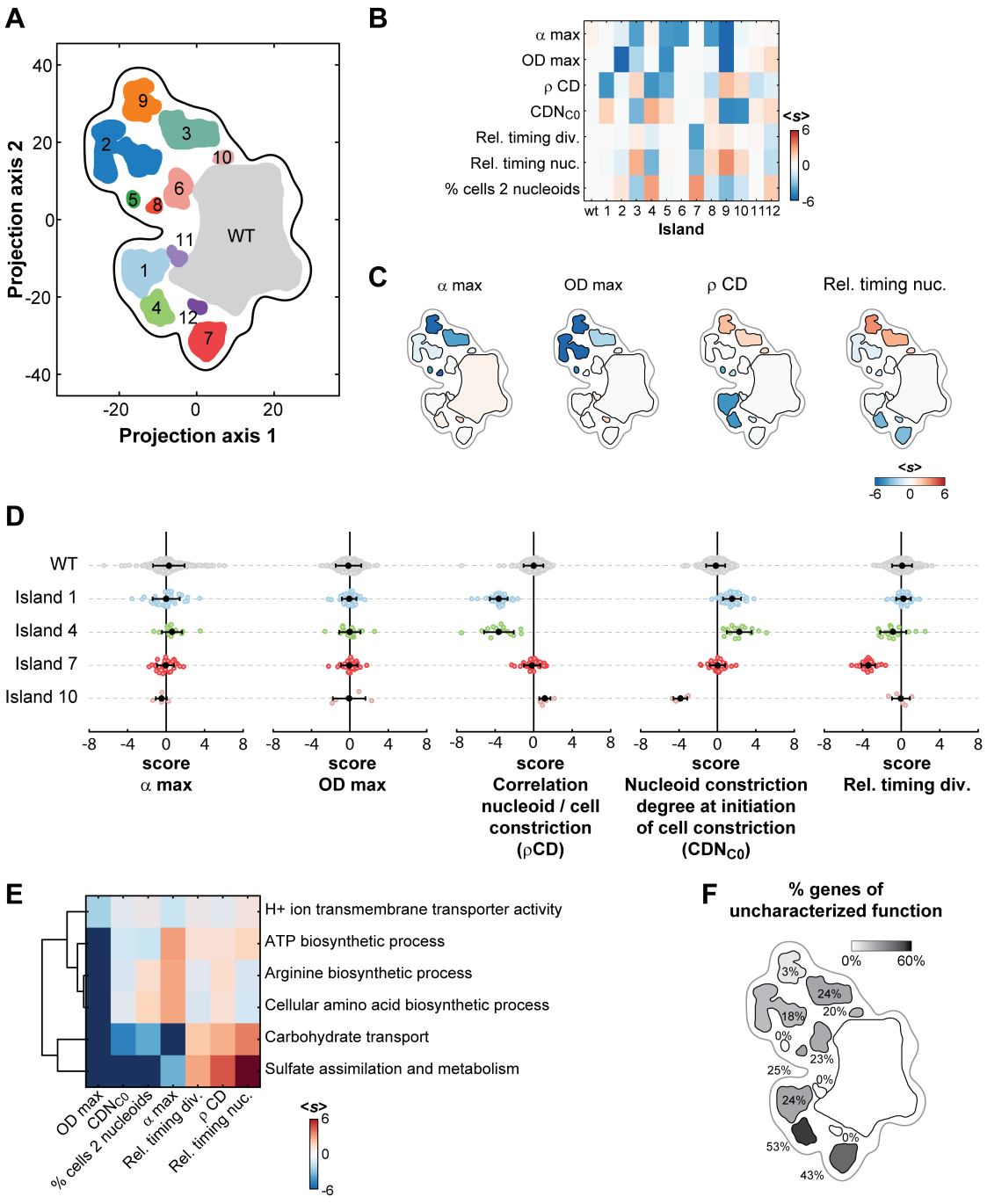

Figure 5

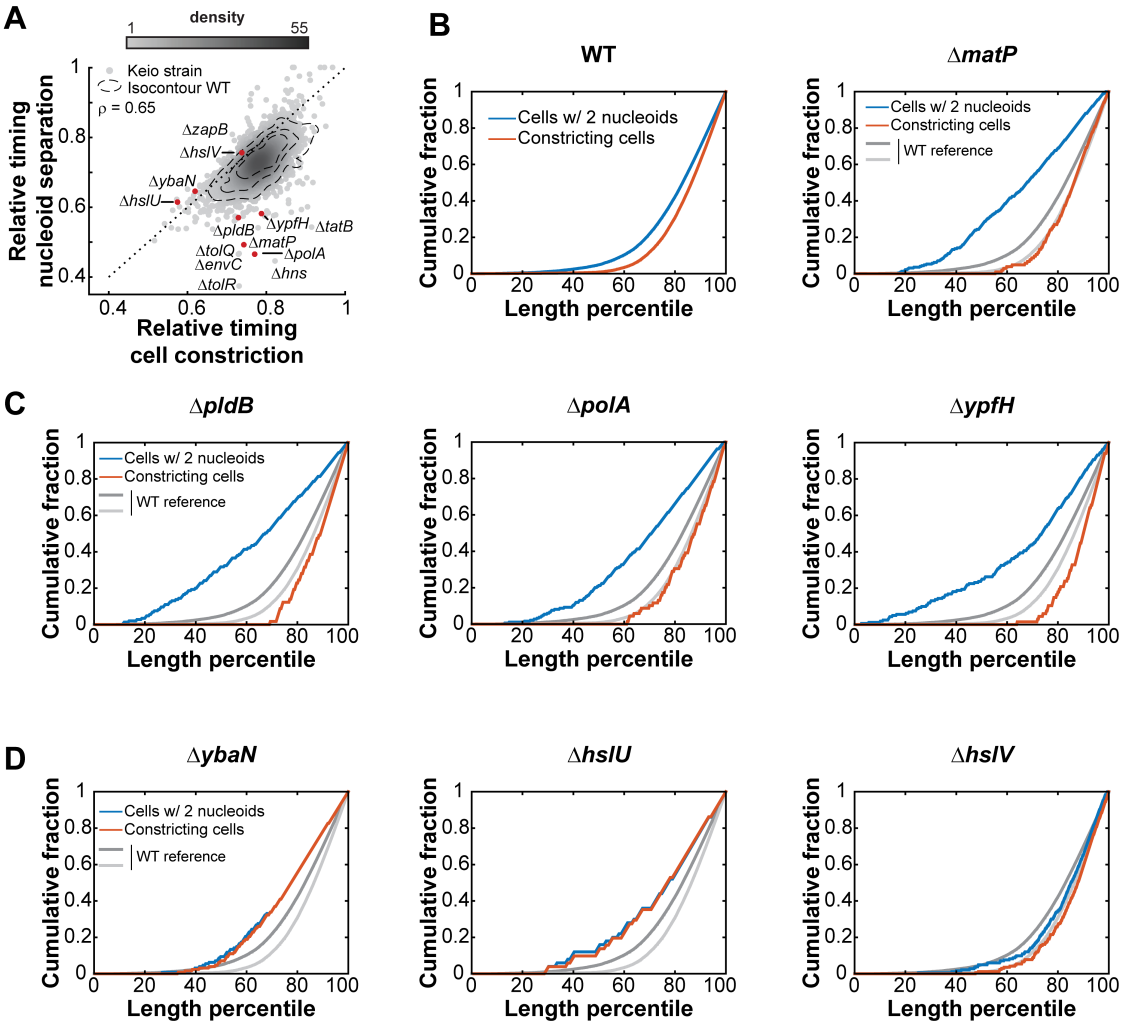

Figure 6

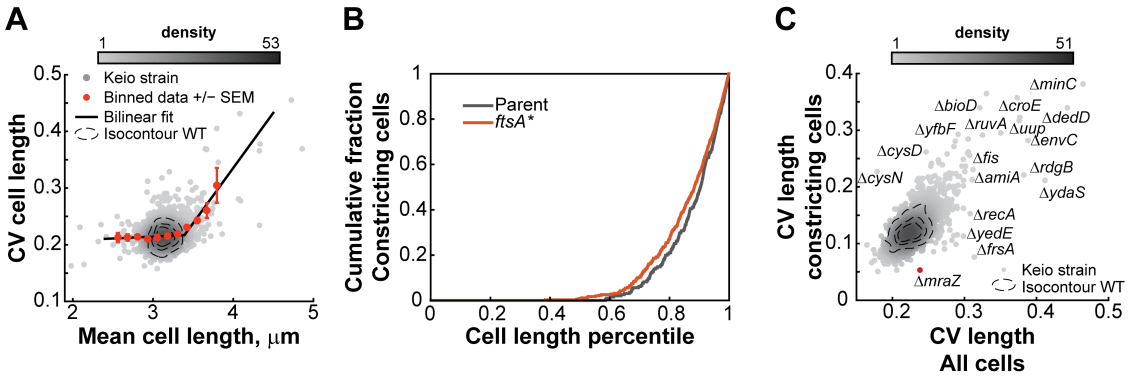

Figure 7

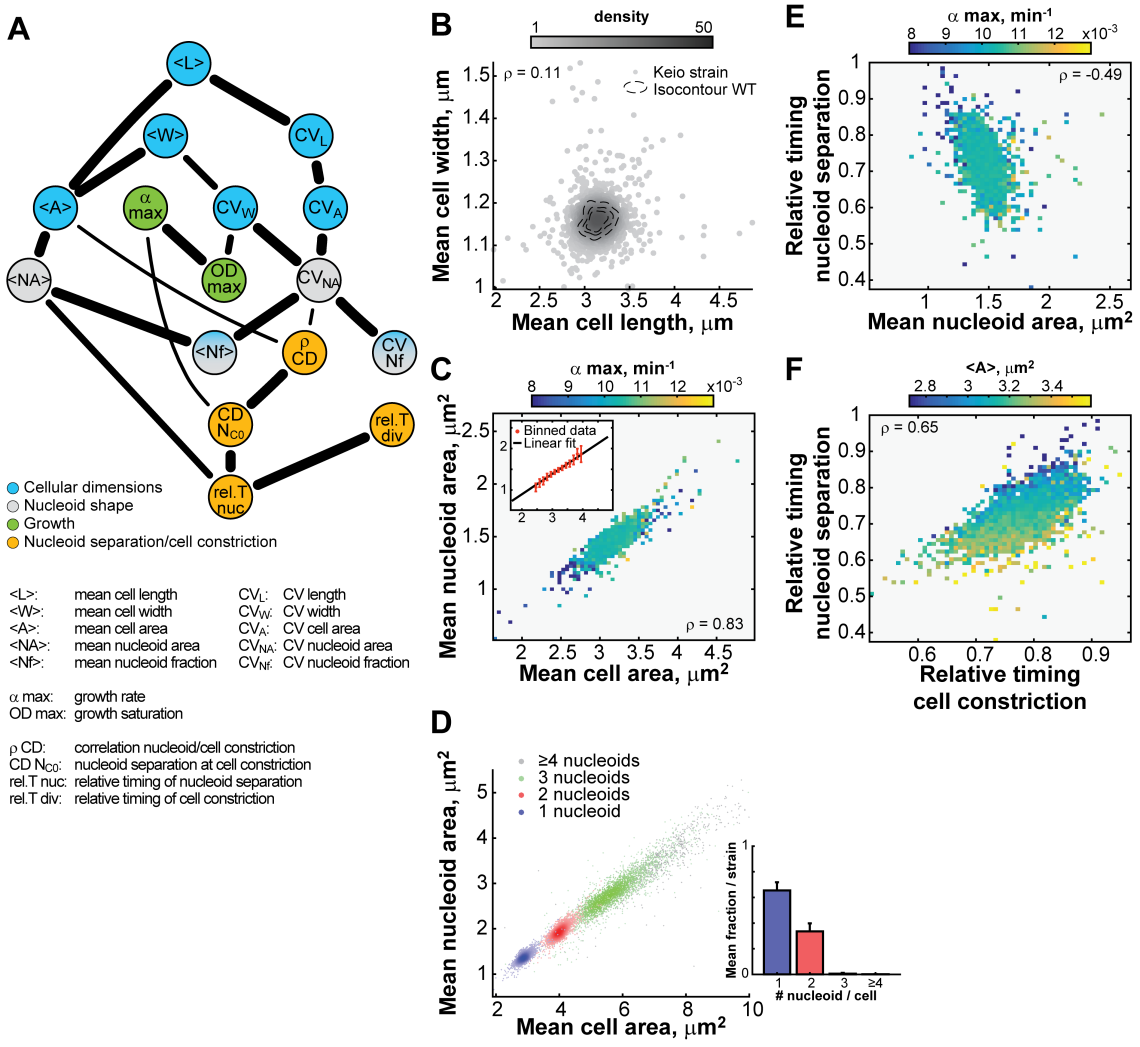

Figure 8

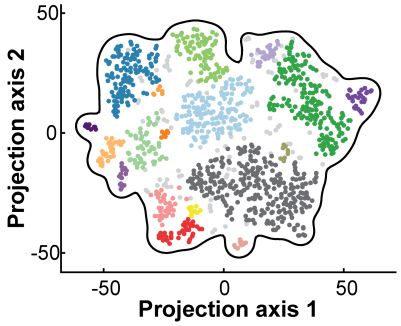

Figure EV1

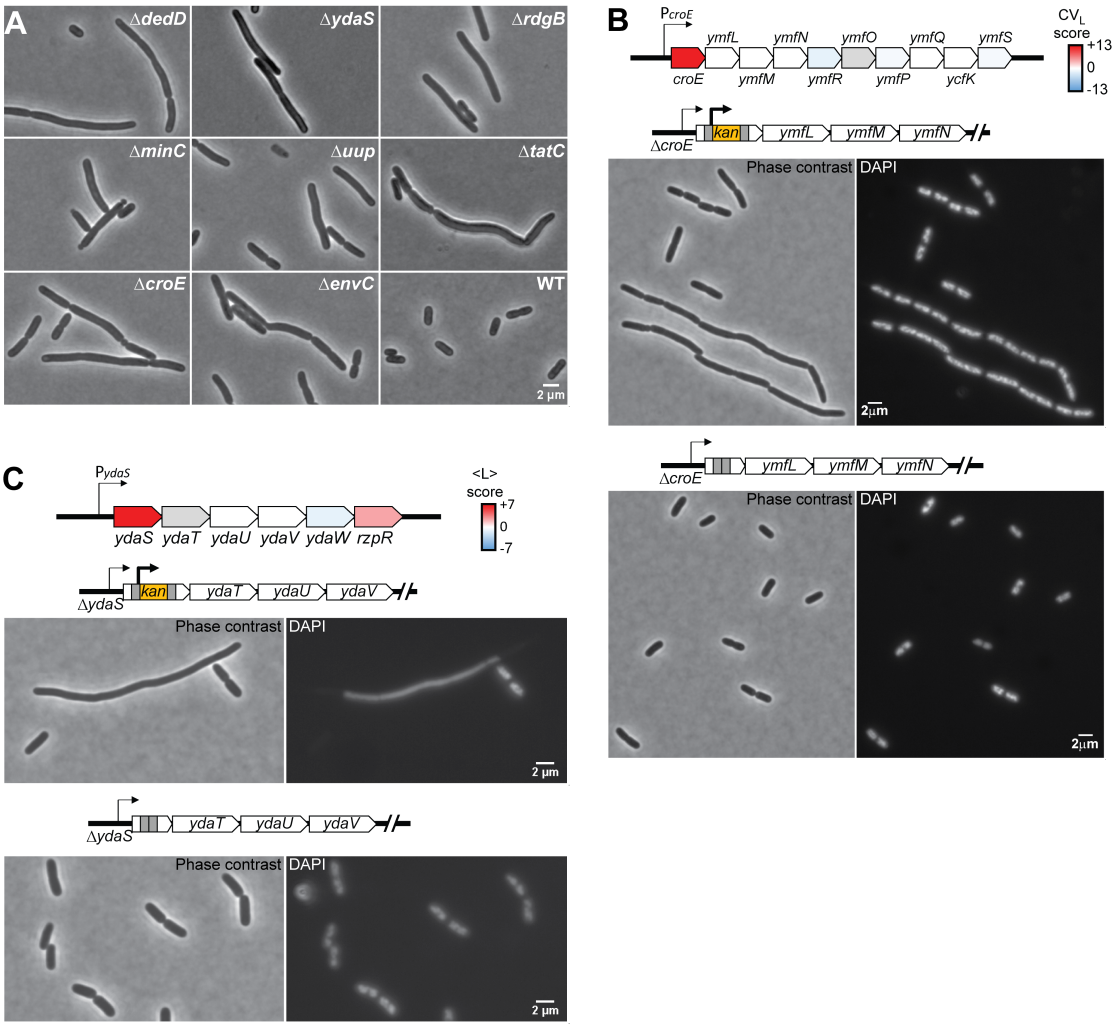

Figure EV2

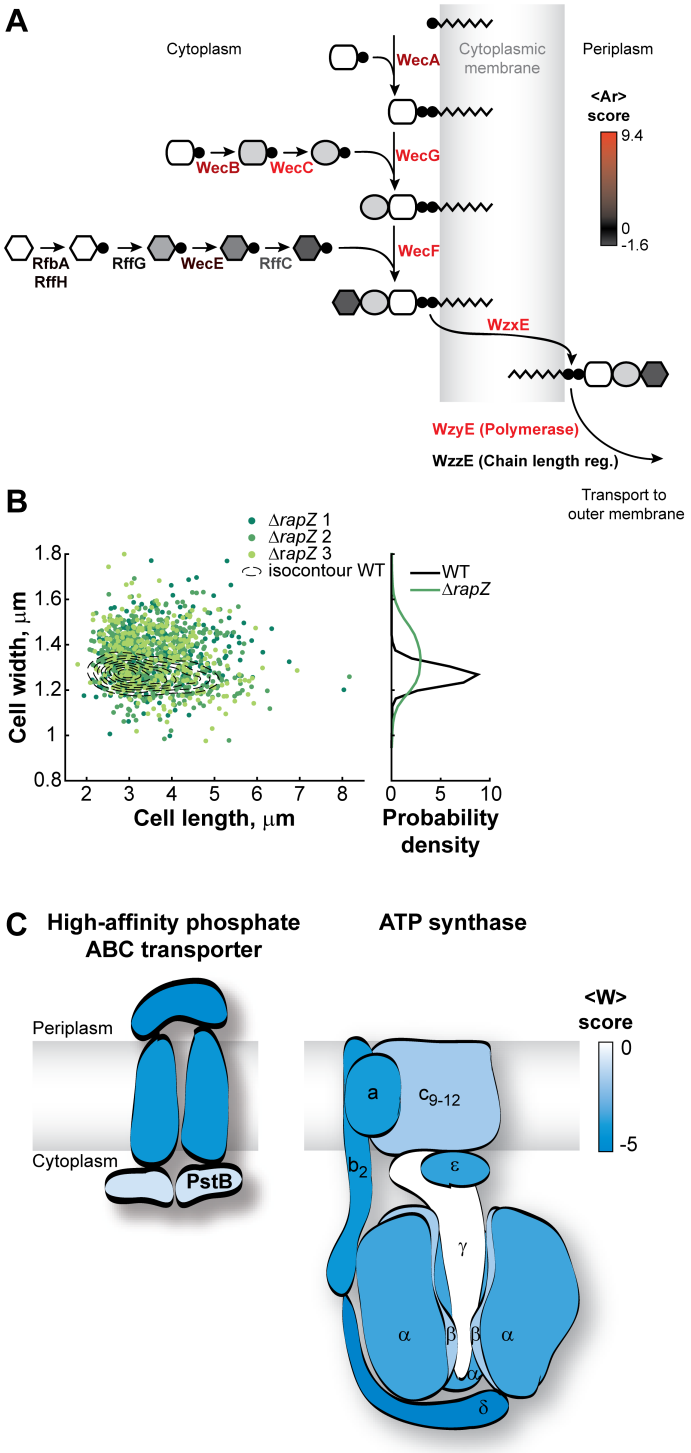

Figure EV3

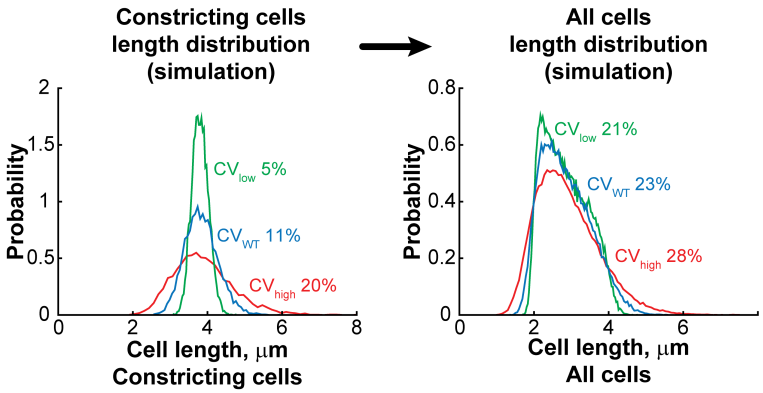

Figure EV4

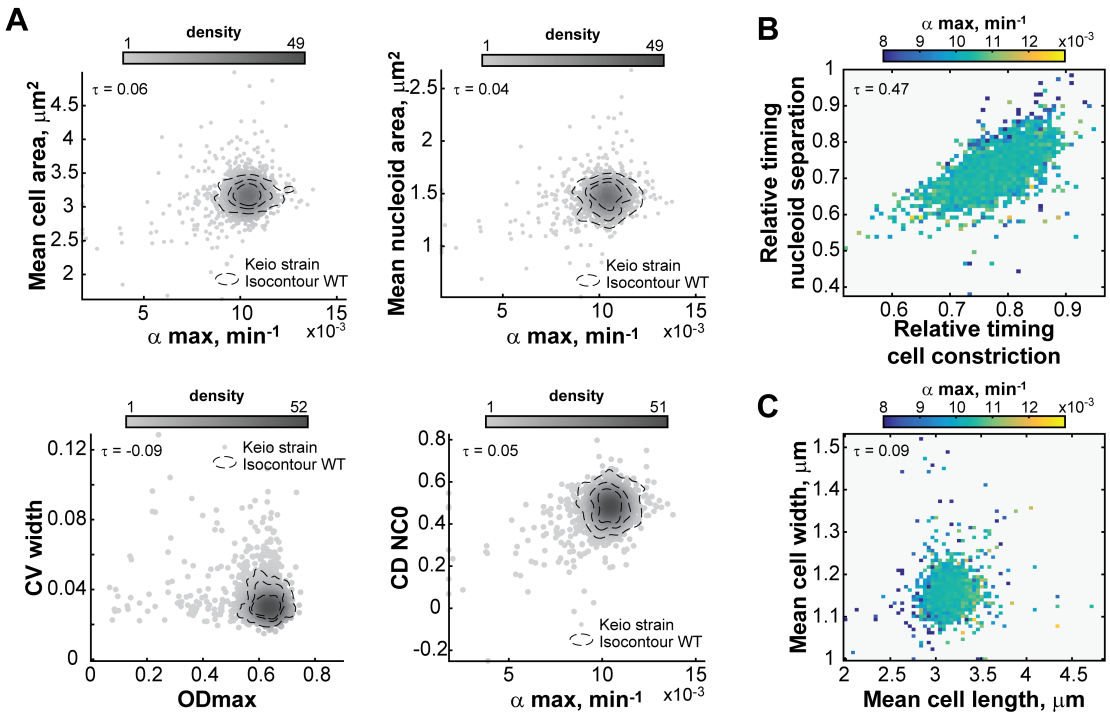

Figure EV5

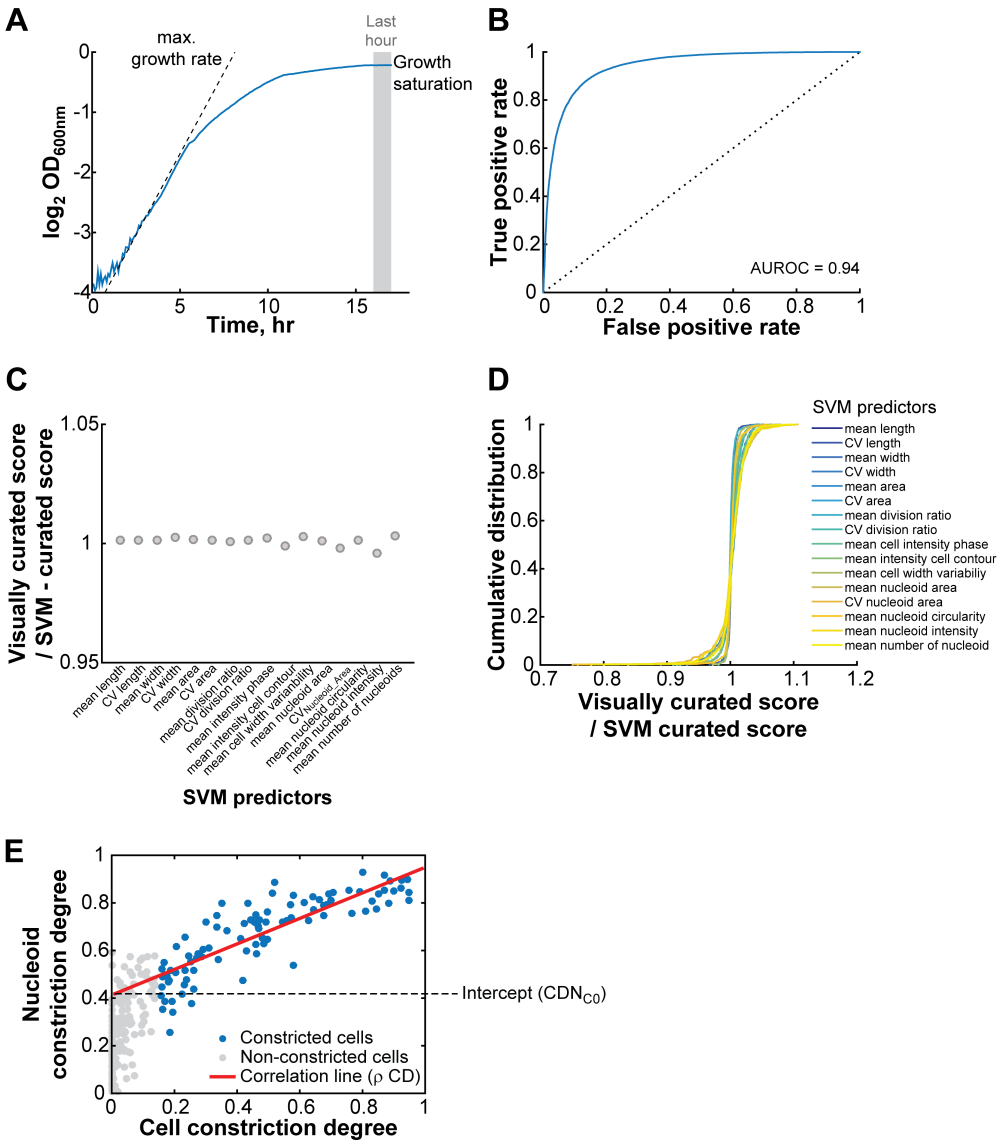

Figure S1

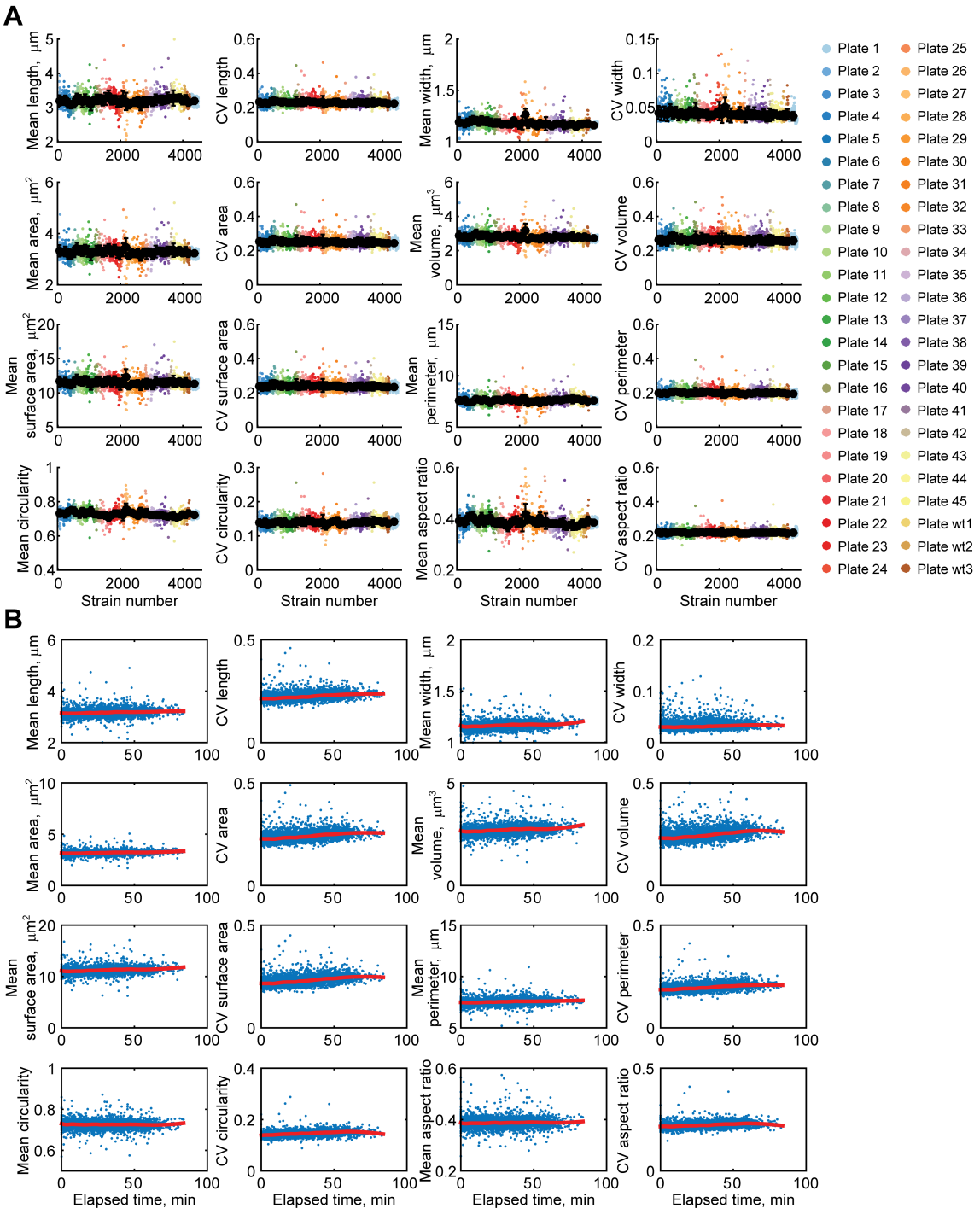

Figure S2

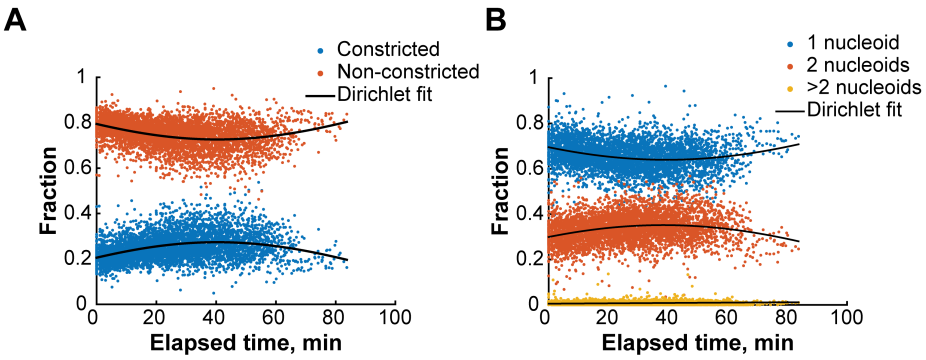

Figure S3

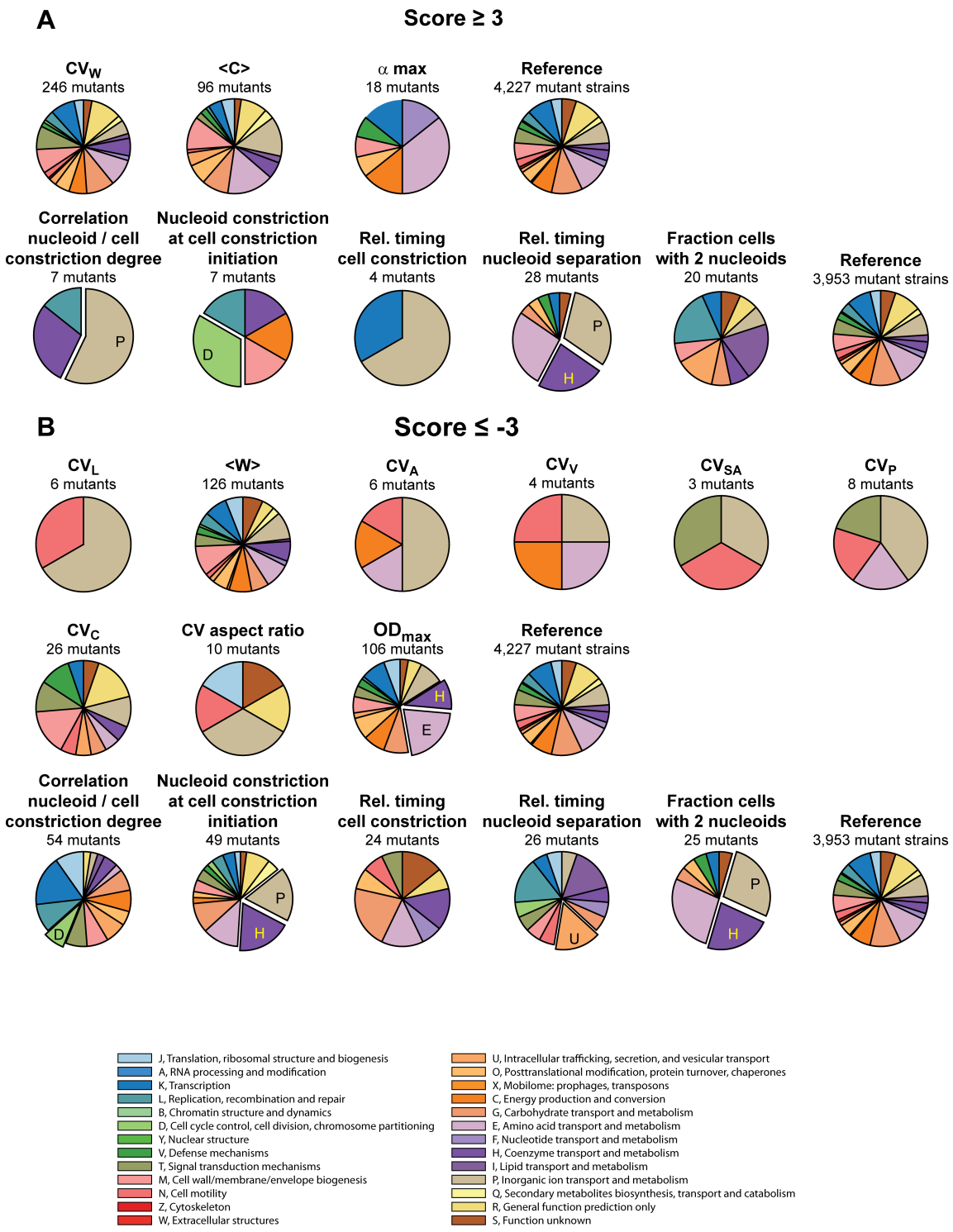

Figure S4