

# Selection of a de novo gene that can promote survival of *Escherichia coli* by modulating protein homeostasis pathways

Received: 11 February 2023

Accepted: 12 September 2023

Published online: 9 November 2023

 Check for updates

Idan Frumkin<sup>1</sup> & Michael T. Laub<sup>1,2</sup>  

Cellular novelty can emerge when non-functional loci become functional genes in a process termed de novo gene birth. But how proteins with random amino acid sequences beneficially integrate into existing cellular pathways remains poorly understood. We screened  $\sim 10^8$  genes, generated from random nucleotide sequences and devoid of homology to natural genes, for their ability to rescue growth arrest of *Escherichia coli* cells producing the ribonuclease toxin MazF. We identified  $\sim 2,000$  genes that could promote growth, probably by reducing transcription from the promoter driving toxin expression. Additionally, one random protein, named Random antitoxin of MazF (RamF), modulated protein homeostasis by interacting with chaperones, leading to MazF proteolysis and a consequent loss of its toxicity. Finally, we demonstrate that random proteins can improve during evolution by identifying beneficial mutations that turned RamF into a more efficient inhibitor. Our work provides a mechanistic basis for how de novo gene birth can produce functional proteins that effectively benefit cells evolving under stress.

A central premise in molecular evolution is that organisms must innovate to survive changing environments. Cellular novelty usually emerges via mutations to existing genes or by mixing-and-matching protein domains<sup>1</sup>. However, evolution may also invent new, functional proteins from scratch, a process termed de novo gene birth<sup>2,3</sup>. Little is known about how often this process occurs and, when it does, how such new proteins provide a benefit to cells<sup>4</sup>.


Recent studies have used comparative genomics and synteny-based methods to identify lineage-specific genes that may represent de novo genes<sup>5–9</sup>. However, the designation of lineage-specific genes as de novo genes suffers from high false discovery rates due to homology detection failure<sup>10,11</sup>. For bona fide cases of de novo genes, some biological effects have been reported<sup>12</sup> but whether they have beneficial functions that confer a selective advantage remains unknown in most cases.

How can a given nucleotide sequence become a gene? The ‘proto-gene’ model for de novo gene birth<sup>13</sup> sets two main requirements: (1) stable expression of a locus and (2) beneficial function of the

emerging gene product. If these conditions are met, natural selection can further improve expression, function and regulation to generate a mature gene integrated into cellular physiology. RNA sequencing and ribosome profiling studies have revealed extensive spurious transcription and translation in species across the tree of life<sup>7,13–17</sup>. These loci could serve as a source for new genes.

A complementary approach to investigating de novo gene birth involves characterizing randomly generated proteins and studying whether they can benefit cells. Natural de novo genes do not necessarily come from purely random sequences because various evolutionary forces shape and bias genomes<sup>18–20</sup>. Nevertheless, finding and characterizing functional proteins with random amino acid sequences can provide a missing rationale for the place of de novo proteins in evolution. Previous work has examined in silico and in vitro properties of such random sequences, including their predicted ability to fold into secondary structures<sup>21</sup>, chaperones-assisted solubility<sup>22</sup>, ATPase activity<sup>23</sup> and potential affinity for different molecules<sup>24–27</sup>.

<sup>1</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>2</sup>Howard Hughes Medical Institute, Cambridge, MA, USA.

 e-mail: [laub@mit.edu](mailto:laub@mit.edu)

However, cellular functions for random genes have rarely been demonstrated *in vivo*. One recent study reported that random proteins in *Escherichia coli* can have beneficial effects on growth<sup>28</sup> but serious caveats in experimental design were subsequently raised<sup>29,30</sup>. Two recent studies found hydrophobic proteins that provide antibiotic resistance to *E. coli* cells<sup>31,32</sup> by membrane depolarization and stimulation of a membrane-bound histidine kinase. Additional studies identified small random proteins that rescue an *E. coli* auxotroph<sup>33,34</sup>, probably by binding to the 5' untranslated region (UTR) of the *his* operon to increase expression of a compensatory enzyme<sup>33</sup>. Another study found a random protein with an unknown molecular mechanism that promotes survival in high concentrations of copper<sup>35</sup>.

Still, the functions that random proteins can assume inside cells remain poorly understood. Here, we screened a library of ~10<sup>8</sup> random genes for their ability to promote growth in the presence of the ribonuclease toxin MazF, finding ~2,000 unique genes that restore growth. Although most function non-specifically to reduce transcription from the promoter driving *mazF*, we found a single random anti-toxin of MazF, RamF, that specifically rescues cells from MazF toxicity. We characterized the function of RamF, its specificity for MazF, and the mutational pathways to becoming a more efficient inhibitor. Our experiments indicate that RamF is a well-tolerated cytosolic protein that remodels the physiology of *E. coli* cells by interacting directly with cellular chaperones, leading to MazF proteolysis. Thus, our work demonstrates how a small, random protein can instantly have a beneficial function, integrate into pre-existing cellular pathways and become improved by mutation and selection—thereby revealing a plausible mechanism for *de novo* gene birth.

## Results

### Selection for functional, random genes that inhibit a toxin

We sought to identify functional and beneficial genes originating from random nucleotide sequences. To this end, we created a library of ~10<sup>8</sup> plasmids, each harbouring a tetracycline-inducible promoter ( $P_{tet}$ ) driving a bicistronic operon with a first open reading frame (ORF) encoding a constant 17-amino acid peptide followed by a second ORF with an ATG start codon and then 50 random NNB codons (Fig. 1a; Methods). This bicistronic design minimizes translation initiation biases due to messenger RNA structures involving the second ORF<sup>36</sup>. Deep sequencing of the initial library demonstrated its high complexity, with 99.42% of the ~370,000 reads being single, unique sequences (Extended Data Fig. 1a). The average length of the random ORFs was 28 amino acids, with 23% of the random genes coding for 51 amino acid proteins (Fig. 1b).

We used this library to select genes that enable cells to grow following induction of the toxin MazF, an endoribonuclease that degrades a range of cellular RNAs to inhibit cell growth<sup>37</sup>. We transformed our library into an *E. coli* strain expressing *mazF* from an arabinose-inducible promoter ( $P_{ara}$ ) on the chromosome. We then induced expression of both the random genes and *mazF* to select those genes that inhibit MazF and promote growth. To enrich for true-positive hits and eliminate case of chromosomal mutations that trivially prevent *mazF* expression (for example,  $P_{ara}$  mutations), plasmids from the first round of selection were harvested and used to transform new cells harbouring  $P_{ara}$ -*mazF* (Fig. 1c).

Deep sequencing of the library after two selection rounds revealed ~2,000 enriched, random genes. We arbitrarily chose five of these genes and tested whether they inhibit two additional toxins: RelE, an unrelated ribonuclease toxin<sup>38</sup>, and Hok, a short hydrophobic toxin that depolarizes cell membranes<sup>39</sup>. All five hits could inhibit these toxins, which were also expressed from the arabinose-inducible promoter (Fig. 1d) and failed to inhibit MazF when the toxin was expressed from a vanillate-inducible promoter,  $P_{van}$  (Fig. 1d). Thus, these random genes are probably not directly inhibiting toxins and instead preventing transcription from the arabinose promoter. Consistent with this conclusion, we found that three of the random genes reduced the levels of

monomeric, super-folding GFP (msfGFP) expressed from  $P_{ara}$  but not from the  $P_{van}$  promoter (Fig. 1e).

To identify random genes that inhibit MazF independent of its promoter, we transformed the pool of ~2,000 candidates into an *E. coli* strain in which *mazF* was expressed from  $P_{van}$  (Fig. 1c). Two successive rounds of selection for growth on vanillate revealed a single random gene that could inhibit MazF driven by  $P_{ara}$  or  $P_{van}$  and that did not inhibit RelE or Hok (Fig. 1d). This gene did not affect levels of msfGFP produced from  $P_{ara}$  or  $P_{van}$  (Fig. 1e). We named this gene *ramF* for random antitoxin of MazF.

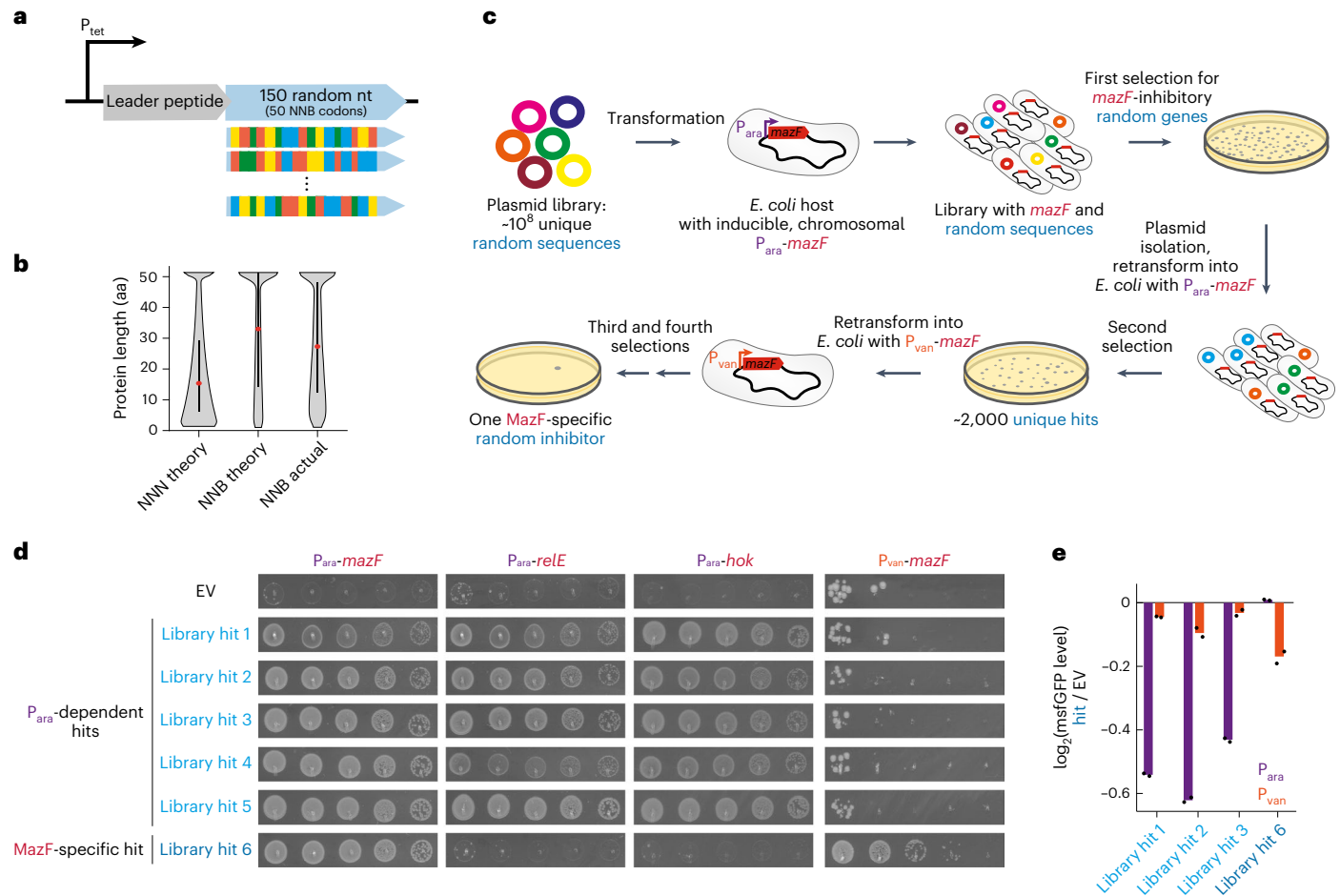
### RamF inhibits MazF by inducing its degradation

We sought to understand the molecular function the random protein RamF performs to rescue cells. The gene *ramF* has an ORF of 51 codons and an amino acid composition intermediate between small *E. coli* cytosolic and membrane proteins (Fig. 2a and Extended Data Fig. 2). No proteins with sequence similarity to RamF were found in existing sequence databases. We first replaced the short ORF upstream of *ramF* with a ribosome binding site (RBS) and confirmed the MazF-inhibitory activity of this new gene architecture (Fig. 2b). To confirm that the MazF-inhibitory activity of *ramF* depends on a small protein, rather than RNA, we mutated the start codon and found that this variant of *ramF* failed to inhibit MazF. We also generated a recoded variant of *ramF* with 46 synonymous mutations (representing changes to 30% of nucleotides in the ORF) and found that it could still inhibit MazF when co-expressed. Additionally, the originally selected *ramF* rescued growth inhibition following expression of a synonymously recoded *mazF* (83 mutations, 25% of the ORF) (Fig. 2b). Finally, *ramF* did not inhibit close homologues of the *E. coli* MG1655 *mazF*, the toxin used in our screen, as it did not rescue cells expressing *mazF* from the ECOR27 strain<sup>40</sup> or MG1655 *chpB*, the closest *mazF* homologue in that strain (Fig. 2c). Together, these findings suggest that *ramF* encodes a new protein that specifically alleviates the toxicity of the MG1655 MazF protein.

We next tested the effects of RamF on MazF levels. We generated a variant of MazF that could be easily used in molecular assays such as immunoblots as it harboured both a C-terminal His<sub>6</sub>-tag, which does not substantially impact function (Extended Data Fig. 3a) and an E24A substitution, which was shown to reduce but not eliminate, RNase activity<sup>41,42</sup>. Cells producing RamF had lower steady-state levels of MazF(E24A)-His<sub>6</sub> compared to cells with an empty vector (EV) (Fig. 2d). Production of MazE, the natural antitoxin of MazF that inhibits its toxicity via direct binding<sup>43</sup>, did not reduce MazF levels (Fig. 2d). Producing RamF also reduced the fluorescence of MazF(E24A) fused to msfGFP in individual cells compared to a control strain (Fig. 2e). These observations suggest that RamF inhibits MazF through a different mechanism than MazE, probably by reducing toxin levels. RamF did not reduce levels of ChpB(E24A)-His<sub>6</sub> (Fig. 2d), consistent with our finding that RamF did not neutralize ChpB toxicity (Fig. 2c).

Because RamF inhibits MazF in a promoter-independent manner, we hypothesized that RamF increases MazF degradation rather than reducing synthesis. To test this possibility, we treated cells producing MazF(E24A)-His<sub>6</sub> with tetracycline to block new protein synthesis and followed MazF(E24A)-His<sub>6</sub> levels over time. Cells producing RamF exhibited faster turnover of MazF(E24A)-His<sub>6</sub> compared to control cells (Fig. 2f), indicating that RamF rescues MazF toxicity by promoting its degradation.

To identify the protease(s) that degrade MazF, we measured MazF(E24A)-His<sub>6</sub> levels in strains producing RamF but lacking each of the major *E. coli* proteases (Fig. 2g). While MazF(E24A)-His<sub>6</sub> levels were not substantially changed in  $\Delta$ *hslV* or  $\Delta$ *htrX* cells,  $\Delta$ *clpP* cells showed an increase in MazF(E24A)-His<sub>6</sub> levels, suggesting that the ClpP protease helps degrade MazF. Because *ftsH* is essential for viability, we could only examine the effects of  $\Delta$ *ftsH* in the presence of the *sfhC* mutation<sup>44</sup>. Cells harbouring  $\Delta$ *ftsH* and the *sfhC* mutation showed substantially elevated levels of MazF(E24A)-His<sub>6</sub> compared to an isogenic



**Fig. 1 | Strategy for selecting functional proteins from a random sequence library. a**, Architecture of the random sequence library. A tetracycline-inducible promoter ( $P_{tet}$ ) drives the expression of a leader peptide followed by an ATG start codon, 150 random nucleotides (50 NNB codons), a stop codon and a transcriptional terminator. **b**, Theoretical protein lengths of 50 NNB codons are lower compared to 50 NNB codons. Actual library distribution as deduced by deep-sequencing preselection is similar to the NNB distribution. Black bar represents 50% of variants and red dot is the median. aa, amino acids. **c**, Selection strategy to identify functional proteins that inhibit the toxin MazF. Approximately  $10^8$  plasmids harbouring unique, random genes were transformed into an *E. coli* strain with a chromosomal, arabinose-inducible  $P_{ara}$ -*mazF* gene. Surviving colonies after *mazF* induction include true hits and false positives due to chromosomal mutations. Plasmids were purified, retransformed

into new cells and selected for a second time. The surviving colonies were then screened twice in a strain expressing *mazF* from the vanillate-inducible promoter ( $P_{van}$ ), resulting in a single gene that passed these selection steps. **d**, Tenfold serial dilution spotting of cells expressing one of the toxins *mazF*, *relE* or *hok* while co-expressing one of the random library hits (numbered 1–6) or an empty vector expressing the leader peptide only. Plasmids harboured the toxins under  $P_{ara}$  or  $P_{van}$  promoters as indicated. Plasmids carrying the random library hits driven by a  $P_{tet}$  promoter. **e**, The  $\log_2$  fold-change of median *mstGFP* fluorescence levels of library hits 1–3 and hit 6 relative to the control strain with an empty vector expressing the leader peptide only. *mstGFP* expressed from either  $P_{ara}$  or  $P_{van}$ , as indicated. Data are the mean of two biological repeats, each black dot is an individual measurement.

*sfhC* control, indicating that FtsH plays a key role in MazF degradation. Both  $\Delta clpP$  and  $\Delta ftsH$  strains demonstrated slower degradation rates of MazF(E24A)-His<sub>6</sub>, compared to control cells when *ramF* was expressed (Extended Data Fig. 4).

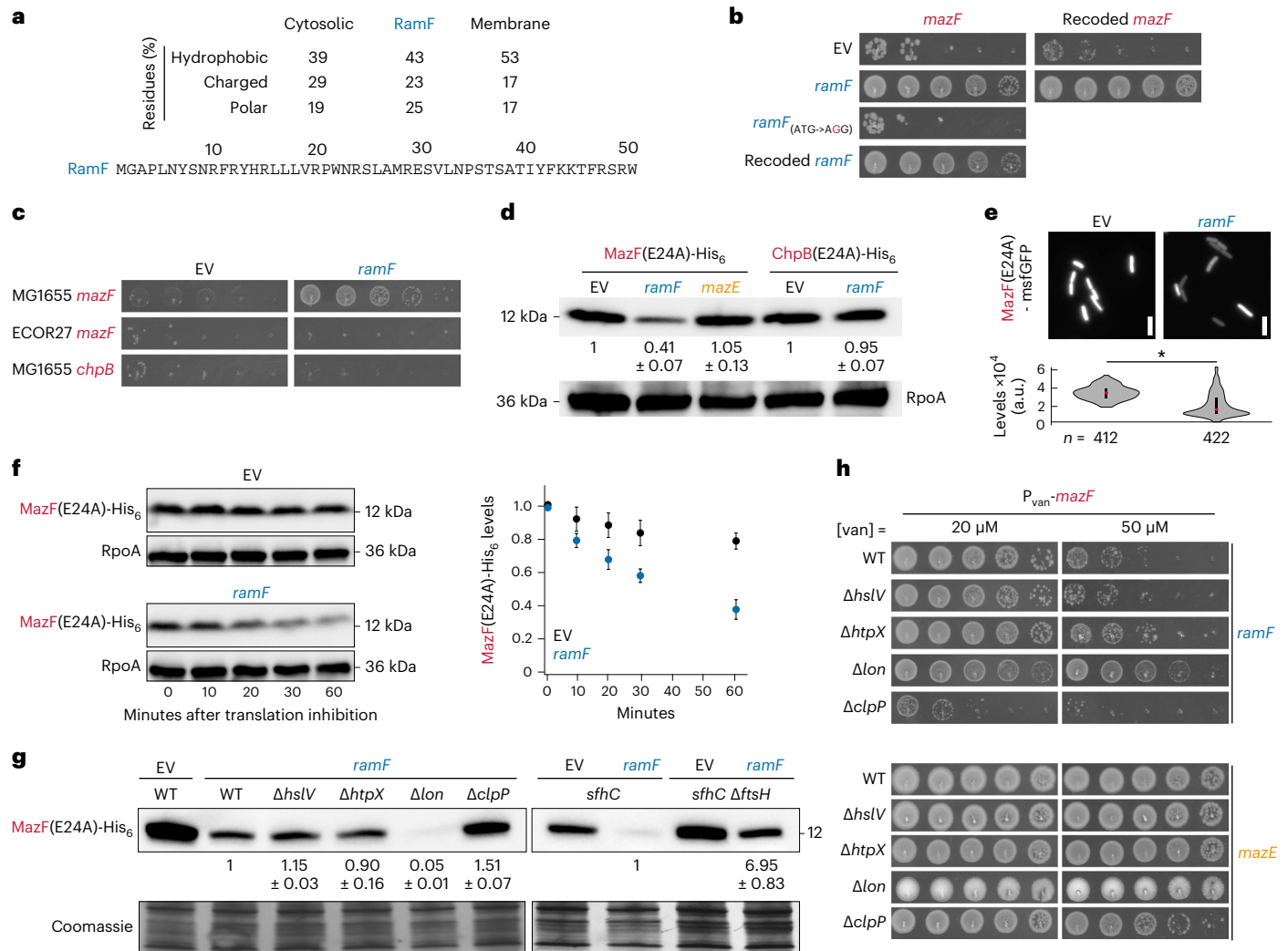
We found that MazF(E24A)-His<sub>6</sub> levels decreased in the  $\Delta lon$  strain. We first considered whether RamF might inhibit Lon, resulting in increased degradation of MazF(E24A)-His<sub>6</sub>, thereby phenocopying the  $\Delta lon$  strain. However, RamF did not decrease Lon activity, RamF could inhibit MazF in cells overproducing Lon and producing the known Lon inhibitor PinA did not inhibit MazF (Extended Data Fig. 5a–d). As an alternative, we proposed that RamF might be a Lon substrate such that RamF levels are increased in a  $\Delta lon$  strain, leading to more rapid degradation of MazF(E24A)-His<sub>6</sub> in  $\Delta lon$  cells. To test this idea, we created a functional, N-terminally FLAG-tagged RamF (Extended Data Fig. 3b) and compared its steady-state levels in control and  $\Delta lon$  cells. Indeed, FLAG-RamF levels increased in a  $\Delta lon$  strain (Extended Data Fig. 5f).

Because the activity of RamF depends on toxin-induced degradation, we predicted that RamF inhibition efficiency would change in

protease deletion strains that altered MazF levels. Indeed, for  $\Delta lon$  cells in which MazF levels were reduced, RamF was functional at higher MazF induction levels than in control cells (Fig. 2h). In contrast, RamF did not inhibit MazF in  $\Delta clpP$  cells as efficiently as in control cells (Fig. 2h) and it was impossible to transform a plasmid harbouring *mazF* into  $\Delta ftsH$  cells, presumably because even leaky expression leads to enough MazF accumulation and toxicity. As controls, we confirmed that deleting either *hslV* or *htpX*, which did not affect MazF levels, did not affect RamF function. Additionally, we showed that the neutralization of MazF by MazE, which inhibits MazF independent of proteolysis, was not substantially affected by protease deletions.

### RamF interacts with chaperones to modify protein homeostasis

Our results demonstrated that RamF prevents MazF toxicity by facilitating its degradation, particularly via the FtsH protease. Known substrates of FtsH also exhibited decreased steady-state levels in RamF-producing cells (Extended Data Fig. 6a), raising the possibility that RamF activates



**Fig. 2 | RamF is a specific MazF inhibitor that induces MazF proteolysis.**

**a**, Amino acid sequence of library hit 6, named *ramF* and its amino acid composition compared to small proteins (<100 amino acids) in *E. coli* that are either cytosolic ( $n = 181$ ) or membrane-localized ( $n = 80$ ). **b**, Tenfold serial dilution spotting of cells expressing *mazF* with either its original nucleotide sequence or a synonymously recoded version from the P<sub>van</sub> promoter. Cells additionally expressed from the P<sub>tet</sub> promoter one of the following: *ramF*, *ramF* with a start codon mutation, synonymously recoded *ramF* or an empty vector. The leader peptide of the library is not expressed in these cells or experiments hereafter. **c**, Tenfold serial dilution spotting of cells expressing MG1655 *mazF* (reference sequence), ECOR27 *mazF* (56% identity) or MG1655 *chpB* (33% identity). Cells also expressed *ramF* from the P<sub>tet</sub> promoter or carried an empty vector. **d**, Immunoblot of MazF(E24A)-His<sub>6</sub> or ChpB(E24A)-His<sub>6</sub>, expressed from P<sub>van</sub>, in cells co-expressing *ramF*, *mazE* or an empty vector. RpoA is a loading control. Quantification is the mean of  $n = 3$  biological repeats and values are normalized to levels in the empty vector strains. **e**, Fluorescence intensities of MazF(E24A)-GFP

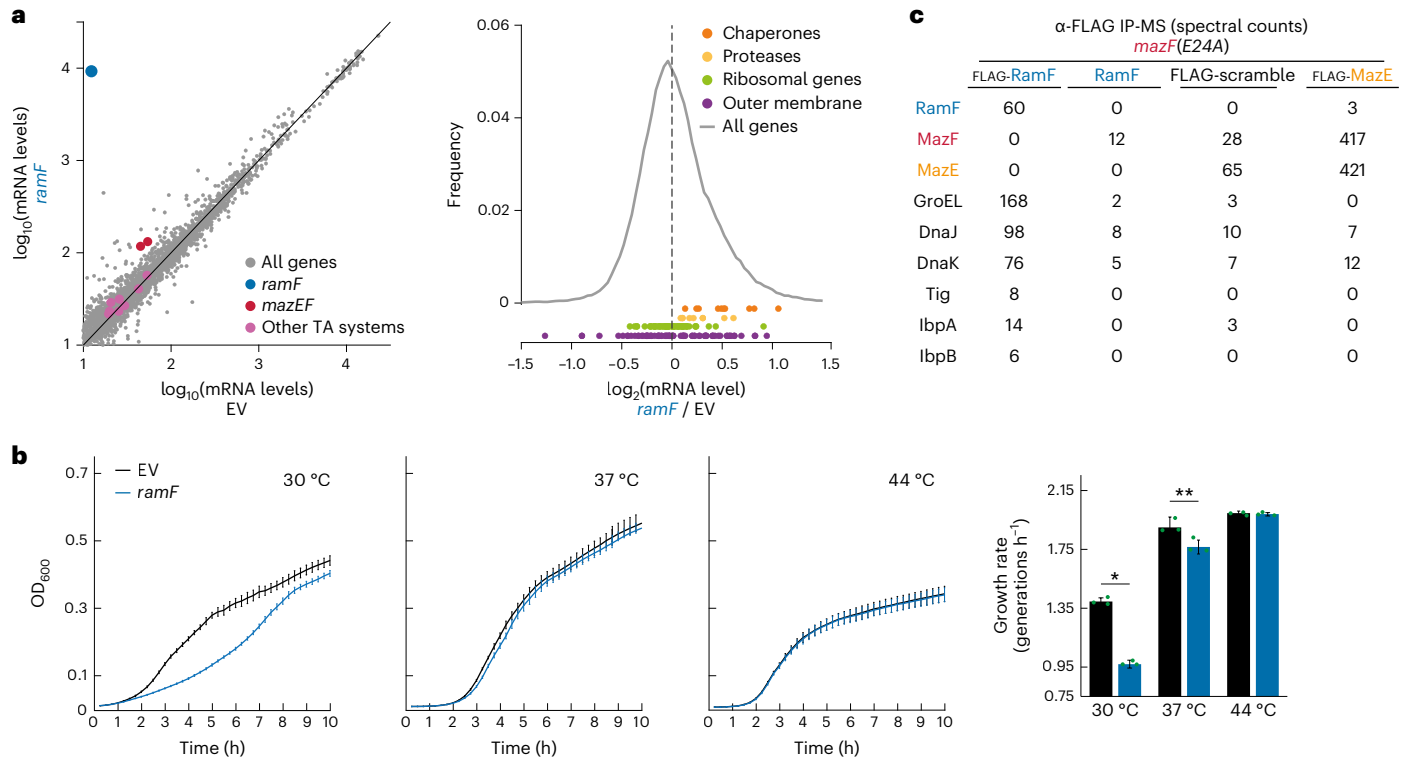
in cells expressing *ramF* or harbouring an empty vector. Violin plots: black bar represents the middle 50% of cells and red dot is the median.  $*P = 4.53 \times 10^{-93}$  based on a two-sided  $t$ -test,  $n = 412$  and 422 cells measured with empty vector or *ramF*, respectively. Scale bars, 2  $\mu$ m. **f**, Immunoblot of MazF(E24A)-His<sub>6</sub> from cells co-expressing *ramF* or harbouring an empty vector. Time points were taken after the addition of tetracycline to stop the translation of new proteins. RpoA is a loading control. Quantification is based on the mean of  $n = 3$  biological repeats, error bars represent s.d. and levels are normalized to  $t = 0$ . **g**, Same as **f** but for strains also lacking one of the major proteases of *E. coli*, as indicated. Loading control is based on Coomassie staining of total protein. Quantification for relevant strains is the mean of  $n = 3$  biological repeats and values are normalized to the *ramF*-expressing strain. Levels in the  $\Delta$ ftsH *sfhC* strain are normalized to the *sfhC* control strain. **h**, Tenfold serial dilution spotting of cells co-expressing *mazF* with either *ramF* or *mazE* in cells also lacking one of the major proteases of *E. coli*. A plasmid harbouring *mazF* could not be transformed into  $\Delta$ ftsH cells. WT, wild type.

FtsH. However, overproducing FtsH in cells lacking RamF was insufficient to inhibit MazF and did not alter RamF efficiency as a MazF inhibitor (Extended Data Fig. 6b), suggesting that RamF does not inhibit MazF by simply activating FtsH.

How, then, can this random 51 amino acid protein mediate MazF proteolysis? To characterize the physiological changes caused by RamF production, we first compared global RNA levels in cells expressing RamF and an empty vector control. We found that RamF does not lead to major transcriptional changes (Fig. 3a). There was, however, an ~2.5-fold upregulation of the native *mazEF* locus (Fig. 3a left, red dots), supporting a model of RamF-dependent degradation

of MazF because the MazEF complex negatively autoregulates *mazEF* expression<sup>43,45</sup>; thus, degradation of MazF leads to upregulation of *mazEF*. In agreement with RamF being a specific MazF inhibitor (Figs. 1 and 2), the mRNA levels of other toxin-antitoxin (TA) systems, which are also autoregulated, were not affected (Fig. 3a left, pink dots,  $P = 0.16$ ,  $t$ -test).

Because RamF production results in MazF proteolysis, we tested if the production of RamF affected protein homeostasis pathways, finding that chaperones and proteases were modestly, but statistically significantly, upregulated (Fig. 3a, right,  $P = 1.94 \times 10^{-4}$  and  $P = 0.04$ , respectively,  $t$ -test). In comparison, the expression of other gene



**Fig. 3 | RamF interacts with chaperones.** **a**, Left,  $\log_{10}$  of mRNA levels in Transcripts Per Million for *E. coli* genes in cells expressing *ramF* or harbouring an empty vector. Right,  $\log_2$  of the mRNA level ratio between RamF-producing cells and cells with an empty vector. Colours: grey, all genes; red, *mazEF*; pink, other TA systems; khaki, ribosomal protein genes; purple, outer-membrane genes; orange, chaperones; yellow, proteases. Data based on two biological repeats. **b**, Left, growth curves for cells expressing *ramF* or harbouring an empty

vector growing at 30, 37 or 44 °C as a mean of  $n = 3$  biological repeats. Right, maximal growth rates calculated from growth curves. \* $P = 1.27 \times 10^{-5}$ , \*\* $P = 0.03$  based on a two-sided  $t$ -test, error bars represent s.d. and each green dot is an individual measurement. **c**, Spectral counts of *E. coli* proteins detected by mass spectrometry following a pull-down with  $\alpha$ -FLAG beads from a lysate of cells producing MazF(E24A) and FLAG-RamF, RamF (negative control), FLAG-scrambled-RamF (negative control) or FLAG-MazE (positive control).

groups, for example ribosomal and outer-membrane gene groups, were unaffected (Fig. 3a, right,  $P = 0.41$  and  $P = 0.21$ , respectively,  $t$ -test).

Our RNA sequencing data suggest that RamF was well tolerated by cells and did not induce a strong stress response. In agreement, producing RamF had a minimal effect (0–2% reduction compared to control cells) on lag times and culture yields at 37 or 44 °C (Extended Data Fig. 7). At 37 °C in LB medium, *ramF* expression led to a small cost in exponential-phase growth rate (Fig. 3b). At 44 °C, *ramF*-expressing cells grew identically to control cells, whereas at 30 °C *ramF* expression caused a severe growth reduction. This temperature-dependent phenotype further indicated that RamF activity may depend on protein homeostasis pathways as chaperone levels are often temperature-dependent<sup>46–52</sup>.

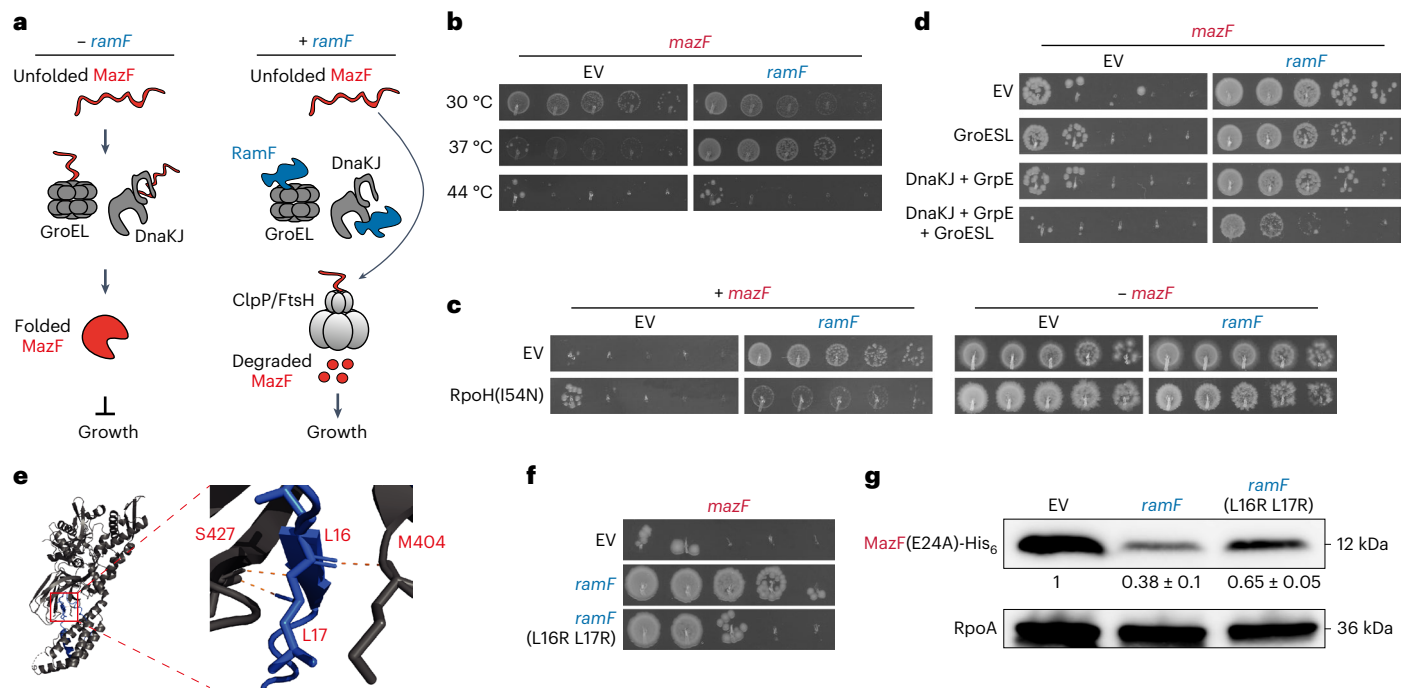
To further investigate how RamF affects cell physiology, we sought to find what proteins RamF interacts within cells. We produced functional FLAG-RamF in cells coproducing MazF(E24A), immunoprecipitated RamF using  $\alpha$ -FLAG beads and then identified co-eluting proteins by mass spectrometry. We did not detect MazF (Fig. 3c). As a control, we showed that the same procedure using a strain producing FLAG-MazE, did detect MazF, as expected. These results support our conclusion that RamF inhibits MazF via a different mechanism than MazE.

Our mass spectrometry data (Supplementary Table 2) revealed enrichment of multiple proteins that immunoprecipitated with FLAG-RamF but not with two negative control experiments: a strain producing untagged RamF and a strain producing FLAG-tagged but scrambled (same amino acid composition but in a randomized order) RamF protein that could not inhibit MazF (Extended Data Fig. 3b). This analysis revealed that RamF strongly interacts with cellular chaperones, including GroEL (Hsp60), DnaK/J (Hsp70), trigger factor and IbpA/B

(Fig. 3c). RamF also appeared to interact with HldD, PepN and SlyD but deletions of each did not affect the ability of RamF to inhibit MazF through induction of toxin proteolysis (Extended Data Fig. 8).

In sum, our results demonstrated that RamF (1) drives increased proteolysis of MazF, (2) promotes increased expression of chaperones and proteases, (3) interacts in vivo with chaperones and (4) results in a growth defect at a temperature where chaperone expression levels are relatively low. On the basis of these findings, we proposed the following model for MazF inhibition by RamF. In cells lacking RamF, chaperones assist MazF to adopt its native, folded state, which can then cleave RNA and thereby inhibit growth (Fig. 4a, left). In cells producing RamF, chaperones become occupied by RamF such that MazF is unable to fold properly, leaving it susceptible to proteolysis, which allows cellular growth (Fig. 4a, right).

To test this model, we first tested if temperature, which is correlated with chaperone levels, affects the ability of RamF to inhibit MazF. Indeed, we found that MazF failed to inhibit growth at 30 °C even in the absence of RamF, possibly because of insufficient chaperone activity to fold MazF. Also, RamF rescued MazF toxicity at 37 °C but not at 44 °C (Fig. 4b). In agreement, we found that MazF expression levels correlate with growth temperature (Extended Data Fig. 9a). Although consistent with our model, growth temperature affects cell physiology in many ways. Thus, to increase chaperone availability in a more controlled manner, we used a strain producing the heat shock sigma factor ( $\sigma^{32}$ ) encoded by *rpoH*, which regulates many *E. coli* chaperones. We used an RpoH variant with an I54N substitution that prevents the degradation of this protein and thus maintains its activity<sup>53</sup>. RamF failed to rescue cells producing both MazF and RpoH(I54N) at various temperatures (Fig. 4c and Extended Data Fig. 9b). We also generated



**Fig. 4 | RamF remodels cellular physiology to change protein homeostasis.**

**a**, Model for RamF function as a MazF inhibitor. In cells not producing RamF, chaperones promote proper folding of MazF, leading to widespread RNA degradation and cell growth arrest. In RamF-producing cells, RamF binds chaperones and prevents MazF maturation, allowing FtsH and ClpP to degrade MazF and restore cell growth. **b**, Tenfold serial dilution spotting of cells expressing *mazF* from  $P_{ana}$  and co-expressing *ramF* or harbouring an empty vector, incubated at 30, 37 or 44 °C, as indicated. **c**, Tenfold serial dilution spotting of cells expressing *mazF* from  $P_{van}$ . Cells also express combinations of *ramF*, *rpoH*(I54N) or empty vector, as indicated. **d**, Tenfold serial dilution

spotting of cells expressing *mazF* from  $P_{van}$ . Cells also express combinations of *ramF*, *groESL*, *dnaKJ* + *grpE*, *groESL* + *dnaKJ* + *grpE* or empty vectors, as indicated. **e**, AlphaFold2 prediction of the interactions between residues M404 and S427 within the substrate-binding domain of DnaK with residues L16 and L17 of RamF. **f**, Tenfold serial dilution spotting of cells expressing *mazF* from  $P_{van}$  and co-expressing *ramF*, *ramF*(L16R + L17R) or an empty vector, as indicated. **g**, Immunoblot of MazF(E24A)-His<sub>6</sub>, expressed from  $P_{van}$ , from cells co-expressing *ramF*, *ramF*(L16R + L17R) or an empty vector. Loading control is RpoA. Quantification is the mean of  $n = 3$  biological repeats and values are normalized to MazF(E24A)-His<sub>6</sub> levels in the empty vector strain.

cells that overproduce the chaperone system DnaK/DnaJ/GrpE or GroEL/GroES or both. Overproducing individual chaperone systems partially reduced the ability of RamF to alleviate MazF toxicity, with a substantial drop in RamF activity when overproducing both systems (Fig. 4d). Consistently, overproduction of RpoH(I54N) marginally alleviated the growth defect of RamF-producing cells grown at 30 °C (Extended Data Fig. 9c). Together, these results demonstrate that cellular availability of chaperones is critical to RamF function.

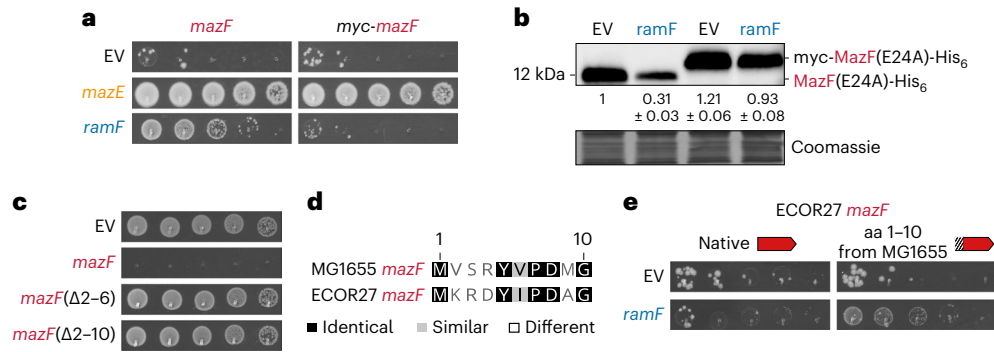
Finally, we asked if the interaction between RamF and chaperones detected in our immunoprecipitation-mass spectrometry (IP-MS) data are important for MazF inhibition. Using AlphaFold2 (refs. 54,55), we modelled the interaction between RamF and DnaK and found that M404 and S427 in DnaK are predicted to bind L16 and L17 in RamF, respectively (Fig. 4e). Notably, these residues in DnaK are found in its substrate-binding domain<sup>56</sup> and were previously shown to bind two contiguous Leu residues of a model peptide<sup>57,58</sup>. A variant of RamF with the substitutions L16R and L17R was not co-immunoprecipitated with DnaKJ as well as the original RamF (Extended Data Fig. 9d). RamF(L16R L17R) also did not inhibit (Fig. 4f) or induce degradation of MazF (Fig. 4g) as efficiently as RamF. Producing RamF(L16R L17R) also resulted in lower overall protein aggregation levels in cells (Extended Data Fig. 9e, see next section). These results are consistent with our model that the interaction of RamF with chaperones is critical to MazF inhibition.

### The N terminus of MazF partially determines RamF specificity

Our results thus far indicate that RamF interacts with central protein homeostasis pathways, which ultimately results in MazF proteolysis.

Using a previously characterized reporter for protein aggregation in *E. coli*<sup>59,60</sup>, we found that producing RamF led to increased protein aggregation (Extended Data Fig. 9e), suggesting that the folding of other proteins is affected by RamF chaperone occupancy. Given this function, how does RamF inhibit *E. coli* MG1655 MazF but not other close MazF homologues (Fig. 1b) that share similar predicted structures (Extended Data Fig. 10a)? We speculated that this specificity might stem from *E. coli* MG1655 MazF, but not its homologues, being recognized by FtsH. The FtsH protease can recognize substrates via unique degron sequences at the N or C termini of proteins or internally<sup>61–64</sup>. Because C-terminal tagging of MazF did not change RamF-dependent inhibition (Extended Data Fig. 3b), we tested the relevance of its N terminus to degradation. We fused an N-terminal myc tag to MazF and found that while inhibition by MazE was maintained, the tag abolished inhibition by RamF (Fig. 5a). This result suggests that tagging MazF on its N terminus prevented degradation, presumably by occluding the degron. Indeed, myc-MazF(E24A)-His<sub>6</sub> levels did not decrease in cells expressing *ramF* (Fig. 5b). Additionally, removing amino acids 2–6 or 2–10 eliminated MazF toxicity (Fig. 5c), suggesting that this region not only mediates MazF degradation but is essential to MazF toxicity.

A sequence alignment of MG1655 MazF and ECOR27 MazF indicated that the first ten amino acids differ at five positions (Fig. 5d and Extended Data Fig. 10b). We hypothesized that replacing these amino acids in ECOR27 MazF with those of MG1655 MazF might make this chimaeric protein a better FtsH substrate and therefore sensitive to RamF inhibition. Indeed, RamF gained the ability to inhibit ECOR27 MazF when its first ten amino acids matched those in MG1655 MazF (Fig. 5e). Taken together, our results explain how (1) a new, random



**Fig. 5 | The N terminus of MazF is essential for its inhibition by RamF.**

**a**, Tenfold serial dilution spotting of cells expressing *mazF* or *myc-mazF* from  $P_{van}$ . Cells also express *ramF*, *mazE* or an empty vector, as indicated. **b**, Immunoblot of MazF(E24A)-His<sub>6</sub> or myc-MazF(E24A)-His<sub>6</sub>, expressed from  $P_{van}$ , from cells co-expressing *ramF* or harbouring an empty vector. Loading control is based on Coomassie staining of total protein. Quantification is the mean of  $n = 3$  biological

repeats and values are normalized to MazF(E24A)-His<sub>6</sub> levels in the empty vector strain. **c**, Tenfold serial dilution spotting of cells expressing *mazF*, *mazF*( $\Delta 2-6$ ), *mazF*( $\Delta 2-10$ ) or an empty vector, as indicated. **d**, Sequence alignment of the first ten positions of MG1655 MazF and ECOR27 MazF. **e**, Tenfold serial dilution spotting of cells expressing ECOR27 *mazF* or ECOR27 *mazF*(1-10) from MG1655) from  $P_{ara}$ . Cells are additionally expressing *ramF* or an empty vector, as indicated.

protein that interacts with central cellular pathways can have a specific effect on a single target and (2) how accumulation of mutations on new targets can make them susceptible to this effect.

### Mutations that improve RamF as a MazF inhibitor are common

Once a de novo gene like *ramF* is established in a genome, natural selection can, in principle, improve its activity via subsequent beneficial mutations. To ask whether RamF can become a better MazF inhibitor, we used PCR-based mutagenesis to create a library of ~60,000 RamF variants. This library was transformed into the same *E. coli* strain used in the initial screen and selected on higher levels of MazF such that MazE rescues growth but the original RamF cannot (Fig. 6a,b; Methods). The library was deep-sequenced pre- and postselection to find mutations enriched by the selection (Fig. 6c). We found five mutations that individually improved the inhibition of MazF: F11L, R12M, T40A, I41T and W51\* by RamF (Fig. 6d). Combinations of these mutations mostly showed additive phenotypes, except for T40A and I41T which exhibited strong negative epistasis (Fig. 6d). We also generated an improved RamF variant harbouring F11L, I41T and W51\*, which was the most efficient MazF inhibitor (Fig. 6d). We confirmed that the RamF(F11L I41T W51\*) variant also reduced MazF(E24A)-msfGFP levels further compared to cells expressing RamF (Fig. 6e).

What mechanisms could underline the beneficial mutations in RamF? The W51\* nonsense mutation replaces the hydrophobic tryptophan with a positively charged arginine at the C terminus of RamF, suggesting that this change stabilizes RamF and increases its steady-state level. Indeed, we observed an ~20% increase in RamF(W51\*) levels compared to RamF (Fig. 6f). We also found that a RamF(R50A W51\*) variant showed an ~40% decrease in expression levels and could not inhibit MazF (Fig. 6f,g), further indicating that RamF levels impact its function. Finally, we found that a RamF variant with the I41T substitution led to higher protein aggregation compared to the original RamF (Extended Data Fig. 9e). The W51\* mutation showed a similar, but less pronounced, increase in aggregation. These results suggest that beneficial mutations that improve RamF functions are common and easily accessible by natural selection.

### Discussion

There is increasing interest in the discovery and characterization of small proteins (<50 amino acids) in biological systems<sup>65-67</sup>. Using new detection methods<sup>68-72</sup>, small ORFs are being discovered across the tree of life, yet their evolutionary origin is enigmatic. The study of randomly generated proteins can support a de novo origin for natural small proteins by demonstrating how the former assume beneficial biological functions.

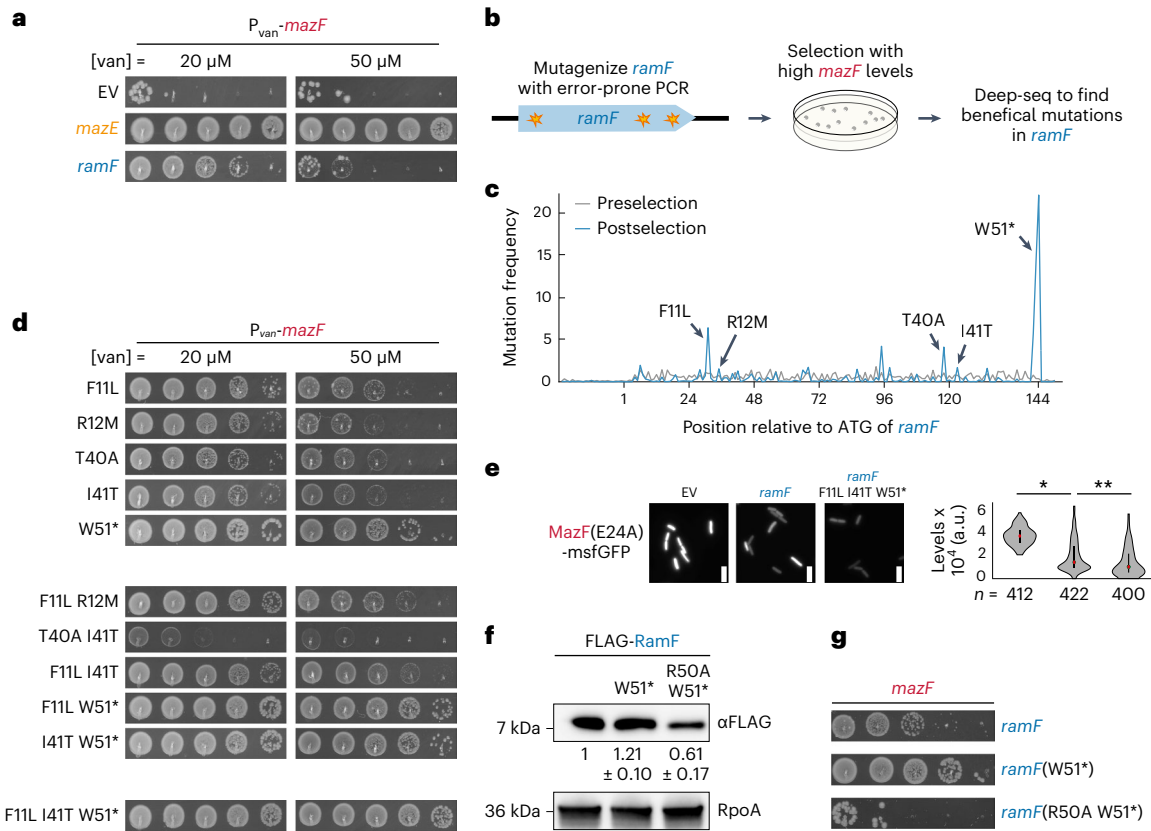
Here, we selected for random proteins that inhibit the toxin MazF. We identified ~2,000 hits that block MazF in a promoter-dependent manner probably by reducing expression from  $P_{ara}$ , although we have not characterized these hits in depth. Why did we find considerably more hits targeting the arabinose promoter than MazF itself? A likely explanation is that the complex arabinose pathway<sup>73-75</sup> simply provides more opportunities for random proteins to prevent activation of  $P_{ara}$ . Additionally, inhibiting the arabinose pathway may be less likely to perturb essential cellular functions, allowing more solutions to emerge. Whatever the case, these hits demonstrate that random proteins can readily adopt beneficial functions inside cells.

We identified one random protein, RamF, that rescued cells in a promoter-independent manner through interactions with cytosolic chaperones that remodel the physiology of *E. coli* cells. RamF was our only promoter-independent hit from a pool of ~10<sup>8</sup> sequences. On one hand, this is surprising given the tendency of random proteins to include hydrophobic regions<sup>5,21</sup> and bind chaperones in vitro<sup>22</sup>. However, other hydrophobic random proteins in our library may have suffered from one of the following shortcomings: (1) a fast turnover that prevents functional interactions with cellular components, (2) a transmembrane domain leading to membrane localization, (3) a hydrophobic amino acid composition that leads to toxic aggregation or (4) activation of the stress responses that offset any beneficial change in cell physiology.

Our results indicated that RamF is specific to MazF, relative to other toxins. However, RamF did result in increased overall protein aggregation levels (Extended Data Fig. 9e), suggesting that the folding of other proteins was affected by the interaction of RamF with chaperones. RamF did not inhibit close homologues of MazF, probably because they lack the N-terminal degron in MG1655 MazF. Alternatively, higher levels of RamF could be required to impact these other toxins, underscoring the notion that the genomic and cellular context in which random proteins emerge can affect their functionality. Whatever the case, to be selected in nature, a de novo gene must cross an expression threshold that allows its function.

### Fitness effects of de novo proteins

Overproduction of some yeast de novo gene candidates positively impacted growth<sup>5</sup>, demonstrating the benefit these genes can have for the fitness of microorganisms. However, other studies have found that random proteins isolated in functional selections can strongly activate the cellular SOS response<sup>34</sup>, reduce cell growth rate<sup>31</sup> or increase growth lag time and decrease culture yield<sup>32</sup>. RamF resulted in a substantial fitness cost at 30 °C, a much lower cost at 37 °C (temperature at which it was selected) and no fitness cost at 44 °C. What does such a cost



**Fig. 6 | Beneficial mutations that optimize the function of RamF as a MazF inhibitor.** **a**, Tenfold serial dilution spotting of cells expressing *mazF* from  $P_{van}$ . Cells are additionally expressing *ramF* or *mazE*, as indicated. **b**, Selection strategy for identifying beneficial mutations that improve RamF activity. Variants of *ramF*, generated by random mutagenesis of *ramF* with error-prone PCR, were selected in the presence of high MazF levels that the original RamF cannot neutralize. **c**, Frequency of *ramF* variants pre- and postselection on high MazF levels. **d**, Tenfold serial dilution spotting of cells expressing *mazF* from  $P_{van}$ . Cells were additionally expressing *ramF* variants, as indicated. **e**, Fluorescence intensities of MazF(E24A)-GFP in cells expressing *ramF*, *ramF*(F11L I41T W51\*) or

harbouring an empty vector. Data for empty vector and *ramF* are as in Fig. 2e. Violin plots: black bar represents the middle 50% of cells and red dot is the median. \* $P = 4.53 \times 10^{-93}$ , \*\* $P = 9.97 \times 10^{-8}$  based on a two-sided *t*-test,  $n = 412$ , 422 and 400 cells measured for cells with empty vector, *ramF* or *ramF*(F11L I41T W51\*), respectively. Scale bars, 2  $\mu\text{m}$ . **f**, Immunoblot of FLAG-tagged RamF, RamF(W51\*) and RamF(R50A W51\*) expressed from  $P_{tet}$ . Loading control is based on RpoA and quantification is the mean of  $n = 3$  biological repeats. **g**, Tenfold serial dilution spotting of cells expressing *mazF* from  $P_{van}$ . Cells were additionally expressing *ramF*, *ramF*(W51\*) or *ramF*(R50A W51\*).

mean for the chance of a new de novo protein to emerge in nature? A proto-gene probably has a better chance of fixing in an evolving population if producing its protein product does not come with a massive growth cost. Yet, many natural genes have been shown to provide a benefit in some conditions while being deleterious in others<sup>76,77</sup>. Additionally, selection could potentially reduce the costs of a new gene in some conditions through beneficial or compensatory mutations or by ensuring that the gene is only expressed at times it is beneficial.

### Relevance of random proteins to the study of de novo gene birth

We screened a library of random proteins against the toxin MazF but when do biological systems face this challenge? Toxin-antitoxin systems are widespread in bacteria and found on both chromosomes and plasmids<sup>78,79</sup>. Notably, antitoxins for the homologues of a given toxin are often not homologous themselves, suggesting that antitoxins can readily change and possibly arise de novo via a pathway similar to that reported here for RamF. Additionally, antitoxins are often short proteins harbouring unstructured domains, which bind their toxin counterparts<sup>80–84</sup>. Random proteins and young genes also tend to be short and unstructured<sup>21,85,86</sup>, further supporting the possibility that some antitoxins have arisen de novo.

Although RamF inhibited MazF toxicity, we did not find a random protein that directly interacted with this toxin, like the natural antitoxin

MazE. One explanation could be that more than  $10^8$  proteins should be screened to find a specific, strong protein–protein interaction and that integration into pre-existing pathways in the cellular system is a more accessible mechanism for random proteins to provide benefits to cells. Such ‘physiology modifiers’ may be used in cellular evolution as initial but pleiotropic solutions until a more specific one is found. In any case, the idea that the expression of random sequences, probably through spurious transcription and translation, can be advantageous is critical for de novo genes to emerge. Our work demonstrates the feasibility of this randomness-to-function process and provides molecular insight into how de novo genes can integrate into existing cellular pathways.

## Methods

### Plasmids, strains and growth conditions

All strains and plasmids used in this study are listed in Supplementary Table 1. *E. coli* was grown in LB medium (10 g l<sup>-1</sup> of NaCl, 10 g l<sup>-1</sup> of tryptone, 5 g l<sup>-1</sup> of yeast extract) or M9 medium (10× stock made with 64 g l<sup>-1</sup> of Na<sub>2</sub>HPO<sub>4</sub>·7H<sub>2</sub>O, 15 g l<sup>-1</sup> of KH<sub>2</sub>PO<sub>4</sub>, 2.5 g l<sup>-1</sup> of NaCl, 5.0 g l<sup>-1</sup> of NH<sub>4</sub>Cl supplemented with 0.1% casamino acids, 0.4% glycerol, 2 mM MgSO<sub>4</sub> and 0.1 mM CaCl<sub>2</sub>). In cases where M9 was used, 0.8% glucose was added to prevent leaky expression from the arabinose-inducible promoter. Media for selection or plasmid maintenance were supplemented with carbenicillin (100  $\mu\text{g ml}^{-1}$ ), chloramphenicol (20  $\mu\text{g ml}^{-1}$ ) or kanamycin (30  $\mu\text{g ml}^{-1}$ ) as appropriate. Overnight cultures were



prepared in the same medium used in a given experiment and cells were grown at 37 °C and 180 rpm in an orbital shaker. The arabinose-, tetracycline- and vanillate-inducible promoters were induced with 0.0002%–0.2% arabinose, 0.1 ng  $\mu\text{l}^{-1}$  of anhydrous tetracycline (aTc) and 15–100  $\mu\text{M}$  vanillate, respectively.

Plasmids were generated by Gibson assembly according to the manufacturer's protocol. Inserts were either amplified from a template by PCR or commercially synthesized by Integrated DNA Technology (IDT) as gBlocks. All plasmids were confirmed by Sanger sequencing of the inserts or by full-length plasmid sequencing by Plasmidsaurus. Plasmids were introduced into cells by either TSS transformation or electroporation. DNA and primers used in this study are found in Supplementary Table 3.

### ***E. coli* genome engineering**

To construct *E. coli* BW27783 *amyA::P<sub>ara</sub>-toxin/msfGFP* (strains ML-4045 to ML-4048) and *E. coli* BW27783 *amyA::P<sub>van</sub>-toxin/msfGFP* (strains ML-4049 to ML-4050), the *P<sub>ara</sub>-toxin*, *kan<sup>R</sup>* or *P<sub>van</sub>-toxin*, *kan<sup>R</sup>* cassettes were PCR amplified from plasmids with primers that included homology to the *amyA* locus. These amplicons were inserted into the genome of the arabinose titratable strain BW27783 (ref. 87) using the lambda red-based recombination<sup>88</sup>. Single insertions were confirmed by PCR and Sanger sequencing for individual colonies.

### **Assembly and transformation of the random gene library**

The random gene library was constructed by cloning 150 random nucleotides into the vector ML-4052 such that they immediately followed an ATG and were followed by two TAA stop codons. Specifically, pooled single-stranded DNA oligos of 50 NNB codons flanked on their 5' end by the sequence GCCTGGCTACCGTCTCGTATG and on their 3' end by TAATGGAGACGAGCAGGCGATG were synthesized by IDT. To avoid frequent premature stop codons, NNB codons, rather than NNN codons, were used; NNB libraries produce similar amino acid composition to NNN libraries. Oligos were PCR amplified using KAPA enzyme according to manufacturer recommendations with 16 amplification cycles. Six independent reactions were performed and combined to minimize PCR bias. Amplicons of the expected size of 193 nucleotides were purified from a gel using a Zymo Gel DNA Recovery kit and ~500 ng of this insert double-stranded DNA were digested and cut using the type IIS restriction enzyme Esp3I at 37 °C for 3 h to reach full digestion. Approximately 500 ng of the vector ML-4052 were similarly cut by BsmBI and both the insert and vector were subsequently purified on a Zymo DNA clean column. Then, 250 fmol of the vector and 1.25 pmol of the insert were combined in a 20  $\mu\text{l}$  ligation reaction with T4 ligase and Esp3I enzyme. The ligation reaction was cycled between 16 °C for 2 min and 37 °C for 2 min for 100 cycles to allow iterative ligation and digestion. This approach increased the ligation efficiency because once an insert was ligated to a vector it could no longer be cut by the restriction enzyme. Ligations were dialysed on Millipore VSWP 0.025  $\mu\text{m}$  membrane filters for 60 min and then the entire volume was electroporated into 20  $\mu\text{l}$  of Invitrogen MegaX DH10B cells, which resulted in  $\sim 10^8$  transformants. Transformants were grown overnight (14 h) in 50 ml of LB + carbenicillin. Then, the culture was split: 25 ml were frozen in 20% glycerol for long-term storage at  $-80$  °C and 25 ml were prepped for plasmids. The plasmid library of random genes was then dialysed and electroporated into *E. coli* strain ML-4045 to yield  $\sim 5 \times 10^8$  transformants.

### **Amplicon sequencing of random library and analysis**

To assess the library complexity pre- and postselection, random sequences were amplified using a forward primer that included the Illumina anchors and indexes as well as a region directly upstream of the random nucleotides and a reverse primer matching a region immediately downstream of the random nucleotides. PCR reactions were performed using KAPA enzyme according to manufacturer

recommendations with ten amplification cycles. Four independent reactions were performed and combined to minimize PCR bias. Amplicons were purified from an agarose gel using a Zymo Gel DNA Recovery kit. Paired-end sequencing was performed on an Illumina MiSeq at the MIT BioMicro Center. Paired-end reads were merged using PEAR with default parameters and identical reads were clustered using usearch with default parameters.

### **Bacterial growth by spotting assay on solid media**

In experiments with *P<sub>ara</sub>* induction, cultures were grown to saturation overnight in M9-glucose supplemented with 5% LB and the appropriate antibiotics. Cultures were then serially diluted tenfold and spotted on appropriate plates supplemented with 0.8% glucose (toxin repressing), 0.0002%–0.2% arabinose (toxin inducing), 100 ng  $\mu\text{l}^{-1}$  of aTc (random gene inducing) or 0.0002%–0.2% arabinose and 100 ng  $\mu\text{l}^{-1}$  of aTc (toxin and random gene inducing). Plates were then incubated at 37 °C for 24–36 h before imaging. A similar approach was used in experiments with *P<sub>van</sub>* induction, except that LB medium and 15–100  $\mu\text{M}$  vanillate as inducer were used.

### **Bacterial growth in liquid**

Cultures were grown overnight at 30 °C in an appropriate medium, back-diluted 1:50 and grown an additional overnight at 30 °C. The next day cultures were diluted 1:200 and seeded into a 96-well plate (160  $\mu\text{l}$  culture overlaid with 70  $\mu\text{l}$  of mineral oil) such that each culture had 12 replicates on the same plate and plates were replicated independently at least three times. Growth was monitored at 15 min intervals with orbital shaking on a plate reader (Biotek) at the indicated temperature. Data presented are the mean and standard deviation of all replicates.

### **Measurements of msfGFP levels with flow cytometry**

Strain ML-4048 or ML-4050 with plasmids ML-4052 to ML-4055 or ML-4058 were grown overnight at 37 °C in LB supplemented with appropriate antibiotics. Cultures were diluted 1:500 in medium supplemented with 100 ng  $\mu\text{l}^{-1}$  of aTc to induce expression of the random genes (or an EV) and grown for 30 min at 37 °C. Then, either 0.2% arabinose or 100  $\mu\text{M}$  vanillate was added to induce the expression of msfGFP. Cultures were grown an additional 4.5 h at 37 °C, then diluted 1:40 into PBS supplemented with a high concentration of kanamycin (0.5 g  $\text{l}^{-1}$ ) to stop translation and incubated at room temperature for 10 min. Fluorescence was measured on a Miltenyi MACSQuant VYB. Two independent cytometry experiments were performed for each strain and 30,000 cells were measured per replicate. FlowJo was used to analyse the data, gating on single live cells and extracting the median of the msfGFP distribution.

### **Western blot analysis of steady-state MazF(E24A)-His<sub>6</sub> levels**

Cultures were grown overnight at 37 °C in an appropriate medium, back-diluted 1:200 the next day and grown at 37 °C until optical density ( $\text{OD}_{600}$ ) ~0.2. Then, 100 ng  $\mu\text{l}^{-1}$  of aTc was added to induce *ramF* (or an EV) and cultures were grown for an additional 30 min. When needed, 100  $\mu\text{M}$  vanillate was added to induce *mazF(E24A)-His<sub>6</sub>* and cultures were grown for an additional 60 min. At  $\text{OD}_{600}$  ~0.4–0.6, 1 ml of cells was pelleted and flash-frozen. Pellets were then resuspended in 1 $\times$  Laemmli sample buffer (Bio-Rad) supplemented with  $\beta$ -mercaptoethanol normalized to the  $\text{OD}_{600}$  of the culture at the moment of collection. Samples were boiled at 95 °C for 10 min, analysed by 4%–20% SDS-polyacrylamide gel electrophoresis and transferred to a 0.2  $\mu\text{m}$  PVDF membrane. To visualize proteins, one of the following primary antibodies was used: (1) anti-His<sub>6</sub> (Invitrogen catalogue no. MA1-21315) at a final concentration of 1:1,000, (2) anti-RpoA (Biolegend catalogue no. 663104) at a final concentration of 1:5,000, (3) anti-FLAG (Sigma catalogue no. F1804) at a final concentration of 1:1,000, (4) Anti-DnaK (Abcam catalogue no. ab69617) at a final concentration of 1:1,000 and (5) Anti-Dnaj (Enzo Life Sciences catalogue no. ADI-SPA-410-F) at a final

concentration of 1:1,000. Primary antibodies were incubated overnight at 4 °C. Then, a secondary antibody was added at a final concentration of 1:15,000: (1) goat anti-mouse IgG, HRP (Invitrogen catalogue no. 32430) or (2) goat anti-rabbit IgG, HRP (Invitrogen catalogue no. 32460). SuperSignal West Femto Maximum Sensitivity Substrate (Invitrogen) was used to develop the blots. Blots were imaged by a ChemiDoc Imaging system (Bio-Rad). Images shown are one of at least three independent biological replicates. Band intensities were quantified using ImageJ (<https://imagej.nih.gov/ij>) and averages and standard errors are based on all replicates. Loading controls were performed using either an anti-RpoA (Biolegend) at a final concentration of 1:5,000 or a Coomassie stain as previously described<sup>89</sup>.

### MazF degradation assay

Cultures were grown overnight at 37 °C in an appropriate medium, back-diluted 1:200 the next day and grown at 37 °C until  $OD_{600} \sim 0.2$ . Then, 100  $\mu$ M vanillate was added to induce *mazF(E24A)-His<sub>6</sub>* and cultures were grown for an additional 60 min. Next, 100 ng  $\mu$ l<sup>-1</sup> of aTc was added to induce *ramF* (or an EV) and cultures were grown for an additional 30 min. At that point, 1 ml of cells was pelleted and flash-frozen. Then 100  $\mu$ g ml<sup>-1</sup> of tetracycline was added to block protein synthesis and samples were collected at time points 10, 20, 30 and 60 min. Immunoblots for samples were performed as described above, using RpoA as a loading control.

### Immunoprecipitation-mass spectrometry (IP-MS)

*E. coli* strains with plasmids ML-4060, ML-4075, ML-4076 or ML-4078 were grown overnight in LB supplemented with appropriate antibiotics at 37 °C. Overnight cultures were back-diluted 1:200 in 50 ml and grown until  $OD_{600} \sim 0.2$  at 37 °C. Then, 100 ng  $\mu$ l<sup>-1</sup> of aTc was added to induce FLAG-RamF or RamF or FLAG-scrambled RamF or FLAG-MazE and cultures were grown for an additional 30 min. Next, 100  $\mu$ M vanillate was added to induce MazF(E24A) and cultures were grown for additional 60 min. Cultures were pelleted at 4,000g for 10 min at 4 °C, supernatant was removed and cells were resuspended in 900  $\mu$ l of lysis buffer (B-PER II, ThermoFisher) supplemented with protease inhibitor (Roche), 1  $\mu$ l ml<sup>-1</sup> of Ready-Lyse Lysozyme Solution (Lucigen) and 1  $\mu$ l ml<sup>-1</sup> of benzonase nuclease (Sigma). Samples were incubated at room temperature for 15 min, normalized by  $OD_{600}$  and centrifuged at 15,000g for 20 min at 4 °C. Next, 850  $\mu$ l of supernatant were incubated with prewashed anti-FLAG M2 magnetic beads (Sigma) for 1 h at 4 °C with end-over-end rotation after which beads were washed three times with a wash buffer free of detergent (25 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA and 5% glycerol). On-bead reduction, alkylation and digestion were performed. Proteins were reduced with 10 mM dithiothreitol (Sigma) for 1 h at 56 °C and then alkylated with 20 mM iodoacetamide (Sigma) for 1 h at 25 °C in the dark. Proteins were then digested with modified trypsin (Promega) at an enzyme/substrate ratio of 1:50 in 100 mM ammonium bicarbonate, pH 8 at 25 °C overnight. Trypsin activity was halted by the addition of formic acid (99.9%, Sigma) to a final concentration of 5%. Peptides were desalted using Pierce Peptide Desalting Spin Columns (Thermo) and then lyophilized. The tryptic peptides were subjected to liquid chromatography with tandem mass spectrometry. Peptides were separated by reverse-phase high-performance liquid chromatography (Thermo Ultimate 3000) using a Thermo PepMap RSLC C18 column over a 90 min gradient before nano-electrospray using an Exploris mass spectrometer (Thermo). Solvent A was 0.1% formic acid in water and solvent B was 0.1% formic acid in acetonitrile. Detected peptides were mapped to *E. coli* MG1655 protein sequences with the addition of the RamF sequence and protein abundance was estimated by the number of spectrum counts. For full IP-MS results of each pull-down, see Supplementary Table 2.

### RNA extraction and sequencing

*E. coli* strains with plasmids ML-4059 or ML-4060 were grown overnight in LB supplemented with appropriate antibiotics at 37 °C. Overnight

cultures were back-diluted 1:200 in 25 ml of cultures and grown until  $OD_{600} \sim 0.2$  at 37 °C. Then, 100 ng  $\mu$ l<sup>-1</sup> of aTc was added to induce RamF or empty vector and cultures were grown for an additional 45 min. At that time, 1 ml of each culture was mixed with stop solution (110  $\mu$ l; 95% ethanol and 5% phenol) and pelleted by centrifugation for 30 s at 16,000g on a tabletop centrifuge. Pellets were flash-frozen and stored at -80 °C. Cells were lysed by adding TRIzol (Invitrogen) preheated to 65 °C directly to pellets, followed by 10 min of shaking at 65 °C and 2,000 rpm on a ThermoMixer (Eppendorf). RNA was extracted from the TRIzol mixture using Direct-zol (Zymo) columns according to manufacturer's protocol. Genomic DNA was removed by adding 2  $\mu$ l of Turbo DNase (Invitrogen) in a 100  $\mu$ l final volume using the provided buffer and incubating for 30 min at 37 °C. DNase reaction products were cleaned up with a Zymo RNA clean and concentrator kit and eluted in 25  $\mu$ l of water.

Libraries were generated as described previously<sup>37</sup>. The library generation protocol was a modified version of the paired-end strand-specific dUTP method using random hexamer primers. Ribosomal RNA was removed using a recently developed do-it-yourself *E. coli* rRNA depletion kit, using 2.5 mg of total RNA as input<sup>90</sup>. Paired-end sequencing was performed on an Illumina MiSeq at the MIT BioMicro Center.

Geneious Prime 2022.2.2 was used to map reads to the *E. coli* MG1655 genome (accession no. NC\_000913) with default parameters and to calculate transcripts per million (TPM) values for all genes. TPM values of each sample were normalized by the median TPM value of a given sample to make all samples comparable<sup>91</sup>. Data shown are based on two independent repeats for each strain. Raw data can be found with NCBI BioSample accessions SAMN32730695 and SAMN32730696.

### Microscopy

*E. coli* strains with plasmid ML-4093 and additional plasmids ML-4059, ML-4060 or ML-4074 were grown in LB supplemented with appropriate antibiotics overnight at 37 °C. Cultures were diluted 1:200, grown at 37 °C for 30 min, supplemented with 100 ng  $\mu$ l<sup>-1</sup> of aTc to induce RamF or empty vector and cells were grown for additional 30 min. Next, 0.2% arabinose was added to induce msfGFP and cells were grown for 2.5 h at 37 °C. Then 1  $\mu$ l of each culture was spotted onto a 1% agarose pad prepared with PBS and placed in a 35 mm glass-bottom dish with 20 mm microwell no. 0 coverglass (Cellvis). Phase-contrast and epifluorescence images were taken using a Hamamatsu Orca Flash 4.0 camera on a Zeiss Observer Z1 microscope using a  $\times 100/1.4$  oil immersion objective and an LED-based Colibri illumination system using MetaMorph software (Molecular Devices). Images were analysed in Fiji using the MicrobeJ plug-in<sup>92</sup>. Individual cells were identified by the phase-contrast image and fluorescence intensity was recorded for each cell, with at least 400 cells for each culture.

### Error-prone PCR mutagenesis of RamF

RamF was mutagenized using error-prone PCR-based mutagenesis, as previously described<sup>93</sup>. The gene *ramF* was amplified using Taq polymerase (NEB) and 0.5 mM MnCl<sub>2</sub> was added to the reaction as the mutagenic agent. PCR products were treated with DpnI, column purified and cloned into plasmid ML-4059 using Gibson assembly. Gibson products were transformed into DH5 $\alpha$ , yielding  $\sim 60,000$  colonies that were grown overnight at 37 °C. Overnight culture was prepped to obtain the mutagenized library, which was then electroporated into strain ML-4049 and plated on medium containing 100 ng  $\mu$ l<sup>-1</sup> of aTc and 100  $\mu$ M vanillate to induce toxin and *ramF* variants, respectively. The mutagenized library was deep-sequenced pre- and postselection to identify enriched RamF variants that inhibit MazF at a high induction level. These variants were further validated by constructing new plasmids with single, double or triple mutations on *ramF*.

## Protein structure prediction with AlphaFold2

The predicted structure of the DnaK-RamF complex was generated using AlphaFold2 (refs. 54,55), modelling both proteins as monomers with default parameters (MSA method: mmseqs2, pair mode, unpaired; number of models, 5; maximum recycles, 3).

## Reporting summary

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

## Data availability

High-throughput data generated in this study are available with NCBI BioSample accessions [SAMN32730695](#) and [SAMN32730696](#). Source data are provided with this paper.

## References

- Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14**, 827–839 (2013).
- Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
- Andersson, D. I., Jernström-Hultqvist, J. & Näsval, J. Evolution of new functions de novo and from preexisting genes. *Cold Spring Harb. Perspect. Biol.* **7**, a017996 (2015).
- McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–578 (2016).
- Vakirlis, N. et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).
- Vakirlis, N., Carvunis, A.-R. R. & McLysaght, A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**, e53500 (2020).
- Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
- Vakirlis, N., Vance, Z., Duggan, K. M. & McLysaght, A. De novo birth of functional microproteins in the human lineage. *Cell Rep.* **41**, 111808 (2022).
- Guerzoni, D. & McLysaght, A. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biol. Evol.* **8**, 1222–1232 (2016).
- Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**, e3000862 (2020).
- Weisman, C. M., Murray, A. W. & Eddy, S. R. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Curr. Biol.* **32**, 2632–2639 (2022).
- Weisman, C. M. The origins and functions of de novo genes: against all odds? *J. Mol. Evol.* **90**, 244–257 (2022).
- Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
- Neme, R. & Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife* **5**, e09977 (2016).
- Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X. & Albà, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* **2**, 890–896 (2018).
- Zhang, L. et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* **3**, 679–690 (2019).
- Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 140–146 (2020).
- diCenzo, G. C. & Finan, T. M. The divided bacterial genome: structure, function and evolution. *Microbiol. Mol. Biol. Rev.* **81**, e00019–17 (2017).
- Koonin, E. V. Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* **41**, 298–306 (2009).
- Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299 (2008).
- Tretyachenko, V. et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci. Rep.* **7**, 15449 (2017).
- Heames, B. et al. Experimental characterization of de novo proteins and their unevaluated random-sequence counterparts. *Nat. Ecol. Evol.* **7**, 570–580 (2023).
- Wang, M. S. & Hecht, M. H. A completely de novo ATPase from combinatorial protein design. *J. Am. Chem. Soc.* **142**, 15230–15234 (2020).
- Wilson, D. S., Keefe, A. D. & Szostak, J. W. The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl Acad. Sci. USA* **98**, 3750–3755 (2001).
- Spangler, L. C. et al. A de novo protein catalyzes the synthesis of semiconductor quantum dots. *Proc. Natl Acad. Sci. USA* **119**, e2204050119 (2022).
- Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
- Betanzos, C. M. et al. Bacterial glycoprofiling by using random sequence peptide microarrays. *ChemBioChem* **10**, 877–888 (2009).
- Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 0217 (2017).
- Weisman, C. M. & Eddy, S. R. Gene evolution: getting something from nothing. *Curr. Biol.* **27**, R661–R663 (2017).
- Knopp, M. & Andersson, D. I. No beneficial fitness effects of random peptides. *Nat. Ecol. Evol.* **2**, 1046–1047 (2018).
- Knopp, M. et al. De novo emergence of peptides that confer antibiotic resistance. *mBio* <https://doi.org/10.1128/mbio.00837-19> (2019).
- Knopp, M. et al. A novel type of colistin resistance genes selected from random sequence space. *PLoS Genet.* **17**, e1009227 (2021).
- Babina, A. M. et al. Rescue of *Escherichia coli* auxotrophy by de novo small proteins. *eLife* **12**, e78299 (2023).
- Digianantonio, K. M. & Hecht, M. H. A protein constructed de novo enables cell growth by altering gene regulation. *Proc. Natl Acad. Sci. USA* **113**, 2400–2405 (2016).
- Hoegler, K. J. & Hecht, M. H. A de novo protein confers copper resistance in *Escherichia coli*. *Protein Sci.* **25**, 1249–1259 (2016).
- Mutalik, V. K. et al. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
- Culviner, P. H. & Laub, M. T. Global analysis of the *E. coli* toxin MazF reveals widespread cleavage of mRNA and the inhibition of rRNA maturation and ribosome biogenesis. *Mol. Cell* **70**, 868–880 (2018).
- Culviner, P. H., Nokedal, I., Fortune, S. M. & Laub, M. T. Global analysis of the specificities and targets of endoribonucleases from *Escherichia coli* toxin-antitoxin systems. *mBio* **12**, e0201221 (2021).
- Brielle, R., Pinel-Marie, M. L. & Felden, B. Linking bacterial type I toxins with their actions. *Curr. Opin. Microbiol.* **30**, 114–121 (2016).
- Ochman, H. & Selander, R. K. Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* **157**, 690–693 (1984).
- Zorzini, V. et al. Substrate recognition and activity regulation of the *Escherichia coli* mRNA endonuclease MazF. *J. Biol. Chem.* **291**, 10950–10960 (2016).

42. Li, G. Y. et al. Characterization of dual substrate binding sites in the homodimeric structure of *Escherichia coli* mRNA interferase MazF. *J. Mol. Biol.* **357**, 139–150 (2006).
43. Kamada, K., Hanaoka, F. & Burley, S. K. Crystal structure of the MazE/MazF complex: molecular bases of antidote–toxin recognition. *Mol. Cell* **11**, 875–884 (2003).
44. Ogura, T. et al. Balanced biosynthesis of major membrane components through regulated degradation of the committed enzyme of lipid A biosynthesis by the AAA protease FtsH (HflB) in *Escherichia coli*. *Mol. Microbiol.* **31**, 833–844 (1999).
45. Zorzini, V. et al. *Escherichia coli* antitoxin MazE as transcription factor: Insights into MazE–DNA binding. *Nucleic Acids Res.* **43**, 1241–1256 (2015).
46. Schramm, F. D., Schroeder, K. & Jonas, K. Protein aggregation in bacteria. *FEMS Microbiol. Rev.* **44**, 54–72 (2019).
47. Tomoyasu, T., Mogk, A., Langen, H., Goloubinoff, P. & Bukau, B. Genetic dissection of the roles of chaperones and proteases in protein folding and degradation in the *Escherichia coli* cytosol. *Mol. Microbiol.* **40**, 397–413 (2001).
48. Mogk, A., Deuerling, E., Vorderwülbecke, S., Vierling, E. & Bukau, B. Small heat shock proteins, ClpB and the DnaK system form a functional triade in reversing protein aggregation. *Mol. Microbiol.* **50**, 585–595 (2003).
49. Chapman, E. et al. Global aggregation of newly translated proteins in an *Escherichia coli* strain deficient of the chaperonin GroEL. *Proc. Natl Acad. Sci. USA* **103**, 15800–15805 (2006).
50. Zhao, K., Liu, M. & Burgess, R. R. The global transcriptional response of *Escherichia coli* to induced  $\sigma$ 32 protein involves  $\sigma$ 32 regulon activation followed by inactivation and degradation of  $\sigma$ 32 in vivo. *J. Biol. Chem.* **280**, 17758–17768 (2005).
51. Patra, M., Roy, S. S., Dasgupta, R. & Basu, T. GroEL to DnaK chaperone network behind the stability modulation of  $\sigma$ 32 at physiological temperature in *Escherichia coli*. *FEBS Lett.* **589**, 4047–4052 (2015).
52. Nonaka, G., Blankschien, M., Herman, C., Gross, C. A. & Rhodius, V. A. Regulon and promoter analysis of the *E. coli* heat-shock factor,  $\sigma$ 32, reveals a multifaceted cellular response to heat stress. *Genes Dev.* **20**, 1776–1789 (2006).
53. Yura, T. et al. Analysis of  $\sigma$ 32 mutants defective in chaperone-mediated feedback control reveals unexpected complexity of the heat shock response. *Proc. Natl Acad. Sci. USA* **104**, 17638–17643 (2007).
54. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
55. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
56. Mayer, M. P., Rudiger, S. & Bukau, B. Molecular basis for interactions of the DnaK chaperone with substrates. *Biol. Chem.* **381**, 877–885 (2000).
57. Zhu, X. et al. Structural analysis of substrate binding by the molecular chaperone DnaK. *Science* **272**, 1606–1614 (1996).
58. Pellecchia, M. et al. Structural insights into substrate binding by the molecular chaperone DnaK. *Nat. Struct. Biol.* **7**, 298–303 (2000).
59. Zutz, A. et al. A dual-reporter system for investigating and optimizing protein translation and folding in *E. coli*. *Nat. Commun.* **12**, 6093 (2021).
60. Govers, S. K., Mortier, J., Adam, A. & Aertsen, A. Protein aggregates encode epigenetic memory of stressful encounters in individual *Escherichia coli* cells. *PLoS Biol.* **16**, e2003853 (2018).
61. Bittner, L. M., Arends, J. & Narberhaus, F. When, how and why? Regulated proteolysis by the essential FtsH protease in *Escherichia coli*. *Biol. Chem.* **398**, 625–635 (2017).
62. Führer, F., Langklotz, S. & Narberhaus, F. The C-terminal end of LpxC is required for degradation by the FtsH protease. *Mol. Microbiol.* **59**, 1025–1036 (2006).
63. Herman, C., Thévenet, D., Bouloc, P., Walker, G. C. & D’Ari, R. Degradation of carboxy-terminal-tagged cytoplasmic proteins by the *Escherichia coli* protease HflB (FtsH). *Genes Dev.* **12**, 1348–1355 (1998).
64. Bittner, L. M., Westphal, K. & Narberhaus, F. Conditional proteolysis of the membrane protein YfgM by the FtsH protease depends on a novel N-terminal degron. *J. Biol. Chem.* **290**, 19367–19378 (2015).
65. Ruiz-Orera, J. & Albà, M. M. Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet.* **35**, 186–198 (2019).
66. Couso, J. P. & Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **18**, 575–589 (2017).
67. Storz, G., Wolf, Y. I. & Ramamurthi, K. S. Small proteins can no longer be ignored. *Annu. Rev. Biochem.* **83**, 753–777 (2014).
68. Orr, M. W., Mao, Y., Storz, G. & Qian, S. B. Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.* **48**, 1029–1042 (2020).
69. Weaver, J., Mohammad, F., Buskirk, A. R. & Storz, G. Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio* **10**, e02819–18 (2019).
70. Stringer, A., Smith, C., Mangano, K. & Wade, J. T. Identification of novel translated small open reading frames in *Escherichia coli* using complementary ribosome profiling approaches. *J. Bacteriol.* **204**, JB0035221 (2022).
71. Sberro, H. et al. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* **178**, 1245–1259 (2019).
72. Miravet-Verde, S. et al. Unraveling the hidden universe of small proteins in bacterial genomes. *Mol. Syst. Biol.* **15**, e8290 (2019).
73. Luo, Y., Zhang, T. & Wu, H. The transport and mediation mechanisms of the common sugars in *Escherichia coli*. *Biotechnol. Adv.* **32**, 905–919 (2014).
74. Schleif, R. AraC protein, regulation of the L-arabinose operon in *Escherichia coli* and the light switch mechanism of AraC action. *FEMS Microbiol. Rev.* **34**, 779–796 (2010).
75. Schleif, R. Regulation of the L-arabinose operon of *Escherichia coli*. *Trends Genet.* **16**, 559–565 (2000).
76. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–592 (2005).
77. Keren, L. et al. Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell* **166**, 1282–1294 (2016).
78. Jurénas, D., Fraikin, N., Goormaghtigh, F. & van Melderen, L. Biology and evolution of bacterial toxin–antitoxin systems. *Nat. Rev. Microbiol.* **20**, 335–350 (2022).
79. Burga, A., Ben-David, E. & Kruglyak, L. Toxin–antidote elements across the tree of life. *Annu. Rev. Genet.* **54**, 387–415 (2020).
80. Loris, R. & Garcia-Pino, A. Disorder- and dynamics-based regulatory mechanisms in toxin–antitoxin modules. *Chem. Rev.* **114**, 6933–6947 (2014).
81. Garcia-Pino, A. et al. Doc of prophage P1 is inhibited by its antitoxin partner Phd through fold complementation. *J. Biol. Chem.* **283**, 30821–30827 (2008).
82. de Gieter, S. et al. The intrinsically disordered domain of the antitoxin Phd chaperones the toxin Doc against irreversible inactivation and misfolding. *J. Biol. Chem.* **289**, 34013–34023 (2014).
83. Cherny, I. & Gazit, E. The YefM antitoxin defines a family of natively unfolded proteins: implications as a novel antibacterial target. *J. Biol. Chem.* **279**, 8252–8261 (2004).
84. Snead, K. J., Moore, L. L. & Bourne, C. R. ParD antitoxin hotspot alters a disorder-to-order transition upon binding to its cognate ParE toxin, lessening its interaction affinity and increasing its protease degradation kinetics. *Biochemistry* **61**, 34–45 (2022).

85. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146 (2017).
86. Kosinski, L. J., Aviles, N. R., Gomez, K. & Masel, J. Random peptides rich in small and disorder-promoting amino acids are less likely to be harmful. *Genome Biol. Evol.* **14**, evac085 (2022).
87. Khlebnikov, A., Datsenko, K. A., Skaug, T., Wanner, B. L. & Keasling, J. D. Homogeneous expression of the P(BAD) promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology* **147**, 3241–3247 (2001).
88. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
89. Welinder, C. & Ekblad, L. Coomassie staining as loading control in western blot analysis. *J. Proteome Res* **10**, 1416–1419 (2011).
90. Culviner, P. H., Guegler, C. K. & Laub, M. T. A simple, cost-effective and robust method for rRNA depletion in RNA-sequencing studies. *mBio* **11**, e00010–20 (2020).
91. Dillies, M. A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).
92. Ducret, A., Quardokus, E. M. & Brun, Y. V. MicrobeJ, a tool for high throughput bacterial cell detection and quantitative analysis. *Nat. Microbiol* **1**, 16077 (2016).
93. Srikant, S., Gaudet, R. & Murray, A. W. Selecting for altered substrate specificity reveals the evolutionary flexibility of ATP-binding cassette transporters. *Curr. Biol.* **30**, 1689–1702 (2020).

## Acknowledgements

We thank the MIT BioMicro Center and its staff for their support in sequencing; the MIT Biopolymers and Proteomics Core and its staff for their help in mass spectrometry experiments; D. Ding and C. McClune for help with library construction; K. Gozzi, S. Mendoza, A. Murray, S. Srikant and C. Vassallo for comments on the manuscript; P. DeWeirdt, C. Doering, K. Forsberg, M. Guzzo, M. LeRoux, C. Weisman, T. Zhang and all members of the Laub laboratory for helpful discussions. I.F. was supported by a long-term fellowship (LT000706/2018) from the Human Frontier Science Program. M.T.L. is an Investigator of the Howard Hughes Medical Institute.

## Author contributions

I.F. and M.T.L. conceived the project and wrote the manuscript. I.F. designed and performed all experiments and data analysis. M.T.L. supervised the project.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41559-023-02224-4>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41559-023-02224-4>.

**Correspondence and requests for materials** should be addressed to Michael T. Laub.

**Peer review information** *Nature Ecology & Evolution* thanks Gisela Storz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

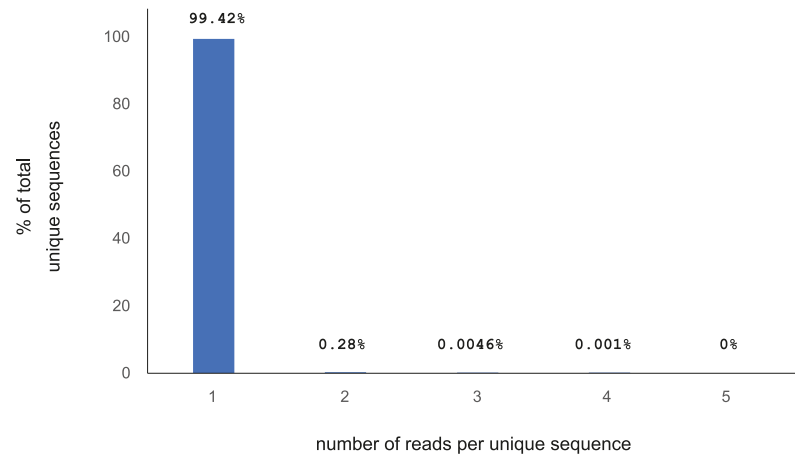
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

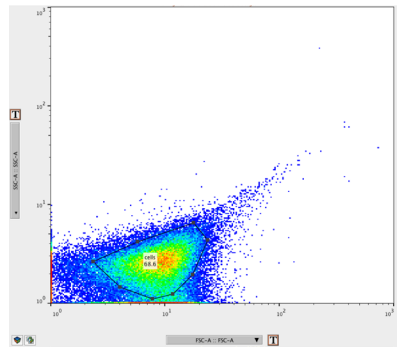
© The Author(s) 2023

A

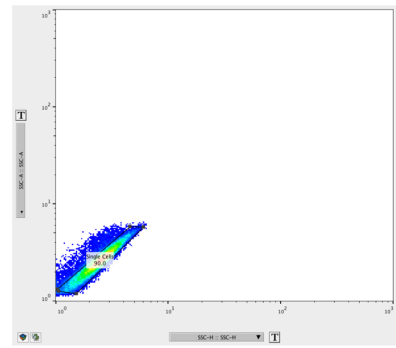


B

Step 1



Step 2

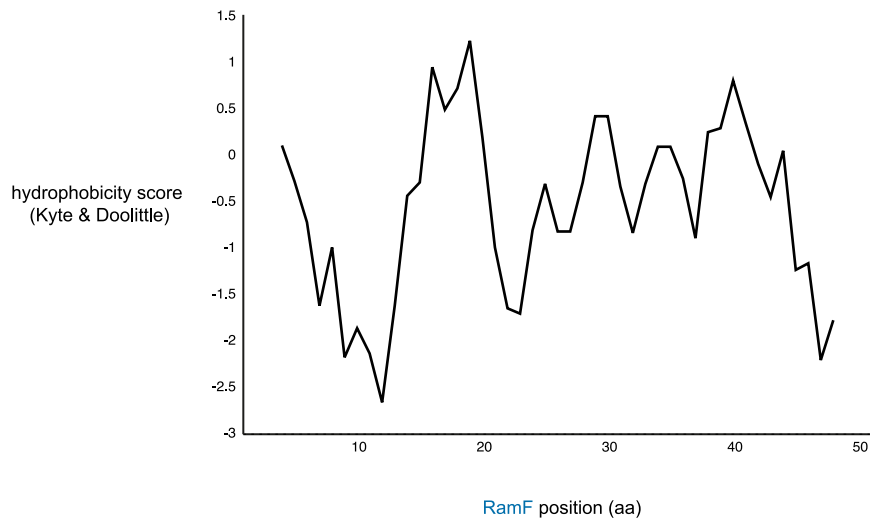


**Extended Data Fig. 1 | Read counts of the library preselection.** (A) Number of reads per unique sequence based on deep sequencing of the random protein library preselection. The total read count was ~300,000. (B) Example for the

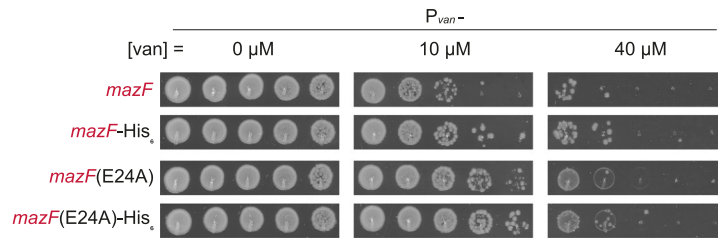
cytometer gating strategy used throughout this study: First, differentiation between cells and non-cell events using SSC-A and FSC-A parameters. Second, selection of singlet cell events using SSC-A and SSC-H parameters.

**A**

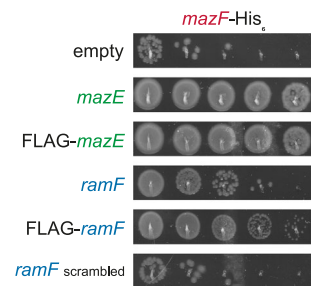
<i>E. coli</i> small proteins (<100 aa)		n=	181	RamF	80
			cytosolic		membrane
% hydrophobic	A		8.7	5.9	9.7
	I		6.0	2.0	8.4
	L		9.2	11.8	13.1
	M		2.0	3.9	3.3
	F		3.2	5.9	5.5
	W		1.1	3.9	2.1
	Y		2.1	5.9	2.4
% charged	V		7.2	3.9	8.4
	R		6.4	5.7	4.4
	H		2.5	2.0	1.5
	K		7.0	3.9	5.1
	D		5.5	0.0	3.3
% polar	E		8.2	2.0	3.1
	S		5.6	1.8	5.3
	T		5.2	5.9	5.3
	N		4.0	7.8	3.2
	Q		4.8	0.0	3.2
	C		1.6	0.0	1.9
	G		6.0	2.0	7.1
	P		3.6	5.9	3.8

**B**

**Extended Data Fig. 2 | Amino acid composition of RamF.** (A) The amino acid composition of RamF compared to cytosolic (n = 181) and membrane (n = 80) proteins in MG1655 *E. coli* whose lengths are each <100 amino acids. (B) Hydrophobicity plot for RamF based on Kyte & Doolittle scale and an average window size of seven amino acids.

**A**

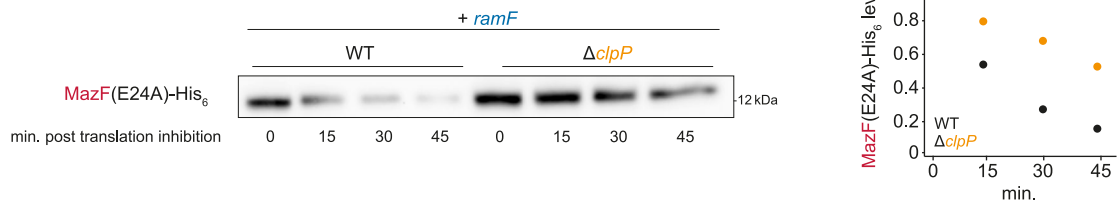
**Extended Data Fig. 3 | Epitope-tagging of MazF and RamF does not interfere with their functions.** (A) 10-fold serial dilution spotting of cells expressing *mazF*, *mazF*-His<sub>6</sub>, *mazF*(E24A), or *mazF*(E24A)-His<sub>6</sub> from  $P_{van^-}$ . (B) 10-fold serial

**B**

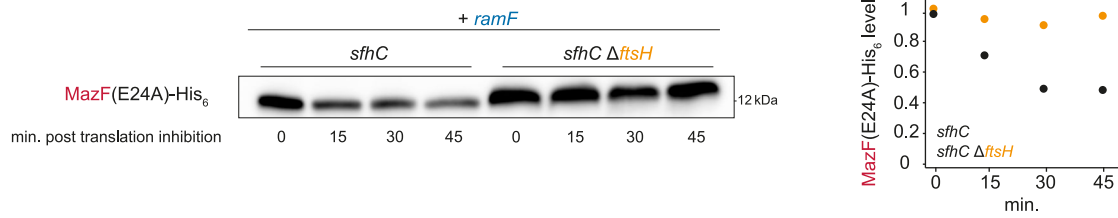
dilution spotting of cells expressing *mazF*-His<sub>6</sub> from  $P_{van^-}$ . Cells were additionally expressing *mazE*, FLAG-*mazE*, *ramF*, FLAG-*ramF*, scrambled *ramF*, or an empty vector.



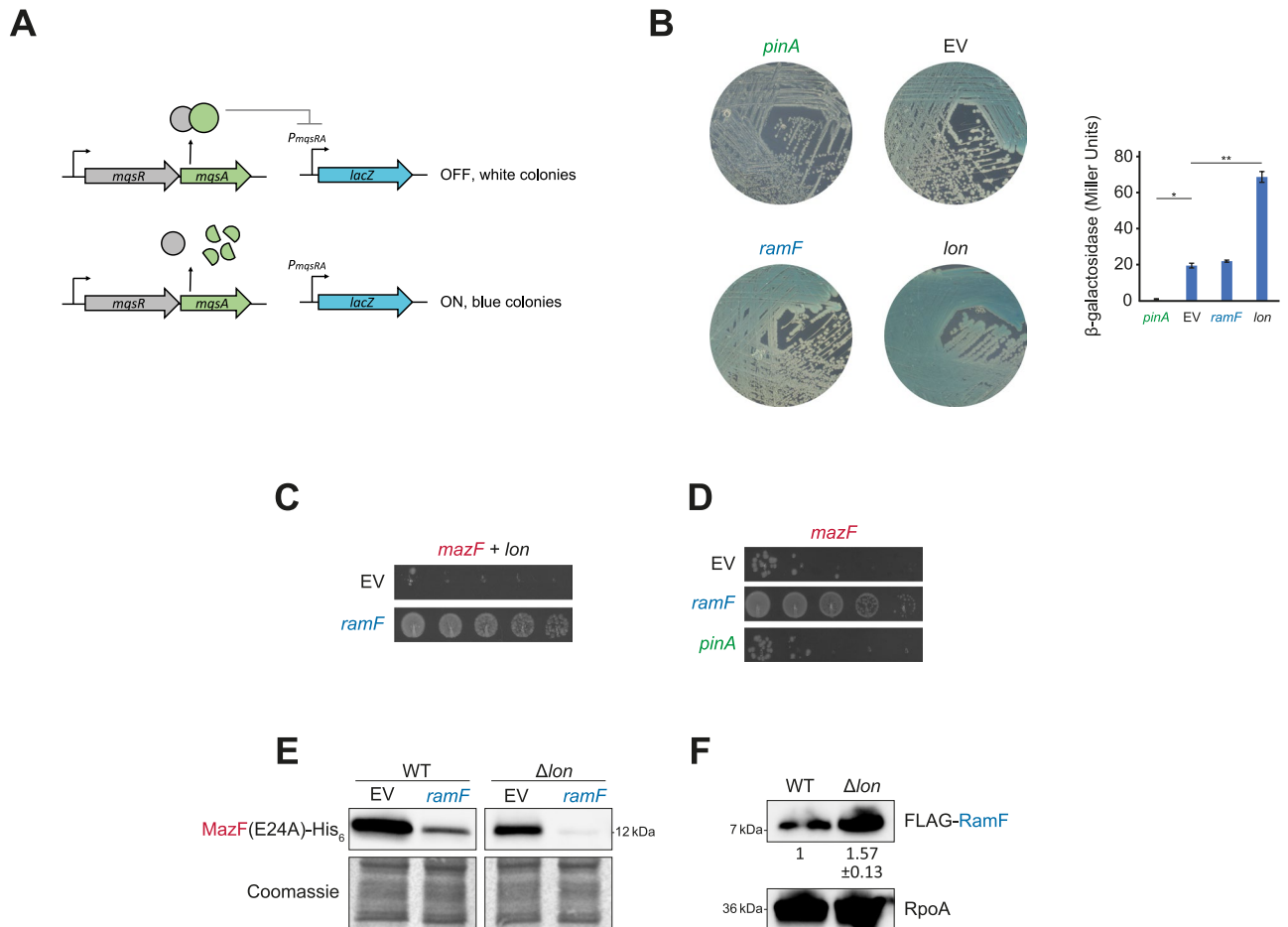
A



B

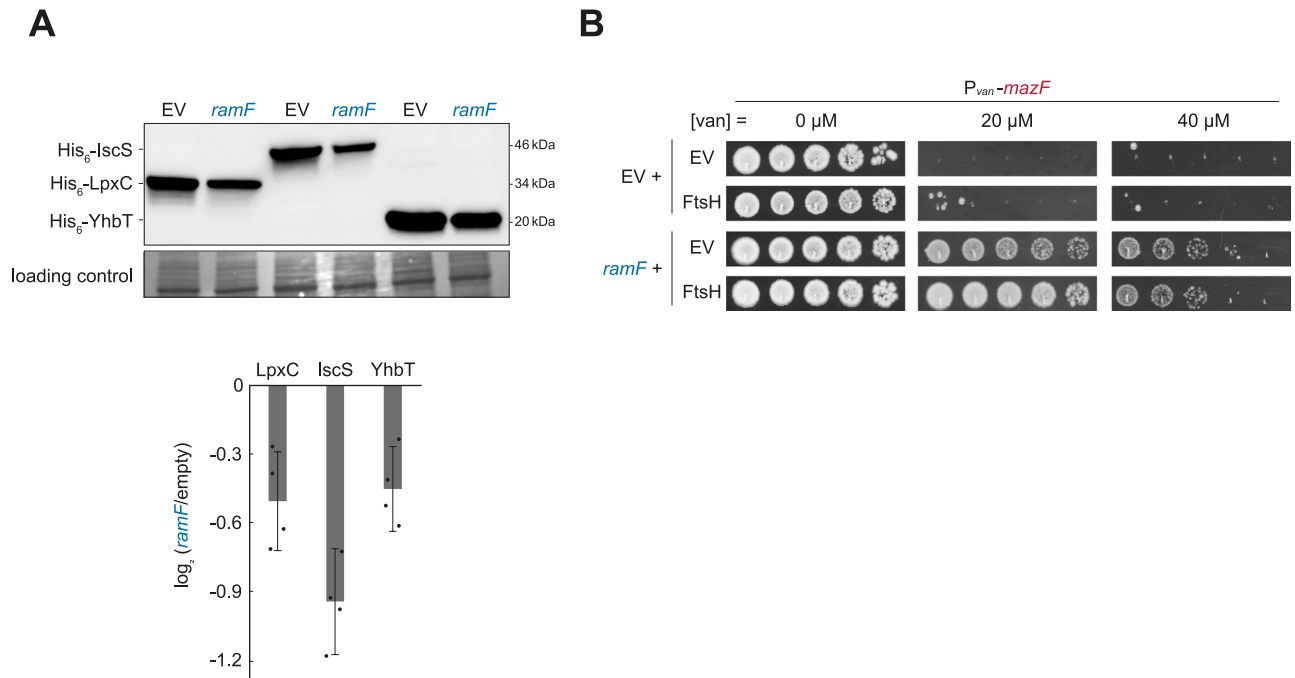


**Extended Data Fig. 4 | MazF(E24A)-His<sub>6</sub> proteolytic in  $\Delta clpP$  and  $\Delta ftsH$  cells.** Immunoblot of MazF(E24A)-His<sub>6</sub>, expressed from  $P_{van}$ , from cells expressing *ramF* in (A)  $\Delta clpP$  or (B)  $\Delta ftsH$  cells. Time points were taken after the addition of tetracycline to stop the translation of new proteins. Quantification is based on two biological repeats and MazF(E24A)-His<sub>6</sub> levels are normalized to t = 0.



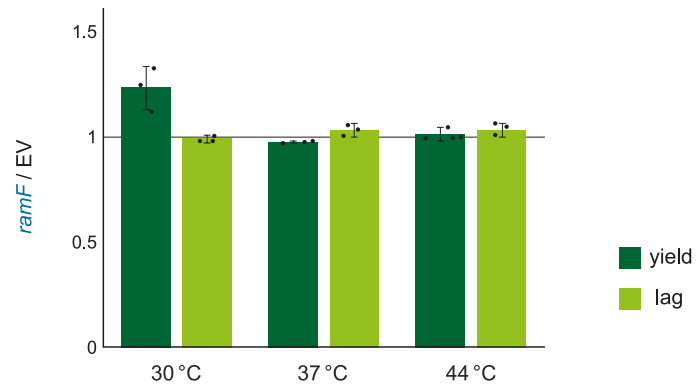
**Extended Data Fig. 5 | RamF does not inhibit the Lon protease.** (A) A system to measure *in vivo* activity of Lon protease: The MqsRA complex inhibits the  $P_{mqsRA}$  promoter driving a *lacZ* reporter. In this system, Lon activity is correlated to LacZ production levels because the antitoxin MqsA is a Lon substrate and upon antitoxin degradation, LacZ is produced and colonies turn blue. (B) Cells harbouring the system described in (A) also expressing *lon*, *ramF*, *pinA* (a known Lon inhibitor), or an empty vector. Quantification of  $\beta$ -galactosidase activity in each strain is based on the mean of  $n = 3$  biological repeats of cells growing at 30 °C overnight. \* $P = 7.15 \times 10^{-6}$ , \*\* $P = 6.36 \times 10^{-6}$  based on a two-sided t-test and error bars represent SD. (C) 10-fold serial dilution spotting of cells expressing *mazF*,

overexpressing *lon* and additionally expressing *ramF* or an empty vector. (D) 10-fold serial dilution spotting of cells expressing *mazF* and additionally expressing *ramF*, *pinA*, or an empty vector. (E) Immunoblot of MazF(E24A)-His<sub>6</sub> expressed from  $P_{van}$  in control cells or cells lacking the protease Lon. Cells additionally expressing *ramF* or harbouring an empty vector. Loading control is based on Coomassie staining of total protein. Results represent  $n = 3$  biological repeats. (F) Immunoblot of FLAG-RamF expressed from  $P_{tet}$  in control cells or cells lacking the protease Lon. Loading control is based on RpoA and quantification is based on two repeats.



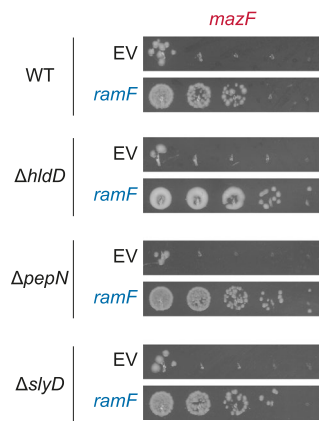
**Extended Data Fig. 6 | Overproduction of FtsH is insufficient to inhibit MazF and does not alter RamF efficiency as a MazF inhibitor.** (A) Immunoblot of His<sub>6</sub>-IscS, His<sub>6</sub>-LpxC, or His<sub>6</sub>-YhbT, known FtsH substrates, from cells co-expressing *ramF* or harbouring an empty vector. Loading control is based on

Coomassie staining of total protein. Bar: error bars represent SD based  $n = 4$  biological repeats and each black dot is an individual measurement. (B) 10-fold serial dilution spotting of cells co-expressing (i) *mazF*, (ii) empty vector or *ramF* and (iii) empty vector or *ftsH*.

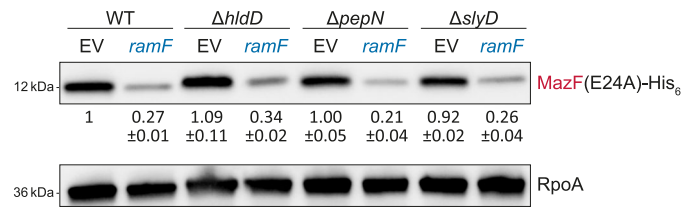


**Extended Data Fig. 7 | Growth characteristics of cells producing RamF.** Lag time (time to reach  $OD_{600} = 0.2$ ) and culture yield (final  $OD_{600}$ ) ratios between cells producing RamF to empty vector at the growth temperatures indicated. Error bars represent SD based on  $n = 3$  biological repeats and each black dot is an individual measurement.

A

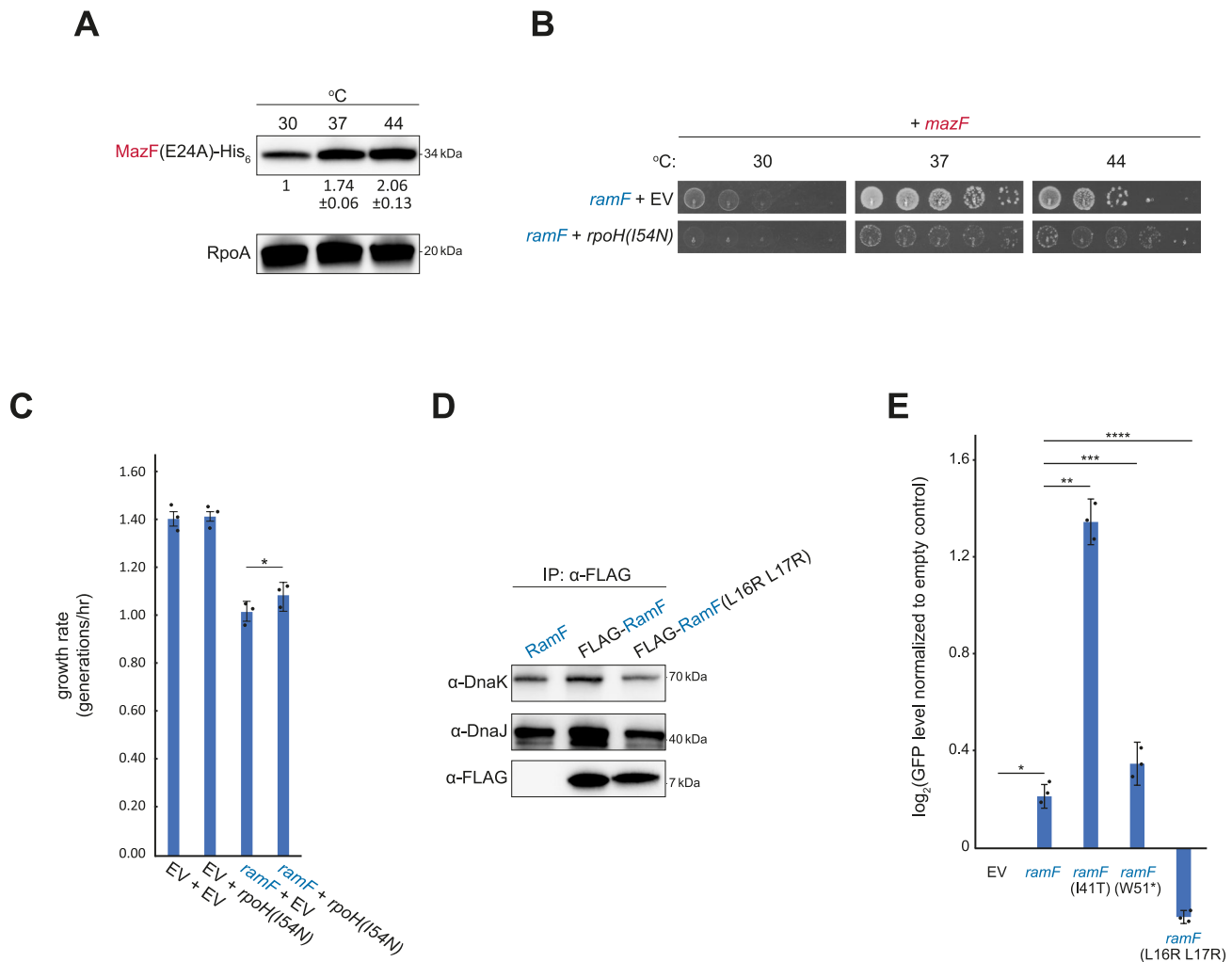


B



**Extended Data Fig. 8 | Non-chaperone proteins that interact with RamF do not affect its ability to inhibit MazF.** (A) 10-fold serial dilution spotting of cells expressing *mazF* in addition to either empty vector or *ramF* in a genetic background of  $\Delta hldD$ ,  $\Delta pepN$ ,  $\Delta slyD$ , or control cells. (B) Immunoblot of

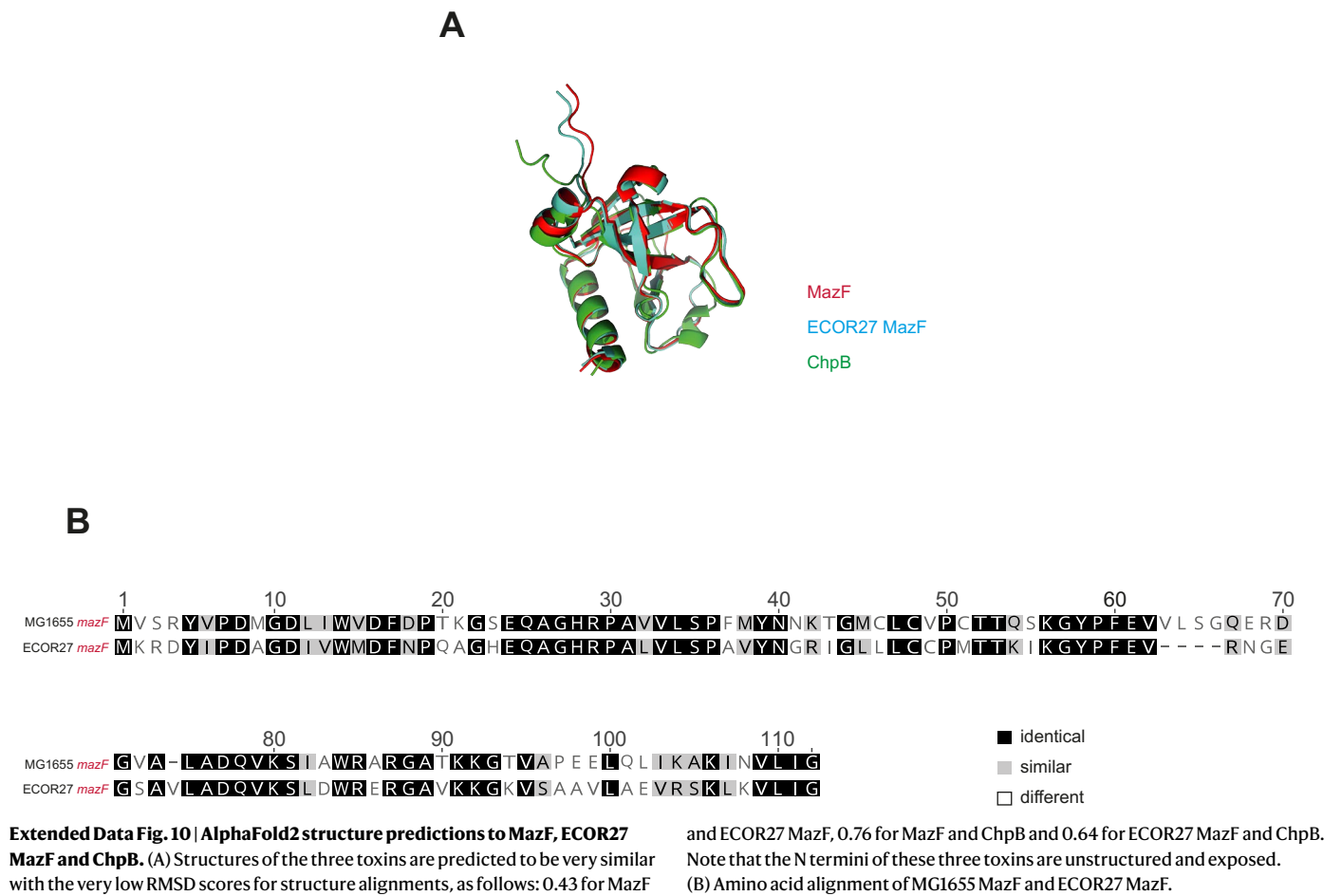
MazF(E24A)-His<sub>6</sub> expressed from  $P_{ram}$  in  $\Delta hldD$ ,  $\Delta pepN$ ,  $\Delta slyD$ , or control cells. Cells additionally express *ramF* or harbour an empty vector. Loading control is based on RpoA and quantification is based on two repeats.



### Extended Data Fig. 9 | MazF and RamF relationship with cellular chaperones.

(A) MazF(E24A)-His<sub>6</sub> steady-state levels increase with temperature. Immunoblot of MazF(E24A)-His<sub>6</sub> expressed from *P<sub>van</sub>* in control cells at growth temperatures 30, 37 and 44 °C. Loading control is based on RpoA and quantification is based on three biological repeats. (B) Overproduction of RpoH(I54N) increases MazF toxicity at a range of temperatures. 10-fold serial dilution spotting of cells expressing *mazF* from *P<sub>van</sub>*. Cells also express combinations of *ramF*, *rpoH*(I54N), or empty vectors, as indicated and were grown at 30, 37, or 44 °C. (C) Overproduction of RpoH(I54N) alleviates growth rate defect of RamF production. Maximal growth rates (generations per hour) of cells producing combinations of *ramF*, *rpoH*(I54N), or empty vectors, as indicated and grown at 30 °C. Quantification is based on *n* = 3 biological repeats. \**P* = 0.04 based on

a one-sided t-test, error bars represent SD and each black dot is an individual measurement. (D) Substitutions L16R and L17R reduce the interaction between RamF and DnaK/J. Cells producing RamF, FLAG-RamF, or FLAG-RamF(L16R L17R) were lysed and used as input for immunoprecipitation using α-FLAG beads. Eluates were then blotted with α-FLAG, α-DnaK and α-DnaJ antibodies. Results represent *n* = 2 biological repeats. (E) RamF and its variants increase protein aggregation levels. Measurements of msfGFP levels expressed from the *P<sub>ibpA</sub>* promoter, whose activity correlates with aggregation levels in *E. coli* cells<sup>53,54</sup>. Cells additionally express *ramF*, *ramF*(I41T), *ramF*(W51\*), or *ramF*(L16R L17R). Values are normalized to empty vector control and are based the mean of *n* = 3 biological repeats. \**P* = 0.003, \*\**P* = 0.00001, \*\*\**P* = 0.07, \*\*\*\**P* = 0.00012, based on a two-sided t-test, error bars represent SD.



## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

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

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

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection MetaMorph (v7.10.2.240) (Molecular Devices LLC) was used to collect microscopy data.  
Biotek Gen5 (v3.02) was used to collect growth curve data.

Data analysis ImageJ (v1.53) and MicrobeJ (v5.13l) were used for image analyses.  
PEAR (v0.9.11) and USEARCH (v11.0.667) were used for Illumina read clustering.  
FlowJo (v10.0.7) was used for cytometry data analyses.  
Geneious Primer (v2022.2.2) was used for RNA-seq analysis.

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



## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

High-throughput data generated in this study is available with NCBI BioSample accessions SAMN32730695 and SAMN32730696.

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

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

Sample size

Sample sizes were chosen based on the number needed to reliably determine differences between groups. All experiments were performed 2-4 times independently.

Data exclusions

No data exclusion was performed.

Replication

All experimental findings were repeated at least twice. Exact biological repeats are indicated. All reported results were successfully reproduced.

Randomization

No experimental groups or control groups were subjectively chosen and there are no covariates to control for as experiments were done in isogenic strains. No experiments required randomization.

Blinding

Blinding was not relevant because all data were obtained objectively and had strong effect sizes.

## Reporting for specific materials, systems and methods

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

## Materials &amp; experimental systems

n/a	Involvement in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

## Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used

6x-His Tag Monoclonal Antibody (clone HIS.H8, Invitrogen Cat#: MA1-21315).  
 Anti-RpoA antibody (Biolegend Cat#: 663104).  
 Anti-FLAG antibody (Sigma Cat#: F1804).  
 Anti-DnaK antibody (Abcam Cat#: ab69617).  
 Anti-DnaJ antibody (Enzo Life Sciences Cat#: ADI-SPA-410-F).  
 Goat anti-Mouse IgG Secondary Antibody, HRP (Invitrogen Cat#: 32430).  
 Goat anti-Rabbit IgG Secondary Antibody, HRP (Invitrogen Cat#: 32460).

Validation

All antibodies used in this study are commercial, standard antibodies routinely used in bacterial studies. Manufacturers specify that antibody reactivity is determined by testing in at least one approved application (e.g., western blot). Antibodies were used according to the manufacturer's guidelines. We performed internal validations of antibodies against negative control strains.

## Flow Cytometry

## Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

## Methodology

Sample preparation

Strain ML-4048 or ML-4050 with plasmids ML-4052 to ML-4055 or ML-4058 were grown overnight at 37 °C in LB supplemented with appropriate antibiotics. Cultures were diluted 1:500 in medium supplemented with 100 ng/μL aTc to induce expression of the random genes (or an empty vector) and grown for 30 minutes at 37 °C. Then, either 0.2% arabinose or 100 μM vanillate was added to induce the expression of msfGFP. Cultures were grown an additional 4.5 hours at 37 °C, then diluted 1:40 into PBS supplemented with a high concentration of (0.5 g/L) kanamycin to stop translation, and incubated at room temperature for 10 min. Then, fluorescence was measured on a Miltenyi MACSQuant VYB.

Instrument

Miltenyi MACSQuant VYB (Miltenyi Biotec)

Software

FlowJo

Cell population abundance

30,000 cells were measured per replicate and at least 18,000 were left post-gating.

Gating strategy

Initial gating was performed using SSC-A and FSC-A and then single bacterial cells were identified using parameters SSC-A and SSC-H.

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