A protein language model unveils the *E. coli* pangenome functional landscape regulating host proteostasis

Daniel Martinez-Martinez[1,2,6,*], Andreea Aprodu[3,6], Cassandra Backes[1,2,3,6,†], Franziska Ottens[3], Aleksandra Zečić[4], Hannah M. Doherty[3], Jonas Widder[3], Andrew Ivan[1], Laurence Game[1], Iliyana Kaneva[1], Georgia Roumellioti[1], Alex Montoya[1], Holger Kramer[1], Thorsten Hoppe[3], Filipe Cabreiro[1,2,3,5,7,*]

1. MRC Laboratory of Medical Sciences (LMS), Du Cane Road, London, W12 0HS, UK
2. Institute of Clinical Sciences, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London, W12 0HS, UK
3. University of Cologne, Faculty of Mathematics and Natural Sciences, Cluster of Excellence Cellular Stress Responses in Aging-associated Diseases (CECAD), 50931 Cologne, Germany.
4. University of Cologne, Faculty of Medicine, Cluster of Excellence Cellular Stress Responses in Aging-associated Diseases (CECAD), 50931 Cologne, Germany.
5. Center for Molecular Medicine Cologne (CMMC) - University of Cologne, Robert-Koch Str. 21, 50931 Cologne, Germany.
6. These authors contributed equally to this work.
7. Lead Contact

*Correspondence: d.martinez@lms.mrc.ac.uk, f.cabreiro@uni-koeln.de

†. Current address: Laboratoire de Bioénergétique et Ingénierie des protéines, CNRS, 31 chemin Joseph Aiguier, 13009 Marseille, France.

ABSTRACT

Understanding how bacterial diversity at strain level resolution shapes host physiology is a central challenge in microbiome research. The vast, functionally unknown genetic diversity within a species pangenome makes it difficult to connect genes to function and their impact on host physiology. Here, we explore how the functional landscape of the *Escherichia coli* pangenome impacts transcriptional responses in *Caenorhabditis elegans* and show that traditional gene-centric methods fail to provide significant functional associations with the host. Thus, we developed a pangenome framework that leverages the protein language model ProtT5 and generates unique strain embeddings representing the functional potential of each 9,558 *E. coli* isolate. Stratification of the pangenome into distinct functional guilds aligned with key host processes such as cell division, metabolism and proteostasis. Further, we identify a critical interplay between the extensive network of bacterial chaperones and proteases in regulating host proteostasis. We find that the bacterial chaperone DNAK/HSP70 and protease ClpX fine-tune the host ubiquitin-proteasome system by controlling propionate and vitamin B12 availability. These findings reveal a conserved ‘co-proteostasis’ mechanism as a key phenomenon modulating host-microbe interactions through metabolic communication. Our pangenome-to-phenotype approach offers a powerful strategy to decode bacterial pangenome functional diversity, directly linking microbial genomic variation to host physiological outcomes.

INTRODUCTION

The metabolic capacity of the host is vastly expanded by its resident microbiome, yet correlating specific microbial signatures with physiological outcomes remains a fundamental challenge. Animal models such as *Caenorhabditis elegans* have been successfully repurposed as biosensors to study the mechanisms underlying host-microbe interactions[1-3]. However, while strain-specific impacts on host physiology are increasingly recognized, the vast genetic diversity within individual bacterial species remains largely underexplored. Current strategies relying on phylogenetic markers or linear reference genomes fail to fully capture this functional potential, leaving a gap in our ability to predict how intra-species genetic heterogeneity drives distinct host phenotypes[4]. The *E. coli* pangenome represents a vast reservoir of uncharted metabolic potential given its ecological ubiquity and open genomic architecture[5,6]. For instance, distinct *E. coli* strains elicit divergent host responses through differential production of metabolites, such as vitamin B12 or betaine[7-9]. Yet, standard laboratory strains often used to study these interactions capture only a fraction of this natural diversity[10]. Consequently, we require analytical frameworks that move beyond sequence identity to capture the latent functional potential of bacterial proteomes and map them directly to host physiology, thereby bridging the gap between reductionist models and the complex reality of natural microbiomes[11].

Here, we bridge this gap by combining high-throughput transcriptomics of *C. elegans* with a novel machine-learning approach that utilizes the protein language model (pLM) ProtT5[12] to recreate the functional landscape of a pangenome of 9,558 *E. coli* assemblies. We integrated the geometrical representation from the pLM with the genetic background of the strains to generate strain embeddings, vector representations encapsulating the total functional potential of a bacterial strain. By exposing *C. elegans* to a diverse library of 592 *E. coli* strains, we demonstrate that the geometry of the bacterial embedding space accurately predicts host phenotypic variance, revealing a profound link between the microbial pangenome and host proteostasis. We identify a cross-domain co-proteostasis mechanism where the bacterial chaperone network (DnaK/ClpX) regulates vitamin B12 and propionate metabolism, dictating metabolic rewiring in the host through B12-dependent or independent metabolic shunt that regulates host ubiquitin-proteasome system (UPS) function.

RESULTS

The *E. coli* pangenome elicits a vast range of transcriptomic profiles in *C. elegans*

E. coli is known to have an open pangenome[6], meaning that the various strains within the *E. coli* species contain unique genes coding for proteins whose functions are essential for their distinct functional properties. We hypothesized that the extensive genetic variation within the *E. coli* pangenome dictates host responses. To interrogate these host responses to individual bacterial strains, we generated high-resolution bulk RNAseq transcriptional profiles for 592 distinct *E. coli* – *C. elegans* mono-association pairs (Fig. 1a). We curated a library combining the EcoRef collection and additional strains with broad phylogenetic coverage (Fig. 1b)[6,10]. This panel spans the major *E. coli* phylogroups, evolutionary lineages defined by specific gene markers that are traditionally linked to distinct ecological roles and primarily comprises commensal strains isolated from human and animal hosts (Fig. 1c; Extended Data Fig. 1a), with roughly 50% belonging to phylogroup B2. Analysis of the strain genomes confirmed an open pangenome architecture: a conserved core of 3,265 gene families (>95% presence), a shell genome of 2,589 gene families (15% - 95% presence), and a diverse and large cloud genome of 20,113 rare gene families (<15% presence) (Fig. 1d). This distribution highlights the immense reservoir of genetic diversity available to influence host physiology. Next, we profiled the host response by raising synchronized *C. elegans* (N2) on each bacterial strain and sequencing total RNA from Day 1 adults. Following rigorous quality filtering and batch effect correction (Extended Data Fig. 1b-d), we established a robust transcriptional dataset comprising 16,410 unique genes. This yielded high-coverage data with an average of 8,545 ± 254 transcripts detected per sample (Extended Data Fig. 1e). Remarkably, we found that the commonly used laboratory strain *E. coli* OP50 used for most studies in classic genetics and aging related publications induces a transcriptional profile in *C. elegans* distinct to the transcriptional signatures of the majority of strains (Fig. 1e), including the K-12 MG1665 lab strain whose genome was one of the first *E. coli* reference sequences to be completed and extensively curated.

Next, we investigated whether grouping strains by phylogroup, as a proxy for bacterial function, would reveal a structure in the worm transcriptional response to the *E. coli* strain panel. The whole transcriptional profiles were correlated to the phylogroup partitioning of the *E. coli* pangenome and variance explained by this functional division was measured. Principal component analysis (PCA) revealed a modest separation between the phylogroups included in this screening (Fig. 1f), consistent with a weak clustering of pairwise Euclidean distances between strains (Extended Data Fig. 1d). Moreover, a permutational analysis of variance (PERMANOVA) for the full transcriptome dataset indicated a significant effect of phylogroup.

However, the model explained only approximately 1.5% variance. Nevertheless, the large Euclidean distance in transcriptional responses observed between strains known to elicit a distinct physiological response in the worm such as OP50 and MG1655 (Extended Data Fig. 1f), suggests that a robust biological signal exists within this transcriptional landscape. To facilitate mapping the worm response onto the *E. coli* pangenome, we reasoned that clustering the normalized read counts to discrete functional categories would improve our ability to map worm response. For this, the curated worm database from Holdorf *et al*.[13] was leveraged and normalized read counts were aggregated for all genes within each functional category at the three defined hierarchical levels defined in the database (Fig. 1a). This yielded three matrices of increased granularity, ranging from 33 broad categories (level 1) to 461 highly specific functional categories (level 3). This stratification generated a dense phenotypic landscape comprising 272,912 phenotypic worm data points at level 3 functional category resolution. PCA for each category (Fig. 1g) revealed a complex landscape where simple patterns remain elusive, underscoring the difficulty of mapping pangenome structure onto worm responses using traditional means. While PERMANOVA testing for level 3 functional categories identified a statistically significant effect (P= 0.041), the low variance explained (1.45%) highlights that evolutionary history alone is a minor driver of host phenotype. Together, these results show that *C. elegans* mounts highly-strain specific responses to *E. coli* strains, both at the whole transcriptional level and at the functional category levels. This confirms that the openness and complexity of the *E. coli* pangenome pose a high-dimensional biological challenge that cannot be fully captured by a standard phylogenetics approach.

The functional content of *E. coli* is encoded in the protein embedding space

The genetic repertoire of commonly used laboratory strains represents only a fraction of the total species diversity [5]. However, there is a consensus supporting that *E. coli* clades, whether classified from multi-locus sequence typing (MLST) or broader phylogroup classification, harbor characteristic functional enrichments associated with their primary ecological niche[6,10,14]. Given the shortcomings of current approaches, to systematically investigate the functional diversity of the *E. coli* pangenome, we established a comprehensive panel of genome assemblies to fully capture this functional richness. We accessed the NCBI Genomes database (Jan 2024) and selected 8,829 high-quality genome assemblies, to which we added assemblies for the strains available in our laboratory, yielding a total of 9,558 high-quality genomes (see Methods, Extended Data Fig.1g). Phylogroup proportions in this extended collection followed a similar pattern as in our laboratory panel (Fig. 2b, R = 0.82, p = 0.014, Extended Data Fig. 2a). This resulted in a total of 92,435 gene families split into a core of 3,005 genes, a shell of 2,917 genes, and a cloud of 86,513 genes (Fig. 2a). Consistent with previous work, the *E. coli* pangenome remained open as indicated by the Heap's law fit (γ =

158 0.28, Extended Data Fig. 2b). However, we observe that new functions start to saturate, with
159 the core genome rapidly stabilizing. Additionally, the pairwise genetic similarity for this larger
160 panel is similar to our laboratory panel (Jaccard similarity of 0.641 and 0.65 respectively,
161 Extended Data Fig. 2c). Phylogeny on the core genome shows that *E. coli* robustly follows the
162 phylogroup evolutionary lineages (Fig. 2c). When considering gene presence-absence
163 patterns within the cloud genome as a proxy for differential functional content, strain
164 relationships strongly follow the same phylogenetic structure (Fig. 2d). This congruence
165 between core genome phylogeny and total gene content demonstrates that a shared
166 evolutionary history shapes the full genomic repertoire of *E. coli*, linking deep ancestry to
167 functional gene repertoire at the strain level.
168
169 To characterize functional content, four gene ontology (GO) annotation strategies were
170 benchmarked by leveraging the linear reference of representative gene families. The most
171 widely used methods rely on sequence similarity which favors accuracy at the cost of limited
172 range. Recently, the development of methods relying on machine learning (ML) models have
173 been proven to achieve significant improvements in quality and coverage[15]. The 4 methods to
174 predict GO terms included Interproscan[16] and eggNOG-mapper[17], which are based on
175 sequence similarity, and ML methods such as Proteinfer[18], a Convolutional Neural Network
176 method, and GoPredSim[19], which leverages the protein language model (pLM) ProtT5-XL-
177 BFD[12] by creating an embedding representation for each protein. ML-based approaches
178 annotated substantially more genes per genome than sequence-similarity based methods
179 (41.7-52.7% increase, Extended Data Fig. 2d, $p<0.001$, one-way ANOVA). This difference
180 was primarily driven by annotations in the cloud and shell genomes where unique sequences
181 are mostly found (Extended Data Fig. 2e,f). Despite this significant improvement, 2,873 gene
182 families remained unannotated. Exploiting the hierarchical information contained in GO
183 annotations, an analysis of the maximum information content per gene and method revealed
184 that the GoPredSim, which leverages the pLM ProtT5 embeddings, produced the most
185 informative annotations (Extended Data Fig. 2g). These results underscore that much of the
186 accessory functional landscape from *E. coli* remains largely uncharacterized but can be more
187 effectively represented using pLM embedding-based models, given the prediction abilities and
188 breadth of information.
189
190 Protein embeddings from the ProtT5-XL-BFD pLM were then used to create a comprehensive
191 functional map of the *E. coli* pangenome (Fig. 1a, Extended Data Fig.1g). PCA of the resulting
192 high-dimensional embeddings revealed a spatial organization aligned with pangenome
193 structure (Fig. 2e, Extended Data Fig. 2h). The core genome occupies a compact, dense
194 region of the space, which expands through the shell genome and into the vast, sparsely

populated region occupied by the cloud genome with significant differences between these compartments (Fig. 2f, Extended Data Fig. 2h, $p < 0.001$, one-way ANOVA). This geometric arrangement demonstrates that the embedding space quantitatively captures functional diversity, transitioning from the conserved core functions to the diverse and mobile cloud genome. Analysis of the COG categories further show that core-associated functions such as energy production and conversion (C) or cell cycle control(D) are enriched in the dense core region. However, functions known to be environment-dependent such as carbohydrate transport (G), defense mechanisms (V) or poorly characterized (S) are shifted towards the periphery (Figure 2g,h,i, Extended Data Fig. 2i,j,k). Collectively, these results establish that the geometry of the protein embedding space created through the pLM ProtT5 acts as a quantitative proxy for the pangenome's functional architecture.

Given that the functional content of a pangenome is encoded in the geometry of the protein embedding space, we extended this framework to establish a microbial strain identity, defined by the combination of its conserved core functions and unique set of accessory functions. Each strain was embedded as a single vector by combining binary gene content (presence/absence) with the corresponding protein embeddings, calculated as an average of all protein vectors considering the whole gene content (see Methods section), yielding a unique representation per strain that determines their functional potential. PCA of these strain embeddings produced a pangenome-level functional landscape in which strains are positioned according to functional capacity (Fig. 2j, Extended Data Fig. 2l). The principal component analysis of the strain embeddings revealed a structured functional landscape that validates our model while highlighting the limitations of pure phylogeny approaches ($p < 0.001$, one-way ANOVA, Extended Data Fig. 3b). PC2 (14.68%) cleanly separated phylogroups, confirming that the embeddings correctly encoded the strains' evolutionary history, whereas PC1 (35.45%) was driven by within-phylogroup variability. This analysis proves that the vast majority of *E. coli* functional diversity is driven by strain-specific adaptations. This pattern is captured by alternative embedding methods, supporting the robustness of these findings (Extended Data Fig. 3a,c). We observe the strain embeddings recaptures the clusters defined by ecology and evolution where phylogroups A, B1 and C formed a clearly defined cluster, followed by a distinct cluster composed of phylogroups D, E, F and G bridging towards the phylogroup B2, which occupied a distinct, distant region. This separation of phylogroups along PC2 supports the evolutionary history between the phylogroups (Fig.2i). Together, these results demonstrate that protein language model embeddings provide a powerful, compact representation of both genes and whole genomes, enabling the encoding of *E. coli* pangenome structure, the delineation of conserved versus accessory functions, and the

representation of entire strains as single, functionally meaningful vectors that can be directly linked to host phenotypes.

Host physiology maps onto the strain embeddings functional landscape

We next hypothesized whether using the *E. coli* strain embeddings would improve our ability to establish causal links between microbial functions and host physiology at the pathway level. To this end, we leveraged the geometry of the PCA projection coordinates derived from the strain embeddings as a functional map to position *C. elegans* transcriptional and functional categories (Fig. 1a). The strain embedding map for the subset of *E. coli* strains used in monoassociation experiments recapitulated the distinctive structure observed for the full pangenome map, revealing a clear separation within and between phylogroups along PC1 and PC2 coordinates, respectively (Fig. 3a). To connect *E. coli* pangenome functional landscape and the worm physiology, we calculated the Spearman correlations between the level 3 pathway categories in the worm transcriptome to the PC1 and PC2 coordinates of the strain embeddings. This analysis revealed a distinct set of host functional categories regulated at the pangenome level (Fig. 3b). Notably, all significant correlations between host pathways and strain embedding axes were linked to PC2, which separates the distinct phylogroups in the pangenome (Fig. 3c, Extended Data Fig. 3d). The significant pathways involved in the host physiology regulation at the pangenome level, clustered into biologically broad processes including cell cycle, central and one-carbon metabolism, proteostasis and response to stress (Fig. 3d, $p < 0.05$, BH correction). These processes have been formerly linked to the regulation of several host phenotypes, including development and aging[20-22]. Together, these findings suggest that functional variation across the *E. coli* pangenome encodes regulatory signals that can modulate core host physiological programs.

Given the *E. coli* pangenome functional landscape - host physiology connections, we sought to validate these associations experimentally. We performed a reporter-based quantitative screen of 13 fluorescent transcriptional or translational reporters representing key genes within the main pathways identified in the functional mapping (Extended Data Fig. 4a). We confirmed the robustness of the experimental pipeline by testing two independent biological replicates over 589 *E. coli* pangenome strains for each reporter (Pearson correlation > 0.7, Extended Data Fig. 4b). We then mapped the reporter gene expression levels to the *E. coli* strain embeddings and found that all 13 gene reporter expression profiles significantly correlated with the strain embedding geometrical projection, mirroring the transcriptional landscape pattern observed (Fig. 3e, $p < 0.05$, BH correction). Next, we leveraged these data to obtain further insights between microbe-host functional relationships. Plotting reporter expression by strain revealed distinct transcriptional programs which were largely independent

of their phylogenetic relatedness (Extended Data Fig. 4c). Similarly, pairwise strain correlations of reporter activity indicated that bacterial functional guilds transcended phylogenetic relatedness on regulating specific host transcriptional responses (Fig. 3f,g, Extended Data Fig. 4d). Interestingly, we observed a global bias towards positive associations between strain pairs (~65%, Extended Data Fig. 4e, f), implying a shared host transcriptional response across the *E. coli* pangenome. Consistent with this, we observed significant pairwise correlations between the 13 reporters (Fig. 3h).

Together, these results show that the *E. coli* functional landscape can be quantitatively mapped onto the host transcriptional and physiological programs by leveraging strain embeddings, uncovering bacterial-driven regulation of fundamental host processes such as central and 1CC metabolism, stress response, or proteostasis. Based on these findings, we next examined how *E. coli* influences host proteostasis, which remain insufficiently characterized in this context.

Propionate and vitamin B12 at the interface of bacterial-host "co-proteostasis"

In eukaryotic cells, the ubiquitin-proteasome system (UPS) plays a central role in maintaining proteostasis by controlling the degradation of damaged proteins. Yet, how the UPS integrates environmental signals to support organismal physiology remains poorly understood. First, we grew the *C. elegans* ubiquitin-proteasome (UPS) reporter strain[23] on individual bacterial isolates from distinct phyla of the *C. elegans* microbiome[24]. Host proteostasis displayed strong bacterial strain-dependent variation (Extended Data Fig. 5a, as observed for the *E. coli pangenome* (Extended Data Fig. 4c) and *E. coli* laboratory strains[25], suggesting fine levels of mechanistic regulation. To identify the underlying mechanism(s), we performed a qualitative screen using the UPS worm reporter strain, together with the single deletion *E. coli* KEIO library and found that deletion of the protease Lon and the homolog of the heat shock protein HSP70/DnaK/DnaJ significantly decreased or increased UPS fluorescence, respectively. This led us to hypothesize that bacterial proteostasis could regulate host proteostasis. Next, we quantitatively tested all known *E. coli* chaperones as well as proteases (Fig. 4a). We confirmed that deletion of the functional DnaK and DnaJ heat shock pair increased UPS fluorescence and respective protein levels, while deletion of Lon, ClpX, BepA proteases and HtpG, HscA, CbpM chaperones (Fig. 4b,c; Extended Data Fig. 5b-d) significantly reduced UPS fluorescence and protein levels.

To better understand the regulatory networks between these chaperones and proteases and the potential mechanisms involved in the regulation of host proteostasis, we performed proteomics of each individual mutant strain. Deletion of *dnaK* led to pronounced changes in

the proteome landscape followed by the deletion of *lon* and *clpX* (Extended Data Fig. 6a). We observed an intricate compensatory mechanism whereby the single deletion of any of these proteases or chaperones leads to significant changes in a network of other chaperones and proteases (Extended Data Fig. 6b), with greater effects observed for *dnaK* and *clpX* mutants. Thus, we tested whether co-regulatory effects of proteases and chaperones could regulate host UPS response through the creation of double-mutant *E. coli* strains of *dnaK*. Notably, only the combined loss of *dnaK* and *clpX* abolished the UPS activation induced by the *dnaK* mutant alone (Fig. 4e, Extended Data Fig. 6c,d) without compromising bacterial fitness (Extended Data Fig. 7a). U.V-irradiation experiments to abolish metabolic activity of *E. coli* and alter their metabolome[26], further show that UPS regulation was dependent on the active metabolism of *E. coli* (Fig. 4f, Extended Data Fig. 6e). To identify the potential mechanism(s) responsible, we performed proteomics in both single and double *E. coli* mutants (Fig. 4g, Extended Data Fig. 7) as well as in worms grown on these bacteria (Fig. 4h, Extended Data Fig. 8). Despite functional rescue at the host level, the *dnaKclpX* double mutant exhibited a unique proteomic signature distinct from both single mutants and wild-type *E. coli* (Extended Data Fig. 7b), as well as in worms (Extended Data Fig. 8a). KEGG analysis in both *E. coli* (Extended Data Fig. 7c-g) and *C. elegans* (Extended Data Fig. 8b-f) with a focus on the comparison between up and downregulated proteins of *dnaKclpX* versus *dnaK* (Fig. 4g-h), which links to the loss of *dnaK* effects on host UPS, showed enrichment of central carbon and amino acid metabolism pathways—particularly branched-chain amino acid (BCAA previously identified as an important UPS regulator[27]) and propionate metabolism, two directly connected metabolic pathways through the sharing of propionyl-CoA—suggesting a role for metabolic cross-talk in modulating host proteostasis. To determine how *E. coli* regulated BCAA and propionate metabolism, we compared all proteins that were significantly down-regulated in *dnaK* mutants while up-regulated in both *clpX* and *dnaKclpX* mutants (Fig. 4i). From the 33 proteins shown to be significant in these comparisons, the tree proteins of the TonB-ExbB-ExbD energy transduction complex were strongly enriched for the Gene Ontology Molecular function of energy transducer activity (FDR=0.0111, Strength=2.1) which are involved in the transport of iron and vitamin B12 (VB12)[28]. Given the well described role of bacterial VB12 in the regulation of BCAA and propionate homeostasis in the host[22], we investigated whether VB12 and/or propionate metabolites regulated UPS proteostasis. First, we created triple mutants of all known iron transporters in *E. coli* and observed that only the deletion of TonB significantly increased the levels of UPS fluorescence when compared to the effect observed when fed *dnaKclpX* mutants (Fig. 4j, p=0.0008, Extended Data Fig. 9a, b) suggesting a regulation of TonB levels by DnaK that are controlled by ClpX. Consistent with this observation, the overexpression of TonB in a *dnaK* or *dnaKclpX* mutant significantly reduced UPS levels (Extended Data Fig. 9c). Overexpression of BtuB protein, a specific transporter of

VB12, also reduced *dnaKclpX* levels to baseline levels (Extended Data Fig. 9d) and this required TonB, confirming that canonical tonB-BtuB-dependent VB12 transport is central to this regulatory axis. Supplementation of VB12 uniformly decreased UPS levels in all bacterial mutant backgrounds (Fig. 4k, Extended Data Fig. 9e) but significant differences in UPS levels between them suggested the role of additional metabolites shaping host proteostasis. Given the role of VB12 in regulating propionate metabolism, we supplemented propionate and found that propionate increased UPS levels in worms fed control bacteria and *dnaKclpX* mutants but not *dnaK* (Fig. 4l, Extended Data Fig. 9f). Together with our proteomic data (Fig. 4g), it suggested a role for propionate as a potential metabolite regulating host proteostasis. Deletion of the *sbm* operon, which encodes enzymes for the "sleeping beauty" mutase pathway that converts BCAAs—particularly isoleucine—into propionate[29], abolished the *dnaK*-induced UPS increase without affecting bacterial fitness (Fig. 4m, Extended Data Fig 9g,h), directly linking bacterial propionate production to host UPS activation.

In *C. elegans*, propionate catabolism proceeds through a VB12-dependent canonical pathway and a VB12-independent "propionate shunt," whose activation can be monitored by the *acdh-1p::GFP* reporter[22]. Worms fed *dnaK* mutants showed increased *acdh-1* levels, indicating elevated flux through the VB12-independent shunt, whereas worms fed *dnaKclpX* bacteria did not (Fig. 4n), consistent with proteomic evidence that *dnaKclpX* suppresses *dnaK*-driven metabolic rewiring. Genetic inhibition of the shunt downstream of *acdh-1* (e.g., *hphd-1, alh-8*) or the canonical pathway (e.g., *mce-1*) elevated UPS activity (Fig. 4o, Extended Data Fig. 10a-c), while VB12 supplementation reduced UPS levels in wild-type and *acdh-1* mutants but not in *mce-1 or hphd-1* mutants, demonstrating that the balance between VB12-dependent and -independent propionate catabolism determines the impact of propionate on proteostasis. Proteomic analyses of worms fed *dnaK* versus *dnaKclpX* bacteria revealed enrichment of pathways linked to ketone metabolism (Fig. 4h), aligning with previous reports that perturbations in propionate catabolism can alter ketone body pathways[30]. RNAi of *suca-1*, which contributes to conversion of acetoacetate to acetoacetyl-CoA (Extended Data Fig 10b), increased UPS activity, whereas exogenous acetoacetate supplementation (independent of its degradation by *atoA* or *atoB*) decreased UPS levels (Fig. 4p, Extended Data Fig 10d) , suggesting that ketone intermediates can directly modulate host proteostasis.

Collectively, these findings define a mechanistic chain in which bacterial DnaK–ClpX–TonB–BtuB control bacterial proteostasis and VB12 transport, thereby shaping propionate production and routing in the host, which in turn determines the balance between toxic VB12-independent catabolites and protective VB12-dependent flux, ultimately tuning UPS activity and proteostasis in *C. elegans*.

DISCUSION

Metagenomic sequencing and other state-of-the art technical advances now enable high throughput, high-resolution scale analyses of microbial strains across diverse and complex ecosystems ranging from the human gut to marine and soil environments. Strain-level resolution has recently been shown to be crucial in microbiome research and in dictating microbe-host interactions. For example, strains of the same species can have diametrically opposed functional, ecological, and clinical manifestations, with species-level identification often leading to erroneous interpretations[31]. Strain-level characterization has also been emphasized in how bacterial strains are transmitted in human populations, highlighting the importance of the need to consider their biological effects [32,33]. *Escherichia coli* has become a canonical example of the diversity displayed by a bacterial species, showing that its vast accessory genome harbored in its open pangenome contains an extensive array of bacterial functions that can potentially alter host physiology[6,10]. Our panel shows that any two given strains can differ by more than 50% of their genetic content. A central challenge in understanding the microbiome is reconciling the inherent genetic diversity contained within bacterial species and how this affects host physiology[34,35].

Here, we present the most comprehensive analysis to date of how strain level variation within a single bacterial species shapes host responses. Given the challenge in defining bacterial functions for the poorly described microbial accessory genes, methods based in the transformers architecture have been trained over large protein databases creating physical representations of the protein space encoded in the microbiome[36,37]. In this work, we leveraged the protein embeddings predicted by the pLM[12] to study the latent functional landscape encoded in the *E. coli* pangenome. By compressing the genomic and functional information encoded in thousands of *E. coli* strains into unified “strain embeddings”, we created a geometric map that captures the potential function per strain, which can be mapped onto the host phenotype it elicits. By building a high-throughput panel of strain-host mono-association spanning hundreds of *E. coli* strains, we have established an experimental platform validating our computational approach and providing additional mechanistic insights. Together, our data uncovers several fundamental principles in microbe-host biology at the strain level. 1) Each strain elicits a unique molecular signature in the host; 2) our data shows that the protein functions shared by every strain (the core) are located in a narrow geometrical space compared to the vast strain-specific functions, 3) strains within the same phylogroup display an ~60:40 of positively to negatively correlated effects on host responses; and 4) phylogenetic relatedness between strains does not predict the host molecular programs they induce. Our data supports the hypothesis that phylogeny, while important, is inadequate as a

single factor to link bacterial functions to host physiology. Future work will be required to determine how strain-level effects manifest in the context of complex microbiota and to test whether these basic principles extend to other bacterial species with open or closed pangenomes.

In line with this, the canonical laboratory microbial source for *C. elegans*, the *Escherichia coli* OP50 strain, elicits a distinct and divergent molecular response in the host compared with other *E. coli* strains (Fig. 1e). This observation is consistent with a growing number of studies incorporating multiple bacterial strains reporting strain-dependent mechanisms underlying diverse host phenotypes, including drug responses [38,39], behavior, reproduction, and lifespan [21,40,41]. Collectively, these findings suggest that the experimental convenience afforded by *E. coli* OP50 may be offset by the specific molecular and physiological signature it imposes on *C. elegans*, potentially failing to reflect the "true" wild-type biology of the host and motivating a critical re-evaluation of the foundational literature of an entire field. This may possibly be better captured using native microbiome *C. elegans* strains or alternatively, commensal *E. coli* strains employed in this study. Using a protein-embedding framework, this work supports these claims as it identifies a broad repertoire of bacterial functions, spanning many COG functional categories with known effects on host physiology, that are regulated at the level of the bacterial pangenome. Among the most significantly enriched categories is proteolysis. Notably, recent work has demonstrated that differences in bacterial-derived RNAs between *E. coli* OP50 and HT115 can trigger a systemic response in *C. elegans* that protect against protein aggregation during aging[42]. Likewise, the present study reveals pronounced differences among *E. coli* strains (Extended Figure S4a) and strains belonging to bacterial species from other phyla (Extended Figure S5a) in their ability to modulate the host ubiquitin–proteasome system. Here, we demonstrate that key bacterial proteostasis regulators control host UPS activity by modulating the availability of vitamin B12 and propionate, which in turn dictates the flux through host propionate degradation pathways. Bacteria that produce or efficiently scavenge B12 can control community composition and metabolic activity by outcompeting B12-dependent neighbors [43,44]. For example, B12 production by *Eubacterium hallii* enables *Akkermansia muciniphila* to convert succinate to propionate, shifting succinate levels, and thereby reshaping the surrounding metabolic network [45]. Here, an unanticipated mechanism is described in which the DnaK/J chaperone system and the ClpX protease act in concert to fine tune B12 and propionate levels. While this regulatory axis may have evolved primarily to modulate microbial community interactions, it also alters host proteostasis, giving rise to what can be conceptualized as microbe–host "co-proteostasis" derived from microbe-host co-metabolism cues.

MAIN REFERENCES

1. Samuel, B. S., Rowedder, H., Braendle, C., Felix, M. A. & Ruvkun, G. Caenorhabditis elegans responses to bacteria from its natural habitats. *Proc Natl Acad Sci U S A* **113**, E3941-3949 (2016). https://doi.org/10.1073/pnas.1607183113
2. Backes, C., Martinez-Martinez, D. & Cabreiro, F. C. elegans: A biosensor for host-microbe interactions. *Lab Anim (NY)* **50**, 127-135 (2021). https://doi.org/10.1038/s41684-021-00724-z
3. Tremaroli, V. & Backhed, F. Functional interactions between the gut microbiota and host metabolism. *Nature* **489**, 242-249 (2012). https://doi.org/10.1038/nature11552
4. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504 (2019). https://doi.org/10.1038/s41586-019-0965-1
5. Cummins, E. A., Hall, R. J., Connor, C., McInerney, J. O. & McNally, A. Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types. *Microb Genom* **8** (2022). https://doi.org/10.1099/mgen.0.000903
6. Touchon, M. *et al.* Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLoS Genet* **16**, e1008866 (2020). https://doi.org/10.1371/journal.pgen.1008866
7. Neve, I. A. A. *et al.* Escherichia coli Metabolite Profiling Leads to the Development of an RNA Interference Strain for Caenorhabditis elegans. *G3 (Bethesda)* **10**, 189-198 (2020). https://doi.org/10.1534/g3.119.400741
8. Watson, E. *et al.* Metabolic network rewiring of propionate flux compensates vitamin B12 deficiency in C. elegans. *Elife* **5** (2016). https://doi.org/10.7554/eLife.17670
9. Revtovich, A. V., Lee, R. & Kirienko, N. V. Interplay between mitochondria and diet mediates pathogen and stress resistance in Caenorhabditis elegans. *PLoS Genet* **15**, e1008011 (2019). https://doi.org/10.1371/journal.pgen.1008011
10. Galardini, M. *et al.* Phenotype inference in an Escherichia coli strain panel. *Elife* **6** (2017). https://doi.org/10.7554/eLife.31035
11. Bergelson, J., Kreitman, M., Petrov, D. A., Sanchez, A. & Tikhonov, M. Functional biology in its natural context: A search for emergent simplicity. *Elife* **10** (2021). https://doi.org/10.7554/eLife.67646
12. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112-7127 (2022). https://doi.org/10.1109/TPAMI.2021.3095381
13. Holdorf, A. D. *et al.* WormCat: An Online Tool for Annotation and Visualization of Caenorhabditis elegans Genome-Scale Data. *Genetics* **214**, 279-294 (2020). https://doi.org/10.1534/genetics.119.302919
14. Tantoso, E., Eisenhaber, B., Sinha, S., Jensen, L. J. & Eisenhaber, F. About the dark corners in the gene function space of Escherichia coli remaining without illumination by scientific literature. *Biol Direct* **18**, 7 (2023). https://doi.org/10.1186/s13062-023-00362-0
15. Barrios-Nunez, I. *et al.* Decoding functional proteome information in model organisms using protein language models. *NAR Genom Bioinform* **6**, lqae078 (2024). https://doi.org/10.1093/nargab/lqae078
16. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014). https://doi.org/10.1093/bioinformatics/btu031
17. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829 (2021). https://doi.org/10.1093/molbev/msab293

18 Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer, deep neural networks for protein functional inference. *Elife* **12** (2023). https://doi.org/10.7554/eLife.80942

19 Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep* **11**, 1160 (2021). https://doi.org/10.1038/s41598-020-80786-0

20 Finger, F. *et al.* Olfaction regulates organismal proteostasis and longevity via microRNA-dependent signaling. *Nat Metab* **1**, 350-359 (2019). https://doi.org/10.1038/s42255-019-0033-z

21 Cabreiro, F. *et al.* Metformin retards aging in C. elegans by altering microbial folate and methionine metabolism. *Cell* **153**, 228-239 (2013). https://doi.org/10.1016/j.cell.2013.02.035

22 Watson, E. *et al.* Interspecies systems biology uncovers metabolites affecting C. elegans gene expression and life history traits. *Cell* **156**, 759-770 (2014). https://doi.org/10.1016/j.cell.2014.01.047

23 Segref, A. *et al.* Pathogenesis of human mitochondrial diseases is modulated by reduced activity of the ubiquitin/proteasome system. *Cell Metab* **19**, 642-652 (2014). https://doi.org/10.1016/j.cmet.2014.01.016

24 Zhang, F. *et al.* Caenorhabditis elegans as a Model for Microbiome Research. *Front Microbiol* **8**, 485 (2017). https://doi.org/10.3389/fmicb.2017.00485

25 Segref, A., Torres, S. & Hoppe, T. A screenable in vivo assay to study proteostasis networks in Caenorhabditis elegans. *Genetics* **187**, 1235-1240 (2011). https://doi.org/10.1534/genetics.111.126797

26 Essmann, C. L. *et al.* Mechanical properties measured by atomic force microscopy define health biomarkers in ageing C. elegans. *Nat Commun* **11**, 1043 (2020). https://doi.org/10.1038/s41467-020-14785-0

27 Ravanelli, S. *et al.* Reprograming of proteasomal degradation by branched chain amino acid metabolism. *Aging Cell* **21**, e13725 (2022). https://doi.org/10.1111/acel.13725

28 Celia, H. *et al.* Cryo-EM structure of the bacterial Ton motor subcomplex ExbB-ExbD provides information on structure and stoichiometry. *Commun Biol* **2**, 358 (2019). https://doi.org/10.1038/s42003-019-0604-2

29 Gonzalez-Garcia, R. A. *et al.* Awakening sleeping beauty: production of propionic acid in Escherichia coli through the sbm operon requires the activity of a methylmalonyl-CoA epimerase. *Microb Cell Fact* **16**, 121 (2017). https://doi.org/10.1186/s12934-017-0735-4

30 Ponomarova, O. *et al.* A D-2-hydroxyglutarate dehydrogenase mutant reveals a critical role for ketone body metabolism in Caenorhabditis elegans development. *PLoS Biol* **21**, e3002057 (2023). https://doi.org/10.1371/journal.pbio.3002057

31 Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat Rev Microbiol* **18**, 491-506 (2020). https://doi.org/10.1038/s41579-020-0368-1

32 Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125-135 (2023). https://doi.org/10.1038/s41586-022-05620-1

33 Martinez-Martinez, D. *et al.* Chemotherapy modulation by a cancer-associated microbiota metabolite. *Cell Syst* **16**, 101397 (2025). https://doi.org/10.1016/j.cels.2025.101397

34 Andreu-Sanchez, S. *et al.* Global genetic diversity of human gut microbiome species is related to geographic location and host health. *Cell* **188**, 3942-3959 e3949 (2025). https://doi.org/10.1016/j.cell.2025.04.014

35 Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol Syst Biol* **13**, 960 (2017). https://doi.org/10.15252/msb.20177589

36 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023). https://doi.org/10.1126/science.ade2574

37 Odrzywolek, K. *et al.* Deep embeddings to comprehend and visualize microbiome protein space. *Sci Rep* **12**, 10332 (2022). https://doi.org/10.1038/s41598-022-14055-7

38 Martinez-Miguel, V. E. *et al.* Increased fidelity of protein synthesis extends lifespan. *Cell Metab* **33**, 2288-2300 e2212 (2021). https://doi.org/10.1016/j.cmet.2021.08.017

39 Scott, T. A. *et al.* Host-Microbe Co-metabolism Dictates Cancer Drug Efficacy in C. elegans. *Cell* **169**, 442-456 e418 (2017). https://doi.org/10.1016/j.cell.2017.03.040

40 Schifano, E. *et al.* Virulence behavior of uropathogenic Escherichia coli strains in the host model Caenorhabditis elegans. *Microbiologyopen* **8**, e00756 (2019). https://doi.org/10.1002/mbo3.756

41 Xiao, R. *et al.* RNAi Interrogation of Dietary Modulation of Development, Metabolism, Behavior, and Aging in C. elegans. *Cell Rep* **11**, 1123-1133 (2015). https://doi.org/10.1016/j.celrep.2015.04.024

42 Kyriakakis, E. *et al.* Bacterial RNA promotes proteostasis through inter-tissue communication in C. elegans. *Nat Commun* **16**, 8650 (2025). https://doi.org/10.1038/s41467-025-63987-x

43 Degnan, P. H., Barry, N. A., Mok, K. C., Taga, M. E. & Goodman, A. L. Human gut microbes use multiple transporters to distinguish vitamin B(1)(2) analogs and compete in the gut. *Cell Host Microbe* **15**, 47-57 (2014). https://doi.org/10.1016/j.chom.2013.12.007

44 Degnan, P. H., Taga, M. E. & Goodman, A. L. Vitamin B12 as a modulator of gut microbial ecology. *Cell Metab* **20**, 769-778 (2014). https://doi.org/10.1016/j.cmet.2014.10.002

45 Belzer, C. *et al.* Microbial Metabolic Networks at the Mucus Layer Lead to Diet-Independent Butyrate and Vitamin B(12) Production by Intestinal Symbionts. *mBio* **8** (2017). https://doi.org/10.1128/mBio.00770-17

46 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

47 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015). https://doi.org/10.1093/nar/gkv007

48 Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* **17**, 137-145 (2020). https://doi.org/10.1038/s41592-019-0654-x

49 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x

50 Waters, N. R., Abram, F., Brennan, F., Holmes, A. & Pritchard, L. Easy phylotyping of Escherichia coli via the EzClermont web app and command-line tool. *Access Microbiol* **2**, acmi000143 (2020). https://doi.org/10.1099/acmi.0.000143

51 Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb Genom* **4** (2018). https://doi.org/10.1099/mgen.0.000192

52 Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* **7** (2021). https://doi.org/10.1099/mgen.0.000685

53 Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020). https://doi.org/10.1186/s13059-020-02090-4

54 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013). https://doi.org/10.1093/molbev/mst010

55 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274 (2015). https://doi.org/10.1093/molbev/msu300

56 Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006). https://doi.org/10.1038/msb4100050

57 Doublet, B. *et al.* Antibiotic marker modifications of lambda Red and FLP helper plasmids, pKD46 and pCP20, for inactivation of chromosomal genes using PCR products in multidrug-resistant strains. *J Microbiol Methods* **75**, 359-361 (2008). https://doi.org/10.1016/j.mimet.2008.06.010

58 Chung, C. T., Niemela, S. L. & Miller, R. H. One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. *Proc Natl Acad Sci U S A* **86**, 2172-2175 (1989). https://doi.org/10.1073/pnas.86.7.2172

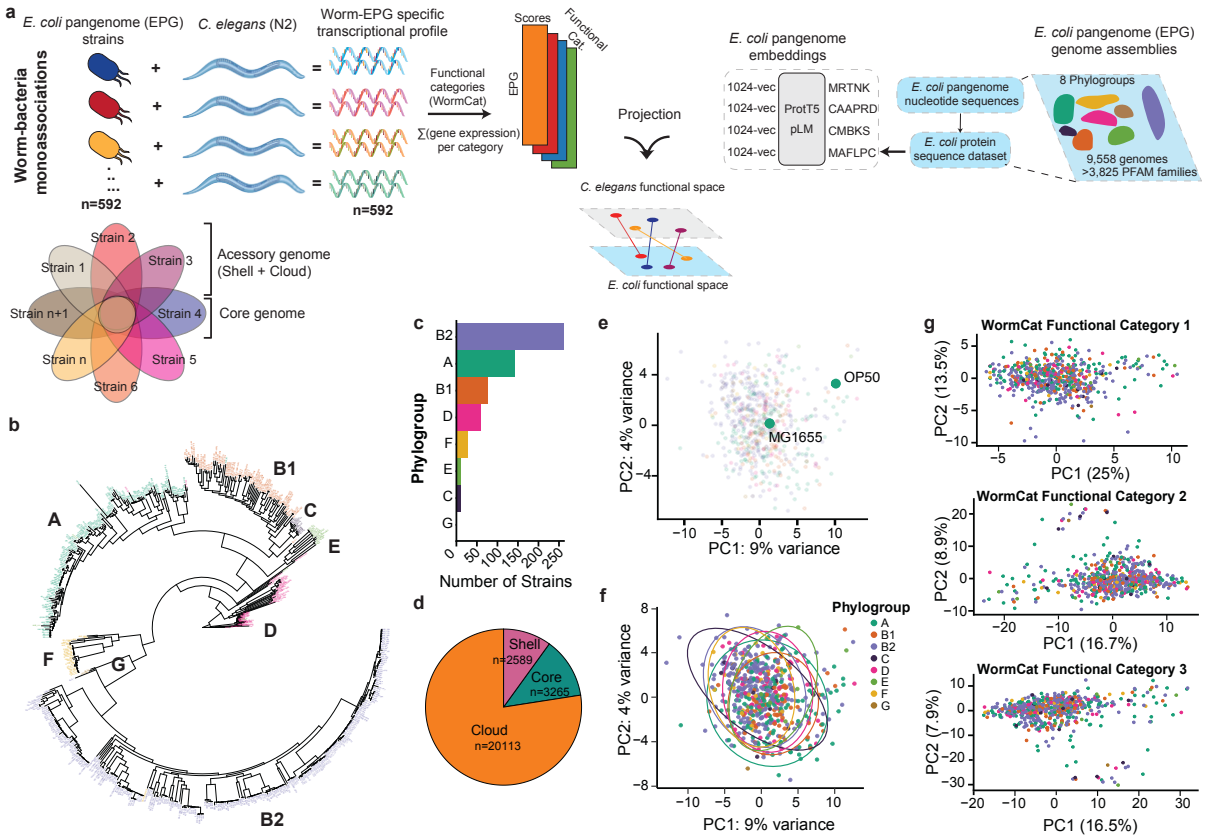
59 Ikeda, H. & Tomizawa, J. I. Transducing fragments in generalized transduction by phage P1. 3. Studies with small phage particles. *J Mol Biol* **14**, 120-129 (1965). https://doi.org/10.1016/s0022-2836(65)80234-0

60 Kitagawa, M. *et al.* Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res* **12**, 291-299 (2005). https://doi.org/10.1093/dnares/dsi012

61 Yang, C., Sun, W. & Ao, X. Bacterial inactivation, DNA damage, and faster ATP degradation induced by ultraviolet disinfection. *Frontiers of Environmental Science & Engineering* **14** (2019). https://doi.org/10.1007/s11783-019-1192-6

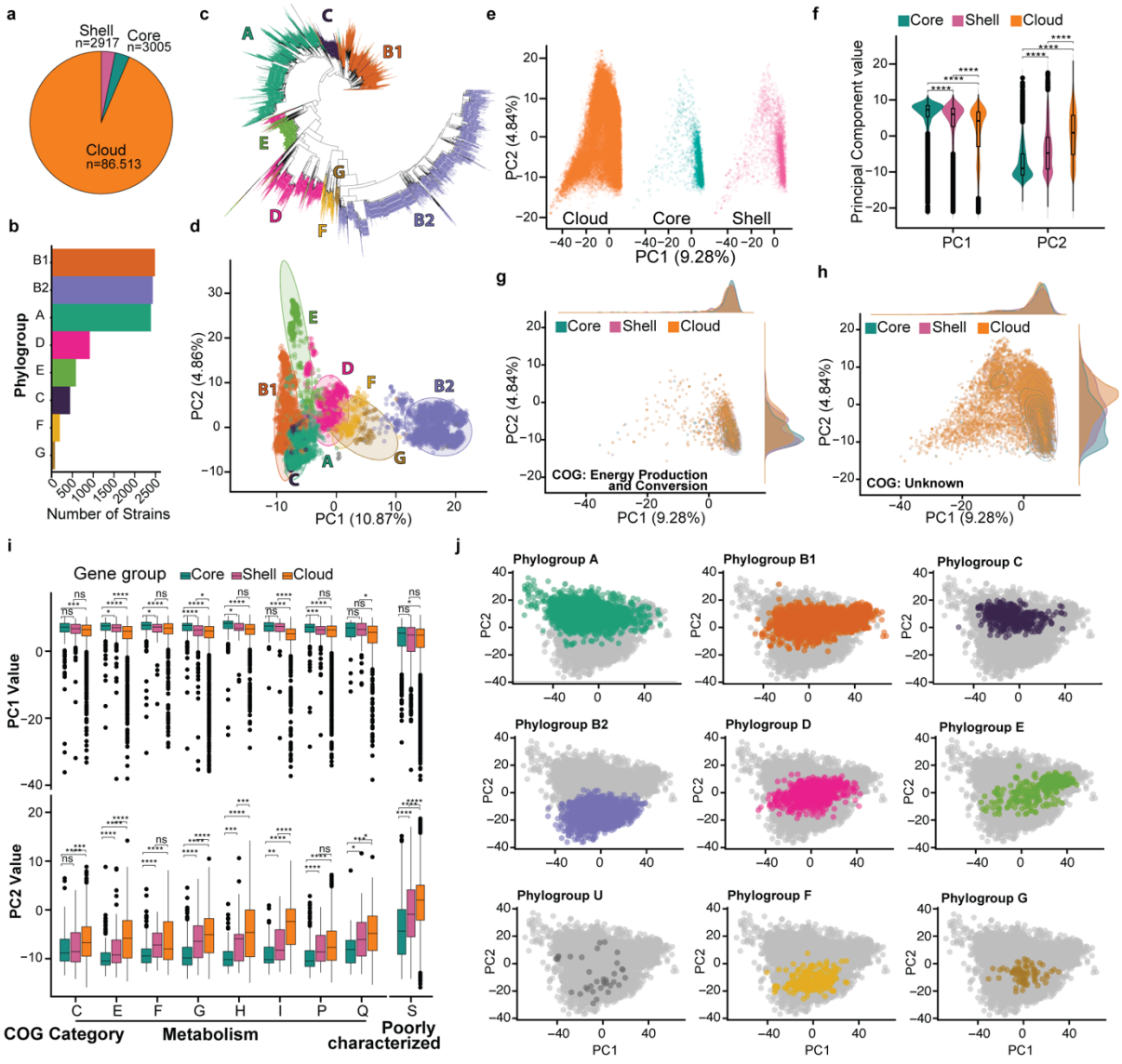
MAIN FIGURES

Figure 1



636 **Figure 2**

637



638

Figure 3

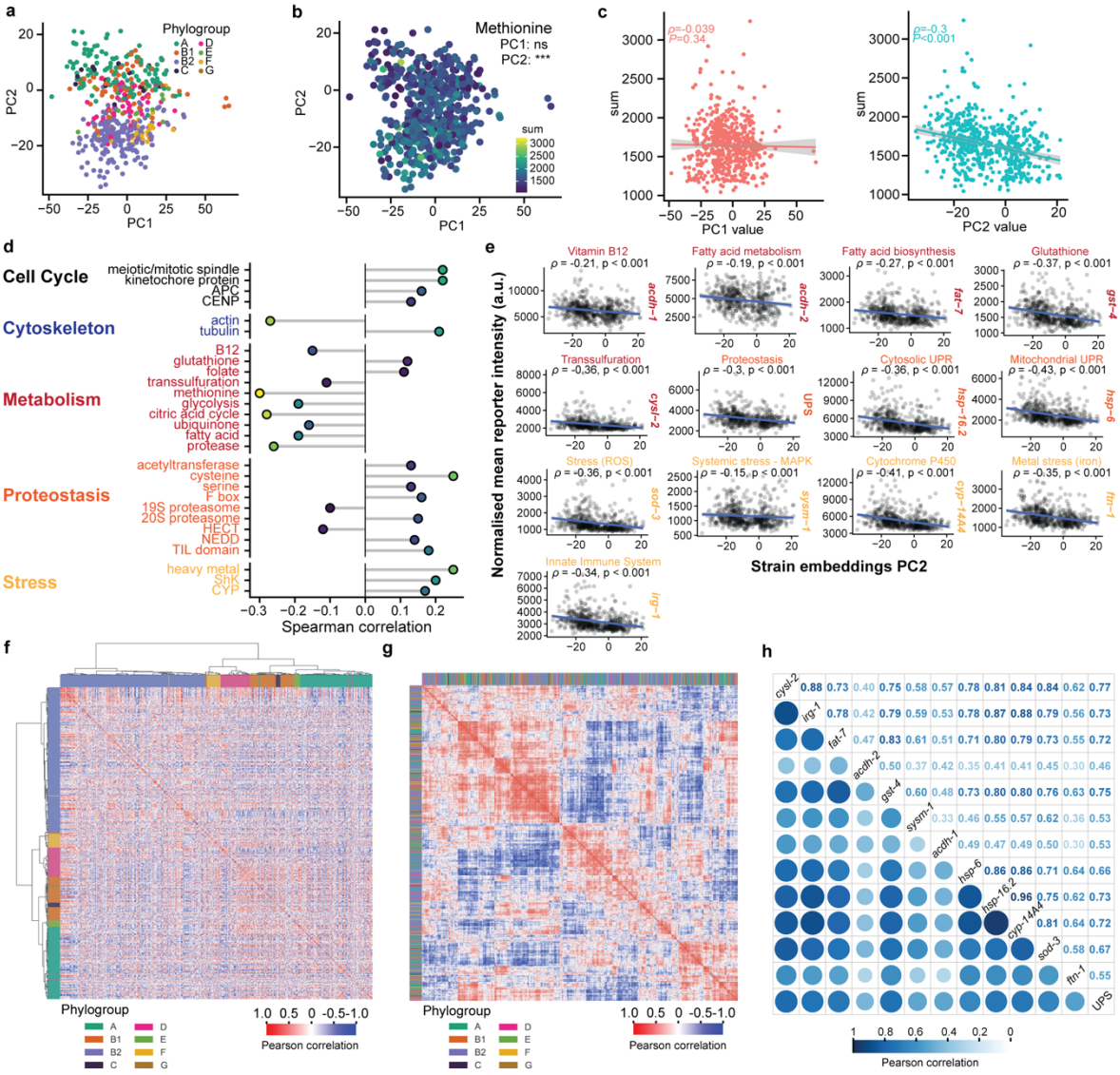


Figure 4

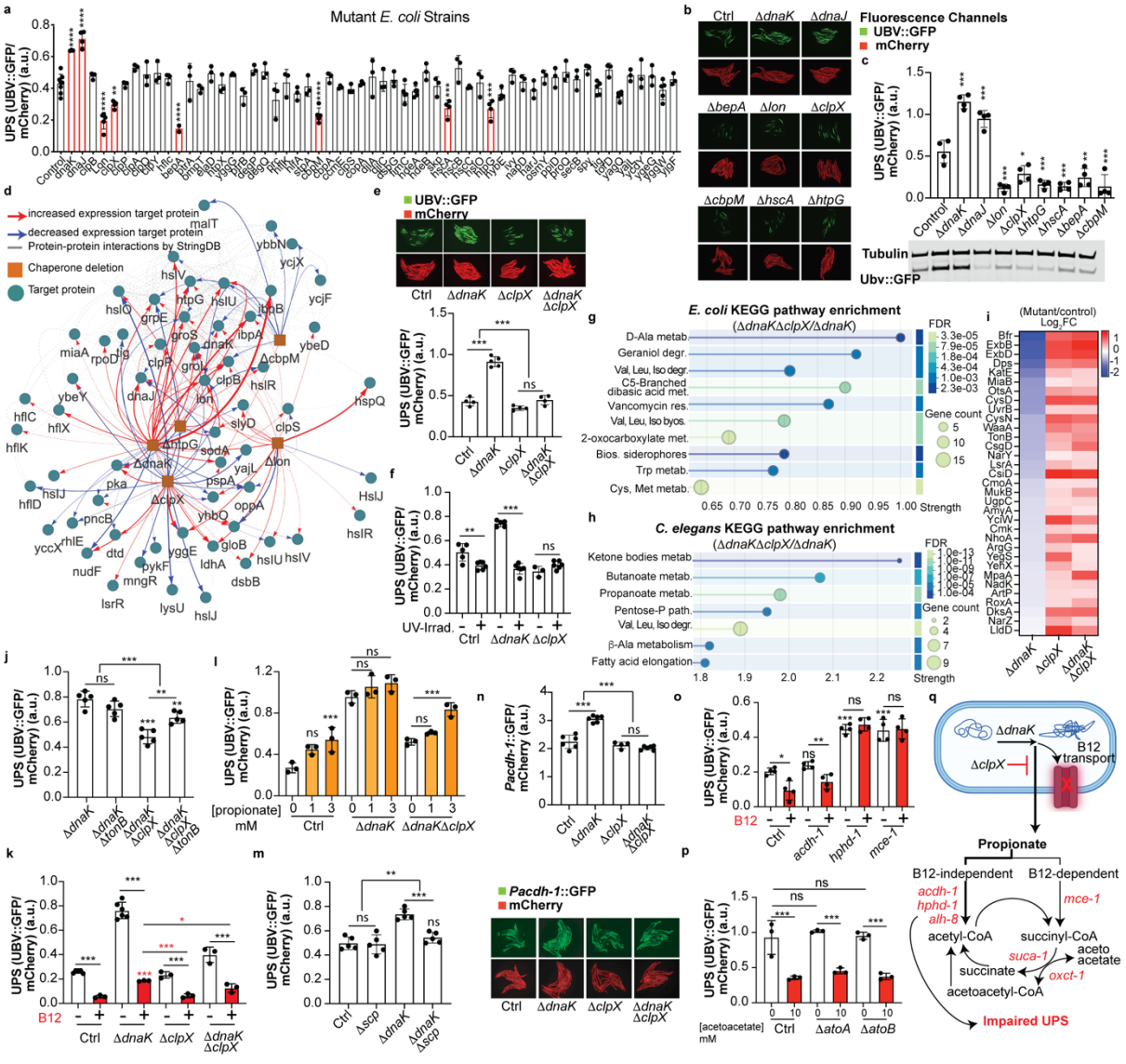


FIGURE LEGENDS

Figure 1. The physiology of *C. elegans* is regulated at the *E. coli* pangenome scale. a, Schematic representation of experimental and analytical workflow: *C. elegans* transcriptomes were profiled for each of the 592 *E. coli* strains and then summarized in WormCat functional categories. The *E. coli* pangenome was calculated for 9,558 strains and its linear reference was used to geometrically represent the functional potential per strain with the pLM ProtT5. Both biological layers were used to map host-microbe functional landscapes. **b**, Phylogenetic tree computed from the core genome fraction of the 592 *E. coli* strains panel with tips colored following the major phylogroups. The tree branch length reflects genetic distances. **c**, Pie chart showing the distribution of gene families belonging to the core (>95% presence), shell (95% to 15% presence), or cloud (<15% presence) genome across the 592 *E. coli* strains. **d**, Distribution of *E. coli* phylogroups across the 592 *E. coli* strains. **e**, Representation of the transcriptional distance between two common *E. coli* lab strains belonging to the same phylogroup, MG1655 and OP50, which are known to induce differences in the worm physiology. **f**, Principal component analysis (PCA) of the whole *C. elegans* transcriptional profiles. Each point representing animals raised on a single *E. coli* strain and colored by the phylogroup of the corresponding strain. **g**, PCA of WormCat level 1, level 2 and level 3 categories, depicting the functional landscape of *C. elegans* transcriptional responses to the *E. coli* strain panel. Each point represents the WormCat functional category for a given strain and colored by the phylogroup of the corresponding *E. coli* strain.

Figure 2. The functional landscape encoded in the *E. coli* pangenome can be leveraged to create a functional map of the species. **a**, Pie chart showing the number of gene families assigned to the core (>95% presence), shell (95% to 15% presence), or cloud (<15% presence) genome across the 9 558 *E. coli* strains. **b**, Distribution of *E. coli* phylogroups across the 9 558 *E. coli* strains. **c**, Phylogenetic tree computed from a set of 275 conserved genes from the core genome of the 9,558 *E. coli* strains. Tips are colored following the major phylogroups and the tree branch lengths reflect genetic distances. **d**, PCA of gene presence/absence across the cloud genome of the *E. coli* strain panel. Each point represents a strain, colored by phylogroup. **e**, PCA projection of protein embeddings with Orange/Pink/Green representing genes belonging to the cloud/shell/core genome respectively. **f**, Violin plots with box plots embedded representing the distribution from the genome fractions per principal component (n = 92,244; p<0.001, One-way ANOVA). **g**, PCA projection of protein embeddings for genes belonging to the COG category C, Energy

Production and Conversion. **h**, PCA projection of protein embeddings for genes belonging to the COG category S, Unknown. **i**, Box plots of the PCA coordinates per COG category belonging to the major category of Metabolism and Poorly Characterized (n = 33-18,050). (NS P > 0.05, *P < 0.05, **P < 0.01, ***P < 0.001, two-tailed pairwise T-test with BH). **j**, PCA projection of the 9,558 *E. coli* strain embeddings, colored and split by their respective phylogroups.

Figure 3. *E. coli* strain embeddings map bacterial functional landscape onto host functional responses. **a**, PCA projection for the 592 *E. coli* strain embeddings included in the RNA-seq screen with points colored by phylogroup. **b**, PCA projection of the 592 *E. coli* strain embeddings with dots colored by Methionine functional scores derived from *C. elegans* transcriptomes. (n = 592, Spearman correlation, *P < 0.05, **P < 0.01, ***P < 0.001). **c**, Spearman correlation plots for the two Principal Components with the Methionine functional scores. Spearman correlation coefficient is represented as the ρ value (n = 592). **d**, Bubble plot summarizing Spearman correlation coefficients (ρ) for significant WormCat functional categories and PC2 of the strain-embedding PCA across the 592 strains. **e**, Spearman correlations between the PC2 of the strain embeddings PCA and biologically relevant *C. elegans* gene reporter phenotypes. Each facet shows the relationship between PC2 of the *E. coli* strain-embedding PCA (x-axis) and normalized mean reporter intensity from the high-throughput imaging screen (y-axis) for each *C. elegans* gene reporter. Spearman correlation coefficients (ρ) and significance are represented for each case. **f-g**, Strain-strain Pearson correlation heat map derived from the *C. elegans* gene reporter. Strains are clustered by their phylogenetic distances (**f**) or by their reporter activity (**g**) and colored by phylogroup (n = 589). **h**, Pearson correlation between *C. elegans* gene reporter data shown as circles (lower triangle) and correlation values (upper triangle) (n = 589).

Figure 4. Proteostasis state in *E. coli* regulates proteostasis regulation at the host level. **a**, Normalized brightness of the worm reporter UBV::GFP as a ratio of GFP over mCherry (UPS fluorescence) fed on *E. coli* knock-out for proteins involved in bacterial proteostasis (n=3-8). **b**, Fluorescent microscope images of the worm reporters UBV::GFP fed on *E. coli* significant mutants identified in **a**. **c**, GFP and tubulin (housekeeping protein) quantification with western blot normalized over mCherry for the significant proteins (n=4). **d**, Bi-partite network representation of the proteome derived from the KO *E. coli* gene Δ*dnaK*, Δ*lon*, Δ*clpX* and Δ*cbpM* compared to the control strain. Nodes are bacterial strains (orange) and significant proteins (grey). Edges represent protein expression (red for increased, blue for decreased)

and protein-protein interactions from STRING (grey). **e**, Normalized UPS fluorescence of worms fed on BW25113 and mutants Δ*dnaK*, Δ*clpX* and Δ*dnaK*Δ*clpX* (n=4). **f**, Normalized UPS fluorescence of worms fed on BW25113 and mutants Δ*dnaK* and Δ*clpX* living bacteria (-) and UV-irradiated bacteria (+) (n=5). **g**,**h** KEGG Pathways enriched from the *E. coli* (**g**) and *C. elegans* (**h**) proteomics from the Δ*dnaK*Δ*clpX* vs Δ*dna* comparison. **i**, Heat map of the significant protein expression from the Δ*dnaK versus* Δ*clpX* and Δ*dnaK*Δ*clpX*. **j**, Normalized UPS fluorescence of worms fed on *E. coli* mutants Δ*dnaK*, Δ*dnaK*Δ*tonB*, Δ*dnaK*Δ*clpX*, Δ*dnaK*Δ*clpX*Δ*tonB* (n=5). **k**, Normalized UPS fluorescence of worms fed on *E. coli* BW25113, Δ*dnaK*, Δ*clpX*, Δ*dnaK*Δ*clpX* in the presence (+) or absence (-) of vitamin B12 (150 nM) (n=3-6). **l**, Normalized UPS fluorescence of worms fed on BW25113, Δ*dnaK* and Δ*dnaK*Δ*clpX* supplemented with propionate (0, 1 and 3 mM) (n=3). **m**, Normalized UPS fluorescence of worms fed on BW25113 and strains Δ*scp*, Δ*dnaK*, Δ*dnaK*Δ*scp* (n=5). **n**, Normalized UPS fluorescence of worms fed on BW25113 and strains Δ*dnaK*, Δ*clpX* and Δ*dnaK*Δ*clpX* (n=4-6). **o**, Normalized UPS fluorescence of worms fed on BW25113 for wild-type *C. elegans* N2 (Ctrl) and worm mutants *acdh-1(0)*, *hphd-1(0)*, *mce-1(0)* in the presence (+) and absence (-) of vitamin B12 (150 nM) (n=4). **p**, Normalized UPS fluorescence of worms fed on BW25113 and strains Δ*atoA*, Δ*atoB* with and without acetoacetate (10mM) (n=3). **q**, Scheme showing that proteostasis regulation at the bacterial level regulates the host response and proteostasis status via propionate and vitamin B12. Stats correspond to comparison against the control (one-way ANOVA) and between nested conditions (two-way ANOVA), represented as *P < 0.05, **P < 0.01, ***P < 0.001, NS P > 0.05.

METHODS

RNA sequencing of *C. elegans* fed on PG *E. coli* strains

Around 30 *C. elegans* animals were grown per well in 96-well microtiter plates, each well containing NGM seeded with a distinct *E. coli* strain. On day 1 of adulthood the worms were transferred in a high-throughput manner to clean 96 well plates using INTEGRA Viaflo 96 liquid handler. Worms were washed twice with RNase-free water to remove bacterial traces and flash-frozen in liquid nitrogen. For lysis, we used bead-based mechanical disruption (Bertin Technologies) in RNA lysis buffer (Zymo Research) and on a bench top Eppendorf Thermomixer C at 2000rpm for 20 min at 4ºC. RNA was concentrated and purified with the RNA Clean & Concentrator-96 kit (Zymo Research). Samples were eluted into microtiter plates and stored at -80ºC prior to library preparation. We quantified and checked the integrity of the RNA and selected batches of 48 RNA samples with similar quality to ensure uniform RNA input. To obtain comprehensive coverage of expressed genes in the *C. elegans* host, we employed Lexogen QuantSeq-Pool Sample-Barcoded 3' mRNA-Seq Library Preparation. Each RNA sample was labelled with a unique 12-nucleotide i1 sequence barcode before conversion to cDNA and pooling. Before amplification, each cDNA pool was dual-indexed with 12-nucleotide i5/i7 index sequences. To validate the RNA extraction and library preparation we prepared a test-pool library from conventional *E. coli* laboratory strains. We performed pair-ended sequencing of the test pool on an Illumina MiSeq sequencer and obtained successful demultiplexing. In total, we prepared 16 libraries that were normalized for final pool sequencing based on library quantification by Qubit 3.0 fluorometer and average library size measured by TapeStation 4200. To remove *E. coli* phylogroup representation biases we randomly distributed strains in the different libraries.

Sequencing was performed on an Illumina MiSeq. Sequences were demultiplexed using DRAGEN GenerateFastQ (v3.7.4) by using the i5/i7 barcodes to separate the different libraries. Each library was further demultiplexed by using idemux (v.0.1.6) and by using each library and sample i1 barcode identifiers. Sequences were quality-cleaned with trimmomatic (v.0.39), removing Illumina adapters and dropping sequences below 65 nucleotides. Lexogen recommends trimming the first 12 nucleotides of each read, a step that can be avoided if the aligner used to map the reads can perform soft-clipping, which was our case with salmon (v.1.10.1). Outlier samples with extremely low read count were discarded at this point. Quality-filtered reads were then filtered to exclude samples with less than 4x10^5 reads, resulting in the discard of 59 samples mainly from libraries 14 and 15. 661 samples were kept for posterior analyses (606 unique strains). Reads have been mapped to the cDNA of the *C. elegans* reference genome in Ensembl (version 111) with salmon (v.1.10.2), which performs soft-

clipping, to build the counts matrix and then imported to R with tximport (v.1.28). were analyzed with DESeq2[46] (v.1.40.1) using the phylogroup as the main group and using the library information to avoid possible batch effect Genes that had less than 10 reads in total were discarded. PCA calculations were performed with the plotPCA function from DESeq2 with the data transformed with the vst function. Outliers from the PCA plot from normalized counts but not batch-corrected were removed from the analysis. Batch correction was performed with the removeBatchEffect function from limma[47] (v.3.64.1). Per-gene variance was modelled with the function modelGeneVar from the scran package[48] (v.1.36.0).

To build the transcriptional landscapes from the worm, the curated database from WormCat[13] containing categories at level 1 to 3 was downloaded (Nov-2021 version). The genes belonging to each category per level were summed by using the normalized and batch-corrected reads from the transcriptional profile. A single value was obtained per category, level and strain, which was used for downstream analyses.

E. coli strain landscape selection

The *E. coli* strains were selected from the NCBI genome database. The metadata was downloaded on January 11th, 2024, and was downloaded with the NCBI dataset download tool (v. 16.2.0). The strains were filtered based on the criteria described here. Genomes that were not at the assembly level of “chromosome”, “complete genome” or “scaffold” were removed. The scaffold N50 was used to filter out genomes with a value lower than 150K. CheckM metrics were used to remove genomes with a completeness lower than 95% and a contamination higher than 1%. Mash distances (mash v. 1.1)[49] were calculated pairwise between all genomes, and those strains whose mean distances were higher than 0.05 were removed. Finally, genomes with a sequence length greater than 7Mbp, and/or genomes with a contig count higher than 300 were removed. This resulted in 8,829 assemblies passing all filters. 517 strains were added from the EcoRef collection[10], where evolution-related strains were discarded from the analysis. 212 commensal strains were added from human isolates from Australia[6]. Phylogroups were assigned with the EzClermont v. 0.7.0 tool[50] (tool based on the approach from ClermonTyping[51]), and genomes belonging to class cryptic, U/cryptic and fails were discarded. The final number of genomes was 9,558 assemblies.

Gene annotation and pangenome analysis

Genome annotation was performed with Bakta[52] (v. 1.9.3) using the full database (v. 5.1) using by default parameters. The pangenome was analyzed with Panaroo[53], selecting a strict clean mode and removing invalid genes. Due to the complexity of this pangenome and the computation time, the pangenome was split into 5 parts containing approximately 2000

random genomes each. Each pangenome was calculated using the same parameters. The output from the 5 calculations was merged into the final output using the Panaroo-merge function from the main pipeline. The reference sequences from the gene families were then translated into proteins (using a custom Python script with biopython v. 1.84). Gene presence/absence matrix was used to calculate the Principal Component Analysis shown in the main text by using the PCA function from scikit-learn v. 1.5.1. The pangenome analysis from the strains used for the *C. elegans* transcriptome and the reporter screening was performed independently by leveraging the annotations obtained with Bakta and running the pipeline as a single process this time (592 strains in total).

Gene presence/absence matrix was used to generate the accumulation curve (ACC) for the full *E. coli* panel of 9,558 strains, and to calculate the Heap's law. The ACC was generated by first removing gene families that were present in more than 99% of the strains, then dividing the total gene count into 50 sampling points and randomly picking genomes for each sampling point to count the number of genes. This process was iterated over 5 times. Heap's law was calculated to fit the following equation:

$$P = k * N^{\gamma}$$

P is the pangenome size, N is the number of genomes, and k and γ are the parameters to fit. Heap's law parameters were fitted using the average of the ACC data per point with the nls function in R.

Pairwise strain genetic similarity was measured as the Jaccard similarity between every pair of strains included in the study. It was calculated as the shared genetic content divided by the union of their genetic content between each strain pair.

Phylogeny

From the core genome extracted from Panaroo, a subset of 275 genes present in all strains was used to build the phylogenetic tree for the full *E. coli* panel. To avoid multicopy bias, only one gene per strain was kept for the alignment. For the laboratory strains included in the smaller panel we used the full core genome. The alignment was done using mafft[54] (v. 7. 526), and the tree was constructed with IQ-Tree[55] (v. 2.3.6) with the GTR+I+G substitution model.

GO term prediction

Sequence-based methods

Proteins were classified using two of the most popular sequence-based methods used in the community: InterPro and eggNOG. Search in the InterPro databases was done using interproscan v. 5.59-91.0[16] with by default parameters. The search in the eggNOG database

was done using eggnog-mapper v. 2.1.12[17] with MMseqs2 to look for novel families options enabled. Results from both methods were filtered to remove entries that had an E-value larger than 1e-5.

Machine learning methods

Reference genes from Panaroo were split into 4 smaller files to fit in memory. Proteinfer[18] source code was downloaded from github and function prediction was done by using 5 ensemble models and a reporting threshold of 0.3.

The reference genes from Panaroo were translated into proteins and sequences were clustered with CD-HIT v. 4.8.1 (similarity threshold of 0.98) to remove similar sequences from the dataset, resulting in 55,942 unique clusters. The resulting file was split into 20 smaller files to fit into memory. Proteins were embedded with bio-embeddings pipeline (v.0.2.2), by using the model ProtT5-XL-U50[12] in half-precision mode. Proteins larger than 3,000 amino acids were discarded to fit in memory. Transfer learning was done using available pipelines under bio-embeddings that used goPredSim[19], using Euclidean distances and a k-nearest-neighbors of 3. ProtT5 h5 file was used as a reference with GOA annotations from 2022. Proteinfer, protein embeddings and transfer learning were carried in a computer with 32Gb of RAM and an RTX 4080 GPU with 16Gb of memory.

Information content calculation

To calculate the information content (IC) of the GO terms predicted by the different tools, we used an adaptation of the method from Barrios-Nuñez *et al*[15]. Given that GO terms have a hierarchical structure, the deeper nodes from the branch will contain a higher functional information. Considering that having a deep node in the branch is less likely than to have a higher node with less information, we can approximate the information content of each node by the negative logarithm of the probability for that node to be inferred:

$$IC = log_2(p(t))$$

Where p(t) is the probability for that node, which can be calculated as:

$$p(t) = 1 - \frac{child\ nodes}{child\ nodes + ancestor\ nodes}$$

IC was calculated by joining all the GO term predictions together to create a joint library of terms for the pangenome. Given that we lack a pre-computed list of GO terms with their probabilities as exist for reference organisms, we had to calculate these probabilities from

scratch. We joined together all GO term predictions from the 4 methods and kept the uniquely present GO terms. This allowed us to create a database whereby to filter the resulting steps. To calculate the number of ancestors and child nodes from each term, OWLTools was used (release 2024-06-12). The database used is the go-basic.obo from geneontology.org (accessed in October 2024). From the joint set of unique GO terms we used the OWLTools-Runner function to get both the ancestors and descendants from each node. As the descendants from a node, especially from the ones up in the tree, can have many different child nodes depending on the final function, we removed all the GO terms that were not present in our joint dataset. The probability was calculated as defined but corrected as $p(t) = p(t)(\frac{1}{ancestor})$ for the cases where no descendant was kept in the list, but the GO term did not reach the bottom of the branch from the obo database.

Strain embeddings calculation

Strain embeddings were calculated based on the gene presence/absence matrix generated by Panaroo, the protein embeddings generated by the pLM model ProtT5-XL-BFD, and the number of genes per strain. To calculate any of the different strain embeddings versions described below, we excluded the core genome set of genes, as they were not useful given that all strains shared them.

Strain embeddings were generated using three distinct aggregation methods: 1) direct summation via matrix multiplication; 2) simple averaging, normalized by the gene count per strain to mitigate genome size bias; 3) weighted averaging, which employs an Inverse Gene Frequency (IGF) metric. The three versions can be visualized in Extended Data Fig. S3a, the strain embeddings have been uploaded to Zenodo (https://doi.org/10.5281/zenodo.18221759)

Matrix multiplication

The simplest form was calculated by multiplying the presence/absence matrix with the embedding matrix with the following form:

$$S = A^T \cdot E$$

Where A is the binary matrix of gene presence/absence with nxm dimensions (genes and genomes), E is the matrix of protein embeddings from the representative genes with nxd dimensions (genes and embeddings), and S is the objective strain embeddings with mxd dimensions (genomes and embeddings).

Average strain embeddings

The average strain embeddings were calculated based on how many genes were encoded in each genome and then applying a diagonal normalization on the matrix multiplication equation. The diagonal normalization is a mxm matrix where the diagonal is the inverse of the number of genes per strain, where N_i is the number of present genes in strain i:

$$D_N = diag\left(\frac{1}{N_1}, \frac{1}{N_2}, \ldots, \frac{1}{N_m}\right)$$

Therefore, the average strain embeddings were calculated as:

$$S_{av} = D_N(A^T \cdot E)$$

Where S_{av} is the objective average strain embeddings with dimensions mxd.

Weighted average strain embeddings

Finally, the contribution for each gene to the strain potential was scaled in terms of their proportion, thus, increasing the importance of rare genes to the final position of the strain embedding. That is, genes that are common have a lower weight than the ones that are rarer. To do so, relied on an adaptation of the *Inverse Document Frequency* metric that can be adapted here as the *Inverse Gene Frequency* (IGF).

First, the Strain Count for a gene family (C_i) was defined as the number of strains in which the gene family i was present over the total number of strains (M). This was equivalent to the sum of the *i*-th row of matrix A:

$$C_i = \sum_{j=1}^{M} A_{i,j}$$

We next defined Weight for Gene Family i (W_i) as the logarithm of the relative presence of a specific gene family, where $W_i = 0$ if $C_i = M$ (gene present in all strains):

$$W_i = \log\left(\frac{M}{C_i}\right)$$

We then defined the diagonal matrix with gene weights calculated from last equation as:

$$W = diag(W_1, W_2, \ldots W_n)$$

Here W is a matrix of nxn dimensions. We then used this matrix to calculate the weighted protein embeddings (E_w) as:

$$E_w = W \cdot E$$

Where E_w and E are matrices with nxd dimensions. Then we calculated the weighted sum of embeddings ($S_{weighted\_sum}$) for each strain as:

$$S_{weighted\_sum} = A^T \cdot E_w = A^T(diag(W)E)$$

Where $S_{weighted\_sum}$ is a mxd matrix. We proceeded by calculating the sum of weights for each strain (W_{sum}) as:

$$W_{sum} = A^T \cdot W_{vec}$$

Where W_{vec} is an $n\ x\ 1$ column vector containing the W_i values; and the sum of weights for each strain j is $W_{sum,j} = \sum_{i=1}^{N} A_{i,j} \cdot W_i$

We created another diagonal normalization matrix (D_w, an mxm matrix), where the diagonal elements are the inverse of the sum of weights for each strain:

$$D_W = diag\left(\frac{1}{W_{sum,1}}, \frac{1}{W_{sum,2}}, \ldots, \frac{1}{W_{sum,m}}\right)$$

Finally, we used all these outputs to do the final calculation and got the weighted averaged strain embeddings:

$$S_{w.av} = D_W \cdot S_{weighted_{sum}} = D_W(A^T(W \cdot E))$$

Where $D_W$is a mxm matrix, $S_{weighted\_sum}$ is a mxd matrix, and the product $S_{w.av}$ is a mxd matrix whose each vector row S_j is the weighted averaged embedding for strain j.

Functional mapping onto host phenotype

Average strain embeddings from the 9,558 *E. coli* strains were used to create a Principal Component Analysis in R using the prcomp function. The PCA coordinates were then leveraged to create the pangenome-host functional mapping. The WormCat-aggregated functions at level 3 were mapped onto the PCA coordinates 1 and 2 of the laboratory *E. coli*

strains. The worm functions were correlated to each Principal Component per separate by using Spearman correlation and by correcting the P-value for multiple comparisons with a FDR calculated with the Benjamini-Hochberg method.

High-throughput imaging and data analysis

C. elegans animals were synchronized by standard hypochlorite method, and around 20 L1 worms were transferred to each well of 96-well plates seeded with *E. coli* pangenome strains. The worms were incubated at 20°C until D1 of adulthood and immobilized for imaging with 5 µL of 2% levamisole per well using INTEGRA Viaflo-96. Images were acquired by an automated protocol that captured 10 images at fixed z-heights per well under identical exposure settings using a Zeiss Axio Zoom V16 microscope system equipped with an AxioCam camera operated by Zen 2 software (Zeiss). GFP filter set (excitation: 450-490 nm; emission: 500-550 nm) or the RFP filter set (excitation: 559-585 nm; emission: 600-690 nm) was used depending on the strain being imaged. All images were exported in CZI format, and the most focused z-stack was extracted in Phyton (v. 3.12). Ilastik (v. 1.4) was used to detect worm pixels and to quantify fluorescence levels per worm/cluster of worms.

The fluorescence data was then filtered such that only single worms (ilastik single worm probability > 0.5) with a pixel size between 1000 and 6000 or clustered worms (ilastik clustered worm probability > 0.5) with a pixel size greater than 6000 were retained. Worm mean fluorescence expressed as the brightness per worm as a whole and corrected per size, was corrected against the background for each case. The mean-fluorescence of technical replicates was then averaged, and biological replicates were normalized. Mean-fluorescence was normalized such that, for each worm reporter, the median of each biological replicate was equal to the global median of all biological replicates. Following this, for each worm-bacteria pair the biological replicates, where n > 2 (up to n=5), were subjected to a z-score analysis using scipy.stats.zscore module (SciPy v1.8) and biological replicates with an absolute z-score > 1.5 for mean-fluorescence were removed. Given that most data had only 2 biological replicates and further replicates were only performed in select cases to replace replicates where some wells reduced data quality, from these sets of biological replicates the two replicates with the lowest deviation from each other were carried forward. Where only 1 biological replicate (8.9% of worm-bacteria pairs) was available; these were carried forward alone. Single replicates arise due to empty wells, where reporter worms escape wells and do not appear in the images; however, differences between single replicates and double replicates were broadly inexistent. The log2 ratio between each biological replicate pairs' mean-fluorescence value was then calculated, and the median log2 ratio calculated for each reporter worm. For each reporter worm dataset, a decreasing threshold was iteratively tested

for the maximum allowed log2 ratio between replicates and the maximum deviation from the median log2 ratio. Here, data above the threshold was removed and the Pearson correlation between the biological replicates calculated using scipy.stats.pearsonr module (SciPy v1.8). The threshold was continuously decreased and worm-bacteria pairs removed until a Pearson correlation of >= 0.7 was achieved. The final mean-fluorescence values were then calculated from the average of the biological replicates. Values were then normalized against the median for each reporter case for the tree representation, which was visualized with tidytree (v. 0.4.6), treeio (v. 1.32.0) and phytools (v. 2.4-4).

Pearson correlations of median ratio profiles were calculated for all strain-strain pairs to produce a correlation matrix using pandas.DataFrame.corr (pandas v2.1.4). A phylogenetic distance matrix for strains was hierarchical clustered, using the UPGMA algorithm, to produce a linkage matrix. Hierarchical clustering was performed here using the scipy.cluster.hierarchy.linkage module (SciPy v1.13.1) and prior to clustering the distance matrix was converted from the vector-from to the square-form using scipy.distance.squareform (SciPy v1.13.1). The median ratio correlation matrix was then clustered either using the phylogenetic linkage matrix or by strain-strain correlation profile similarity and displayed as a clustered heatmap (Seaborn v0.13.2, matplotlib v3.10.3).

Pearson correlations of median ratio profiles were then calculated for all strain-strain pairs within the same phylogroup, as above. The percentage of strains with positive or negative correlations within each phylogroup, as well as for the pangenome, were calculated for a range of thresholds between 0 to 1 using steps of 0.05, excluding same-strain pairs. Where correlations were greater than the threshold, they were classed as positively correlated. Where correlations were less than the negative of the threshold, they were classed as negatively correlated. Clustered heatmaps were produced for each phylogroup correlation matrix, hierarchically clustering by strain-strain correlation profile similarity (Seaborn v0.13.2, matplotlib v3.10.3). Chord plots were calculated as the pairwise Pearson correlation and P-values corrected by Benjamini-Hochberg. The significance threshold was set at an alpha of 0.05, and results were represented as the symmetrical relation of the significant correlations existent per phylogroup with the library circlize (v. 0.4.16)

Proteostasis strains and culture conditions

E. coli BW25113 single gene deletion mutants were obtained from the KEIO collection and confirmed by PCR. The reaction was performed with GoTaq mix and the PCR was carried out in a PCRmax Alpha Cycler 2 as follows: 2min at 98°C for the initial activation of enzymes, 30

cycles of 30s at 98°C, 30s at 58°C and 1 min/Kb at 72°C. Each strain was grown in LB broth overnight and 120 µL were plated on nematode growth medium (NGM) plates and kept at 20°C for 2 days.

The *C. elegans* UBV reporter, PP607 (hhIs64[unc-119(+); sur-5::UbV-GFP]; hhIs73[unc-119(+); sur-5::mCherry]) was provided by Hoppe Lab, Germany. This strain allows to quantify the proteasomal activity *in vivo* thanks to the GFP fused to a non-cleavable ubiquitin (UbV-GFP) under the control of the ubiquitous *sur-5* promoter[20,23,25]. The following strains were made for fluorescence studies: FGC72 *nIs470[Pcysl-2::GFP];wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65];* FGC73 *agIs17 [myo-2p::mCherry + irg-1p::GFP] IV;wbmIs67 [eft 3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65]*; FGC74 [rtIs30(pfat-7::GFP);wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65]; FGC76 *wwIs25[Pacdh-2::GFP + unc-119(+)*];*wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65]*; FGC77 *dvIs19 [(pAF15)gst-4p::GFP::NLS] III;wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65];* FGC78 *agIs219 [T24B8.5p::GFP::unc-54 3' UTR + ttx-3p::GFP::unc-54 3' UTR] III;wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65]*; FGC79 *wwIs24 [Pacdh-1::GFP + unc-119(+)];wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65]*; FGC80 *zcIs13[hsp-6::GFP];wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65]*; FGC81 *dvIs70 [hsp-16.2p::GFP + rol-6(su1006);wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65];* FGC82 *mgIs73 [cyp-14A4p::gfp::cyp-14A4 3'UTR + myo-2p::mCherry] V.;wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65];* FGC83 *muIs84 [(pAD76) sod-3p::GFP + rol-6];wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65];* FGC84 *wuIs177 [Pftn-1::gfp lin-15(+)];wbmIs67 [eft-3p::3XFLAG::wrmScarlet::unc-54 3'UTR *wbmIs65]*; FGC89 *acdh-1(ok1489), hhIs72[unc-119(+); sur-5::mCherry], hhIs64 [unc-119(+); sur-5::UbiV-GFP]III*; FGC120 *hphd-1(ok3580); hhIs72[unc-119(+); sur-5::mCherry], hhIs64 [unc-119(+); sur-5::UbiV-GFP]III*; FGC121 *mce-1(ok243) I; hhIs72[unc-119(+); sur-5::mCherry], hhIs64 [unc-119(+); sur-5::UbiV-GFP]III*. Worms were maintained at 20°C, on nematode growth medium NGM seeded with different bacterial strains. We supplemented NGM with homocysteine (final concentration 1 and 5 mmol/L), propionate (final concentration 1 and 3 mmol/L) and cobalamin (vitamin B12, final concentration 150 nmol/L) solubilized in water and filter sterilized.

Double deletion bacterial strain construction

Double gene deletion has been generated by using strains from the KEIO Library[56]. This library is based on the *Escherichia coli* strain BW25113. The Kanamycin cassette has been removed by using the plasmid pCP20. This plasmid encodes the yeast Flp recombinase gene, chloramphenicol and ampicillin resistant genes, and temperature sensitive replication[57]. *E. coli* BW25113 strains containing a single mutation were transformed following the TSS enhanced chemical transformation[58] and were plated on chloramphenicol 30ug/ml incubated at 30°C overnight. Clones were selected and streaked on LB with no selection and LB-Kanamycin (50 µg/mL) plates, incubated at 30°C overnight. Kanamycin sensitive clones were streaked on LB agar plates and incubated at 37°C overnight to stop the replication of the pCP20. Clones were then streaked on LB, LB-Kanamycin and LB-chloramphenicol and incubated at 37°C overnight. Sensitive clones to chloramphenicol and kanamycin were selected and kanamycin cassette removal was confirmed by PCR. From this, we obtained mutant with a single mutation not carrying a kanamycin cassette. The secondary mutations were then extracted from another mutant of the Keio library. The strain of interest was lysed by using P1 phage and transduced in *E. coli* kanamycin sensitive strain according to the protocol from Thomason *et al*. 2007[59,60] and was then selected for his resistance to kanamycin. Finally, the presence of both mutations was confirmed by PCR.

Bacterial growth assay

The optical density (OD) at 595nm was monitored using NuncTM 96-well polystyrene round bottom microwell plates containing LB overnight at 37°C (previously grown overnight in LB and diluted 1,000-fold). Plates were placed in the BioTeK BioSpa 8 automated incubator (Agilent), and OD595 was measured every 30 minutes by the BioTek Citation 3 plate reader (Agilent) for 24h. Growth curves were extracted and area under the curve (AUC) calculated by using an in-house Python code (https://github.com/Cabreiro-Lab/cell_dynamics). Growth curves and stats were performed in Prism 8 (v8.4.0) and in RStudio.

Bacterial overexpression mutant generation

We used strains from the ASKA library, based on the *E. coli* K-12 strain[60]. The expression of the ORF of interest is under the control of an IPTG-inducible promoter (isopropyl β-D-1-thiogalactopyranoside) on the plasmid pCA24N carrying chloramphenicol resistance. Clones overexpressing *btuB* and *tonB* were grown in LB broth supplemented with 30 µg/mL of chloramphenicol at 37°C shaking at 200 rpm, plasmids were then extracted with the kit Miniprep GenElute (Simga Aldrich PLN350) and resuspend in water. Plasmids were transformed into strains of interest using the TSS enhanced protocol[58]. Once the transformation was confirmed by PCR, we grown these strains in LB broth supplemented with 1 mmol/L of IPTG at 37°C shaking at 200 rpm for 16 hours.

UV-irradiation of bacteria

Bacteria strains were irradiated with UV to inactivate them[61]. To prepare UV-irradiated *E. coli*, an overnight culture was grown in LB broth at 37°C with shaking at 200 rpm for 16 hours. A CL-1000 UV crosslinker equipped with UV-B lamps was sterilized by wiping with 70% ethanol and irradiating the chamber for 5 minutes alternatively. The overnight culture was diluted in fresh sterile LB at a 1:3 ratio and placed in petri dishes. Plates were placed inside the UV chamber without lids and irradiated for a total of 60 minutes, swirling every 10 minutes to ensure uniform exposure. To prevent heat shock-induced bacterial death, the chamber was allowed to cool for 5 minutes between intervals. Following UV treatment, bacteria were collected into a new sterile 50 mL Falcon tube, centrifuged at 4000 rpm for 10 minutes at 4°C, and the supernatant was carefully removed. The bacterial pellet was resuspended in LB and placed on NGM plates for worms.

Protein identification and quantification by LC-MS/MS

Bacterial samples preparation

E. coli BW25113, wild type, Δ*lon*::*kan*, Δ*htpG*::*kan*, Δ*dnaK*::*kan*, Δ*clpX*::*kan*, Δ*cbpM*::*kan*, Δ*dnaK*Δ*clpX*::*kan* were grown in LB broth overnight at 37°C shaking 200 rpm. NGM plates were seeded with 120 μL of overnight bacterial cultures and lawns were left to grow at 25°C for 2 days. 5 biological replicates were included per condition. Bacteria were collected from plates with PBS 1X buffer using a sterile glass scraper in Diagenode tubes. Samples were centrifuged at 14000 rpm for 90s at room temperature. The supernatant was removed, and pellets were resuspended with lysis Buffer (8 mol/L urea, 20 mmol/L hepes pH 8). Samples were flash frozen in liquid nitrogen and kept on ice from this point onward. Pellets were then lysed via sonication for 5 minutes at 100% amplitude by using the sonicator waterbath QSonica Q700. Samples were centrifuged at 20000g for 15 minutes at 4°C to separate the cellular debris and proteins. Supernatants containing the extracted protein were transferred to clean tubes and protein concentrations were determined by the Quick start Bradford protein assay (Biorad) at 565 nm. The BSA was used for standard curves. We proceeded to two proteomic analyses, the first one with *E. coli* BW25113, wild type, Δ*lon*::*kan*, Δ*htpG*::*kan*, Δ*dnaK*::*kan*, Δ*clpX*::*kan*, Δ*cbpM*::*kan.* The second one has been proceeded with *E. coli* BW25113 wild type, Δ*dnaK*::*kan*, Δ*clpX*::*kan*, and Δ*dnaK*Δ*clpX*::*kan.*

Worm samples preparation

N2 worms were cultivated on NGM plates seeded with *E. coli* BW25113 wild type for 5 days. Eggs were harvested and L1 were seeded on NGM seeded with bacterial strains of interest that have been incubated 2 days at 25°C. 5 biological replicates were included per condition. After 4 days, worms were harvested and washed 5 times with PBS 1X buffer and transferred in Diagenode tubes. Worms were then resuspended in the lysis buffer (8 mol/L urea, 20 mmol/L hepes pH 8.0). Samples were flash frozen with liquid nitrogen then sonicated 2 times 5 minutes at 100% amplitude by using the sonicator waterbath QSonica Q700. Samples were centrifuged 20 min 20000 rpm 4°C. Supernatants containing the extracted protein were transferred to clean tubes and protein concentrations were determined by the Quick start Bradford protein assay (Biorad) at 565 nm. BSA was used for standard curves.

Sample preparation for bacterial proteomics

Protein samples (100 µg per sample) were processed using an in-solution digestion procedure. Briefly, samples were sequentially reduced and alkylated at room temperature and in the dark, to final concentrations of 10 mmol/L dithiothreitol (DTT) and 50 mmol/L 2-chloroacetamide (2-CAM), respectively. Samples were diluted two-fold for the first analysis with 20 mmol/L HEPES (pH 8.0), reducing the urea concentration to 4 mol/L, and diluted 8-fold for the second analysis, reducing the urea concentration to 1 mol/L. This was followed by the addition of 2 µg of trypsin (Promega, V528A) and incubation overnight for the first analysis. For the second analysis, an initial LysC (Wako, 121-05063) digestion at a 1: 500 proteases to protein ratio, for 5 hours at 37°C. Samples were then further diluted to a final urea concentration of 2 mol/L with 20 mmol/L HEPES (pH 8.0), followed by the addition of trypsin (Serva, 37286.03) at 1:50 protease to protein ratio. Samples were incubated at 37°C for 16 hours. The digestion of the first analysis was stopped by acidification with a final concentration of 1% trifluoroacetic acid (TFA) against 0,2% for the second one and protein digests were desalted using Glygen C18 spin tips (Glygen Corp, TT2C18.96). Tryptic peptides were eluted with 60% acetonitrile, 0.1% formic acid (FA). Eluents and dried by vacuum centrifugation.

Sample preparation for worm proteomics

Protein samples (100µg/sample in 8M urea) were processed using an in-solution digestion procedure. Briefly, samples were sequentially reduced and alkylated at room temperature and in the dark, to final concentrations of 10mM dithiothreitol (DTT) and 50mM 2-chloroacetamide (2-CAM), respectively. Samples were diluted 8-fold with 20mM HEPES (pH 8.0), reducing the urea concentration to 1.5M. This was followed by addition of 2µg of trypsin (Promega, V528A). Samples were incubated over-night at 37ºC. The digestion was stopped by acidification with 10% trifluoroacetic acid (TFA) to a final concentration of 1% and protein digests were desalted

using Glygen C18 spin tips (Glygen Corp, TT2C18.96). Tryptic peptides were eluted with 60% acetonitrile, 0.1% formic acid (FA). Eluents and dried by vacuum centrifugation.

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis

Dried tryptic digests were re-dissolved in 0.1% TFA and each sample injected at 2µg LC-MS/MS analysis was performed using an Ultimate 3000 RSLC nano liquid chromatography system (Thermo Scientific) coupled to a coupled to a Q-Exactive mass spectrometer (Thermo Scientific) via an EASY spray source (Thermo Scientific). For LC-MS/MS analysis re-dissolved protein digests were injected and loaded onto a trap column (Acclaim PepMap 100 C18, 100 µm × 2cm) for desalting and concentration at 8 µL/min in 2% acetonitrile, 0.1% TFA. Peptides were separated on-line to an analytical column (Acclaim Pepmap RSLC C18, 75 µm × 75 cm for the bacterial samples, and C18, 75 µm × 50 cm for the worm samples) at a flow rate of 200 nL/min and 250 nL/min for the bacteria and worm samples respectively). For bacteria samples, peptides were separated using a 120 minutes gradient, 4-25% of buffer B for 90 minutes followed by 25-45% buffer B for another 30 minutes (composition of buffer B – 80% acetonitrile, 0.1% FA). For worm samples, peptides were separated using a 90 minutes gradient, 1-22% of buffer B for 60 minutes followed by 22-44% buffer B for another 30 minutes (composition of buffer B – 75% acetonitrile, 5% DMSO and 0.1% FA). Eluted peptides were analyzed by the mass spectrometer operating in positive polarity using a data-dependent acquisition mode. Ions for fragmentation were determined from an initial MS1 survey scan at 70000 resolution for bacterial samples and 120000 for worm samples, followed by HCD (Higher Energy Collision Induced Dissociation) of the top 12 most abundant ions for bacteria samples and 30 most abundant ions for worm samples at 17500 resolution. MS1 and MS2 scan AGC targets were set to 3e6 and 5e4 for maximum injection times of 50ms and 50ms respectively. A survey scan m/z range of 375 – 1800 was used, normalized collision energy set to 27%, charge exclusion enabled with unassigned and +1 charge states rejected and a minimal AGC target of 1e3. Dynamic exclusion was set to 45-50 seconds.

Data analysis for proteomics

Raw proteomic files were analyzed by using the Perseus software (version 1.6.2.3 for the bacterial samples analysis and version 1.6.10.43 for the worm samples) which is part of MaxQuant to obtain statistical and bioinformatic analysis, as well as for visualization (the perseus computational platform for comprehensive analysis of proteomics data). LFQ intensities were located as columns. The data matrix was filtered based on categorical columns to remove reverse decoy hits, potential contaminants and protein groups which were 'only identified by site'. Gene annotations were done by using *E. coli* K12 (version 20200915)

or *C. elegans* (version 20210628) GOBP, GOMF, GOCC, and KEGG database. Data were log2 transformed. The 5 biological replicates for each mutant were then pooled, compared to each other and visualized as Volcano plots. Volcano plots were generated based on LFQ intensities with the following settings: T-test; side: both; number of randomizations: 250; preserve grouping in randomizations: <none>; FDR: 0.05; S0: 0.1. Then, significant differences between mutants were exported for a Hierarchical clustering analysis (HCA). This was carried out after filtering rows based on a minimum of two valid values in at least one group, Z-scoring of values in rows. The HCA was generated with the following settings for both rows tree and columns tree: distance: Euclidean; linkage: average; constraint: none; preprocess with k-means selected (number of clusters: 300; maximal number of iterations: 10; number of restarts: 1). Further data representation and plotting was carried out in R programming language.

Given that both Δ*clpX* and Δ*dnaK*Δ*clpX* behave in a similar way opposed to the Δ*dnaK* deficient strain, we subtracted the differences between groups to study the proteins that were unique to each cluster. We were specifically interested in the set of proteins that were downregulated in Δ*dnaK* opposed to the upregulated in Δ*dnaK*Δ*clpX*, we used the double mutant as a control and subtracted the proteins found in Δ*dnaK*. Therefore, the effects shown in the distinct proteins between both groups can be described as the unique signature of the differential proteostasis capabilities of both groups. In a similar way, the set of proteins expressed in Δ*dnaK* but not in the other groups was studied using the same logic. Thus, the set of proteins that conferred protein stability was also captured here. Groups were extracted from the significant proteins using R programming language and the UpSet library v. 1.4.0.

Western blot

Worms were grown on plates seeded with *E. coli* BW25113, Δ*lon*::*kan*, Δ*htpG*::*kan*, Δ*dnaK*::*kan*, Δ*clpX*::*kan*, Δ*cbpM*::*kan*, Δ*hscA*::*kan*, Δ*dnaJ*::*kan*, Δ*hybE*::*kan* from the L1 to day1 adult stage at 20°C. 75 worms were collected in 100 μL 1X SDS loading buffer. Then samples were boiled 5 minutes at 95°C at 1400 rpm, sonicated for 5 minutes at 100% of amplitude by using the sonicator waterbath QSonica Q700, and boiled again for 5 minutes at 95°C at 1400 rpm. Samples were then centrifuged for 5 minutes at 14000 rpm. For the western blot, proteins from the lysate worms were separated by size using an Invitrogen precast SDS-Page gel 4-12%. Separated proteins were transferred on a nitrocellulose membrane by a dry blotting system (iBlot 2 dry blotting system) with a setting according to manufacturer's instructions. For the detection of GFP, mCherry and Tubulin, the membranes were probed with primary Mouse monoclonal antibodies anti-GFP at a 1:5000 dilution (clone JL-8), anti-

mCherry at a 1:2000 dilution (clone 1C51), anti-alpha tubulin at a 1:10000 dilution (clone B-5-1-2) respectively. Then membranes were exposed to the secondary antibody, Li-Cor anti-Mouse 800CW/680 from Donkey at a 10000 dilution.

The intensity of each GFP band was normalized by the intensity of its corresponding mCherry and Tubulin bands. 3-4 biological replicates were included per condition. Statistical analysis was done by using a one-way ANOVA with multiple comparisons (Tukey's multiple comparison test) with the software PRISM8 (version 8.4.0).

Nematode fluorescence microscopy

PP607 worms (UBV worms) were cultivated on NGM plates seeded with *E. coli* BW25113 wild type for 5 days. Eggs were harvested and L1 were seeded on NGM previously seeded with bacterial strains of interest incubated 2 days at 25°C. After 4 days at 20°C, a minimum of 11 worms were anesthetized with 2% levamisole on NGM plates and were imaged under a 40x objective using a Zeiss Axio Zoom V16 microscope system equipped with an AxioCam MRm camera operated by Zen 2 software (Zeiss). Either the GFP filterset (excitation: 450-490 nm; emission: 500-550 nm) or the mCherry filterset (excitation: 559-585 nm; emission: 600-690 nm) was used. All images were exported in CZI format and fluorescence levels were quantified using Volocity 5.2 software (PerkinElmer) run on a Surface tablet (Microsoft).

The fluorescence intensity of worms was calculated as the pixel density of the entire cross-sectional area occupied by worms from which the pixel density of the background had been subtracted. 3 independent replicates were carried out with a minimum of 11 worms imaged per condition per replicate. The fluorescence intensity was calculated automatically by setting a minimum threshold intensity that excluded the background.

3-6 biological replicates were included per condition. Statistical analysis was done by using a one-way ANOVA with multiple comparisons (Tukey's multiple comparison test) with the software GraphPad PRISM8 (version 8.4.0).

DATA AVAILABILITY

The raw sequences for the transcription profiles of the mono-association experiments with *C. elegans* and the *E. coli* pangenome reported in this study can be accessed in GSE315953. The raw proteomics profiles reported in the experimental validation can be accessed in PRIDE under the IDs PXD071769, PXD071818 and PXD071867.

ACKNOWLEDGEMENTS

F.C. was supported by the Wellcome Trust/Royal Society (102532/Z/12/Z and 102531/Z/13/A), the DFG German Research Foundation (EXC 2030 -390661388) and (IPFP-B02 Filipe Cabreiro) of the Center for Molecular Medicine Cologne. T.H is supported by Research Unit FOR5762 (project HO 2541/18-1 to T.H.) and (ERC, Cellular PQCD, 101141579). D.M.M was supported by the Leverhulme Trust (RPG-2022-299). We acknowledge computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232). We would like to thank Saul Moore for helping with developing a method to select focused microscope images, Jennifer Van der Laan and Evgeny Galimov for technical support. We would like to express our gratitude to our colleagues and to Dr. Mária Džunková which helped to shape this manuscript.

AUTHOR CONTRIBUTIONS

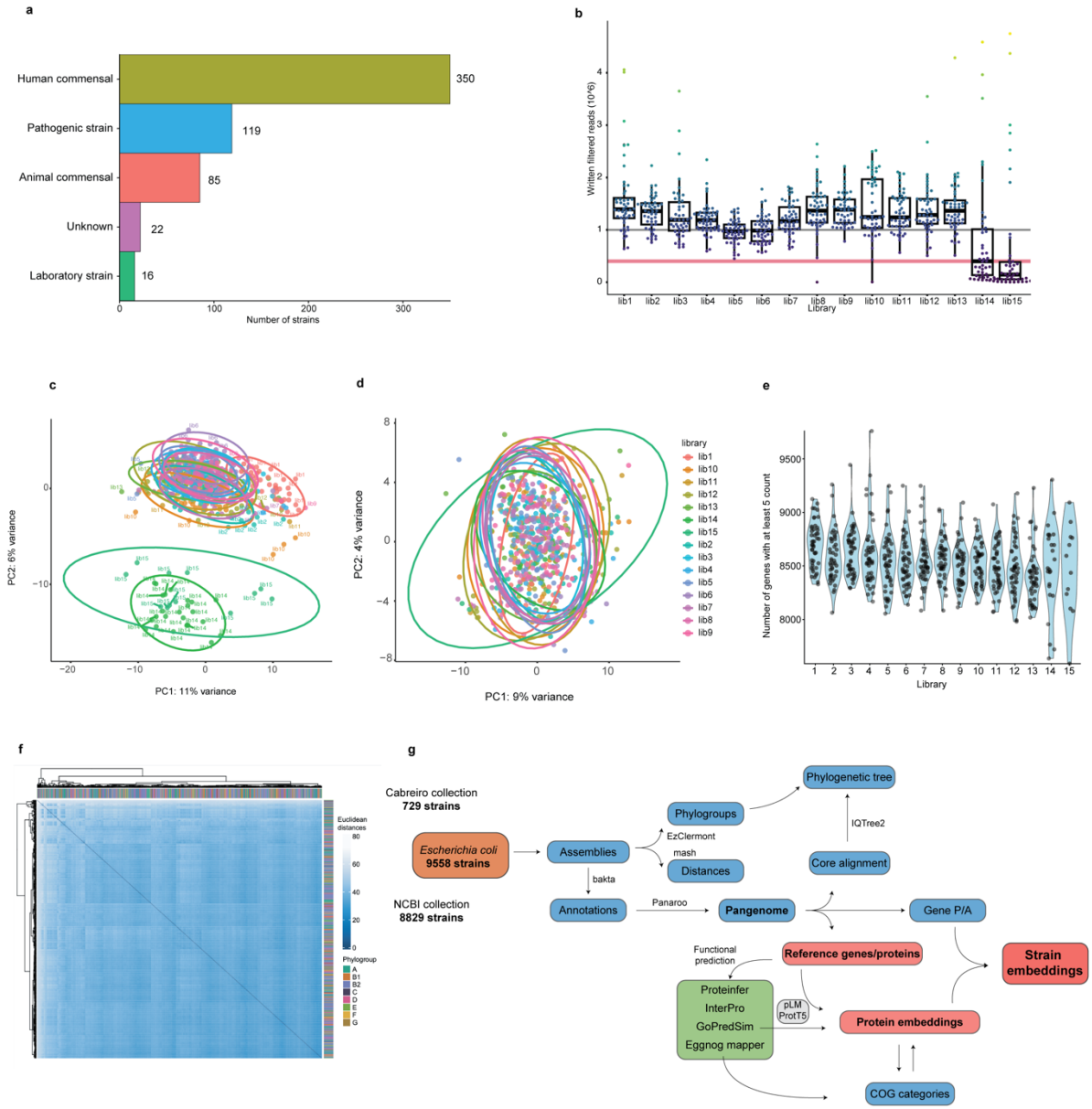
D.M.M., F.C. conceptualized the research. D.M.M. performed the computational analysis of the pangenome. D.M.M, A.A., C.B., H.M.D., J.W. and F.C. analyzed the data. A.A. performed the reporter validation and transcriptome studies. C.B., A.Z., F.O., J.W. and F.C performed the experiments for the worm proteostasis. A.I. and L.G. sequenced transcriptomes and genomes. I.K., G.R, A.M. and H.K. performed proteomics identification and analysis. D.M.M., and F.C. wrote the manuscript. D.M.M., C.B., A.A. and F.C. participated in editing the manuscript. D.M.M., T.H. and F.C. participated in the interpretation of the main findings. D.M.M and F.C. supervised the research. All authors read and approved the final manuscript.

COMPETING INTERESTS

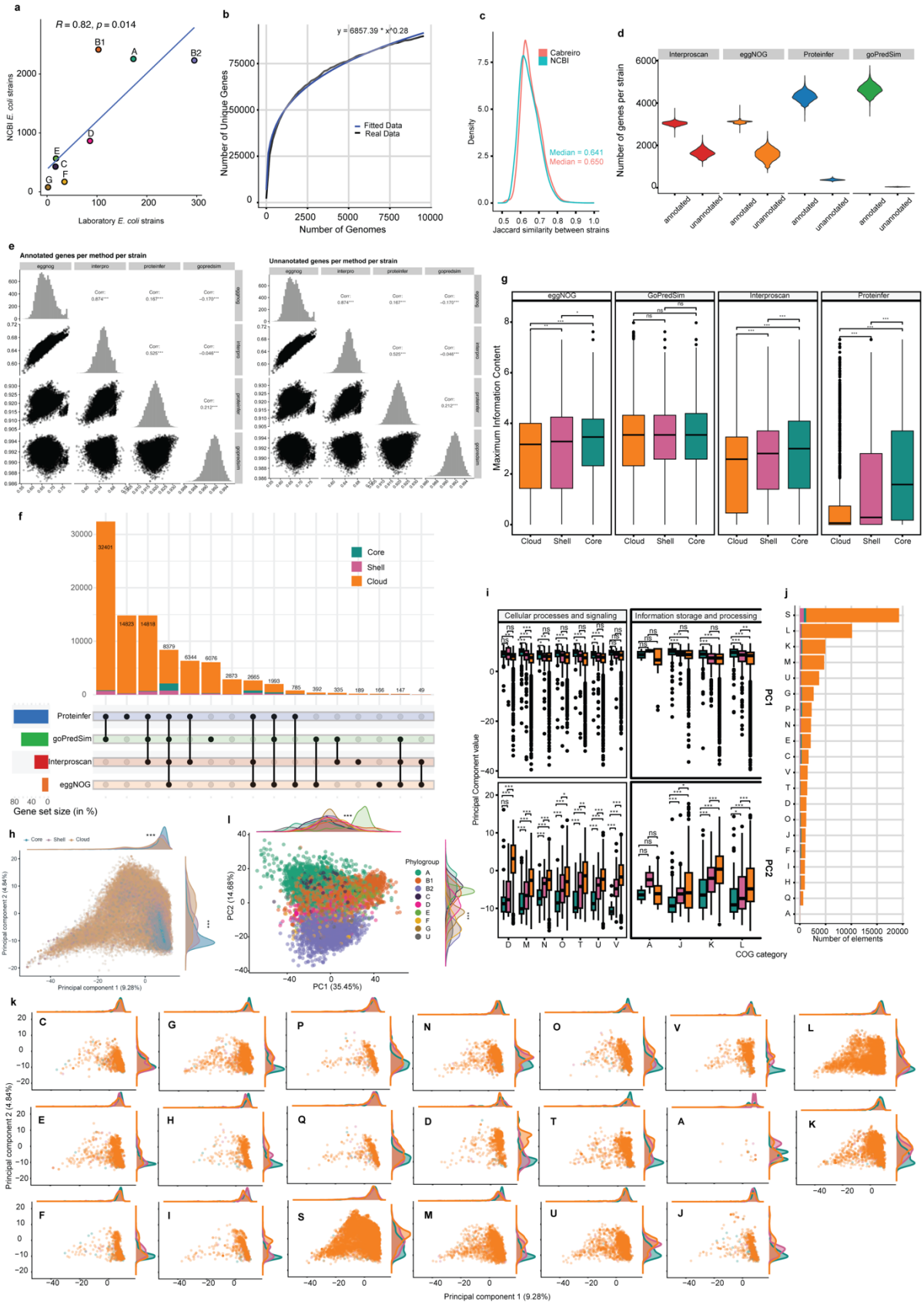
The authors declare no competing interests.

EXTENDED DATA FIGURES

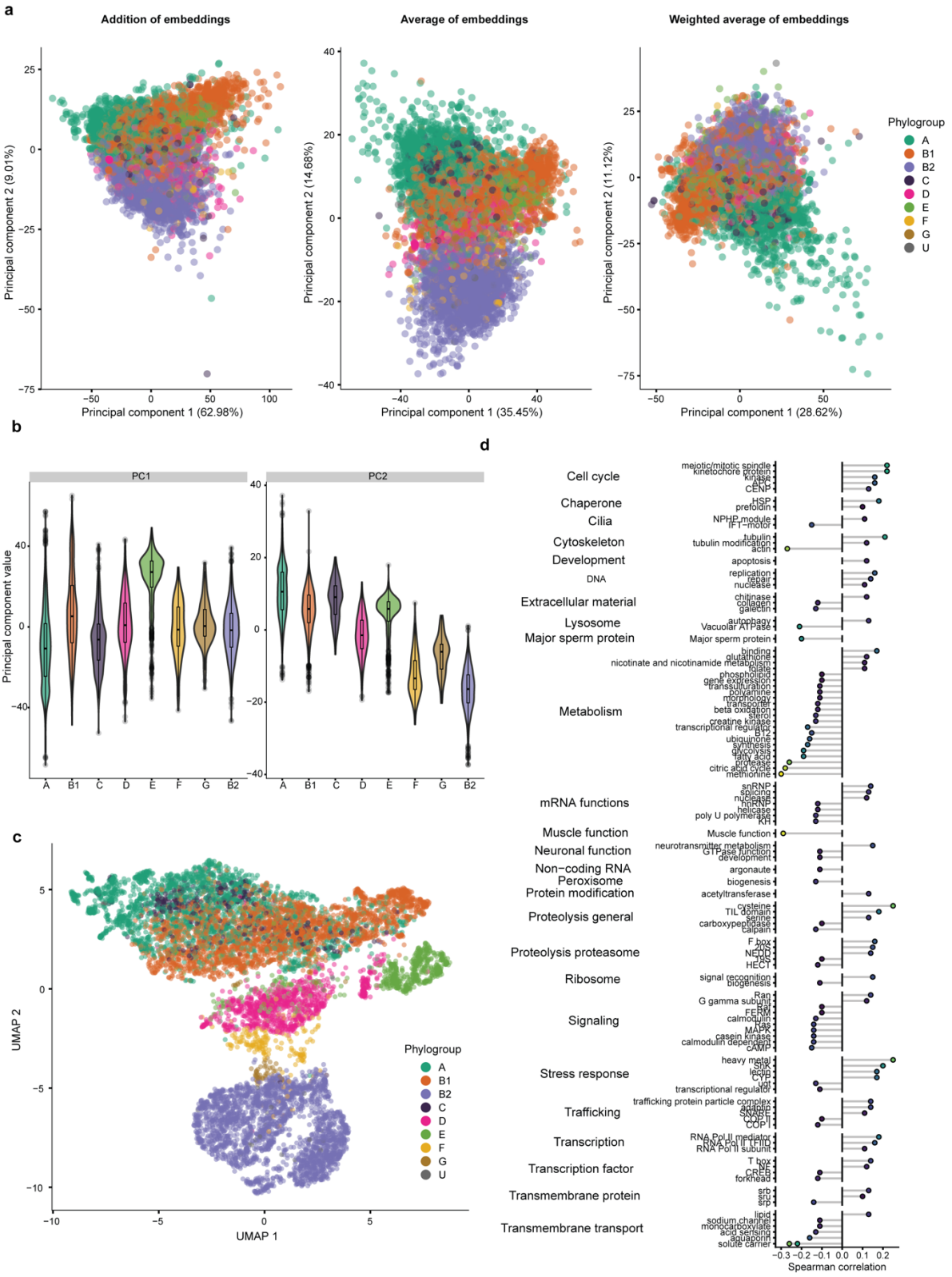
Extended Data Figure 1



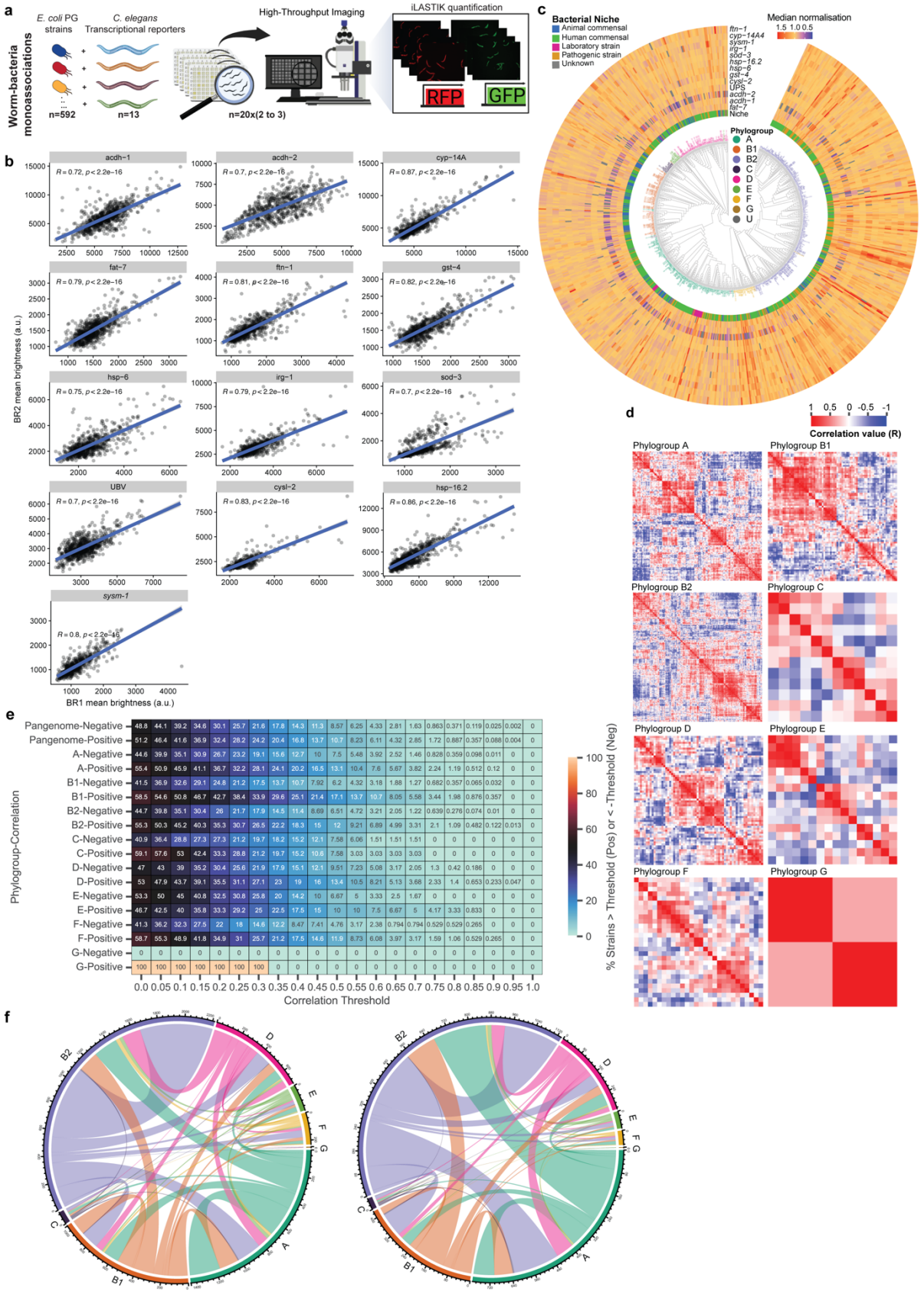
Extended Data Figure 2



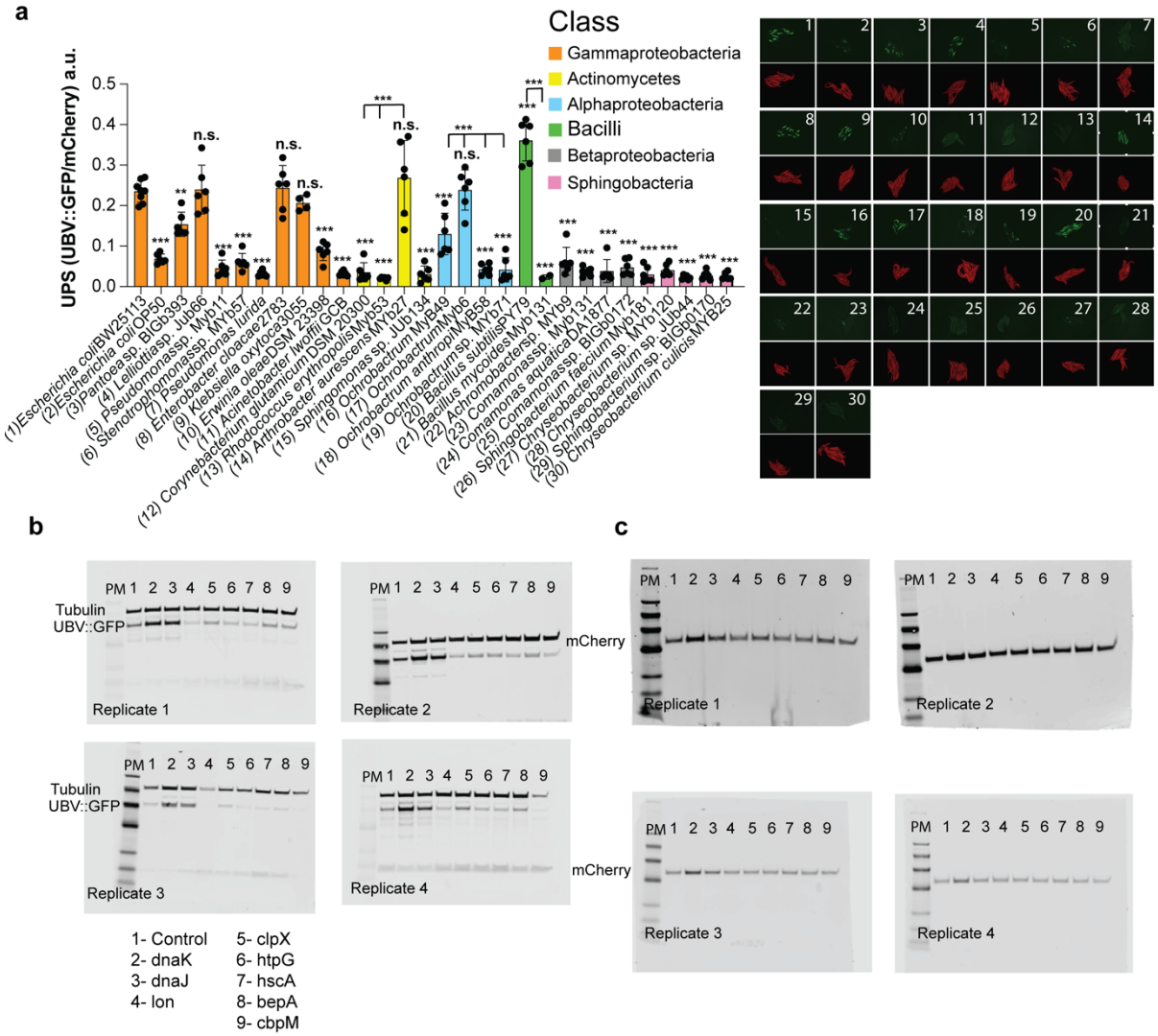
Extended Data Figure 3



Extended Data Figure 4

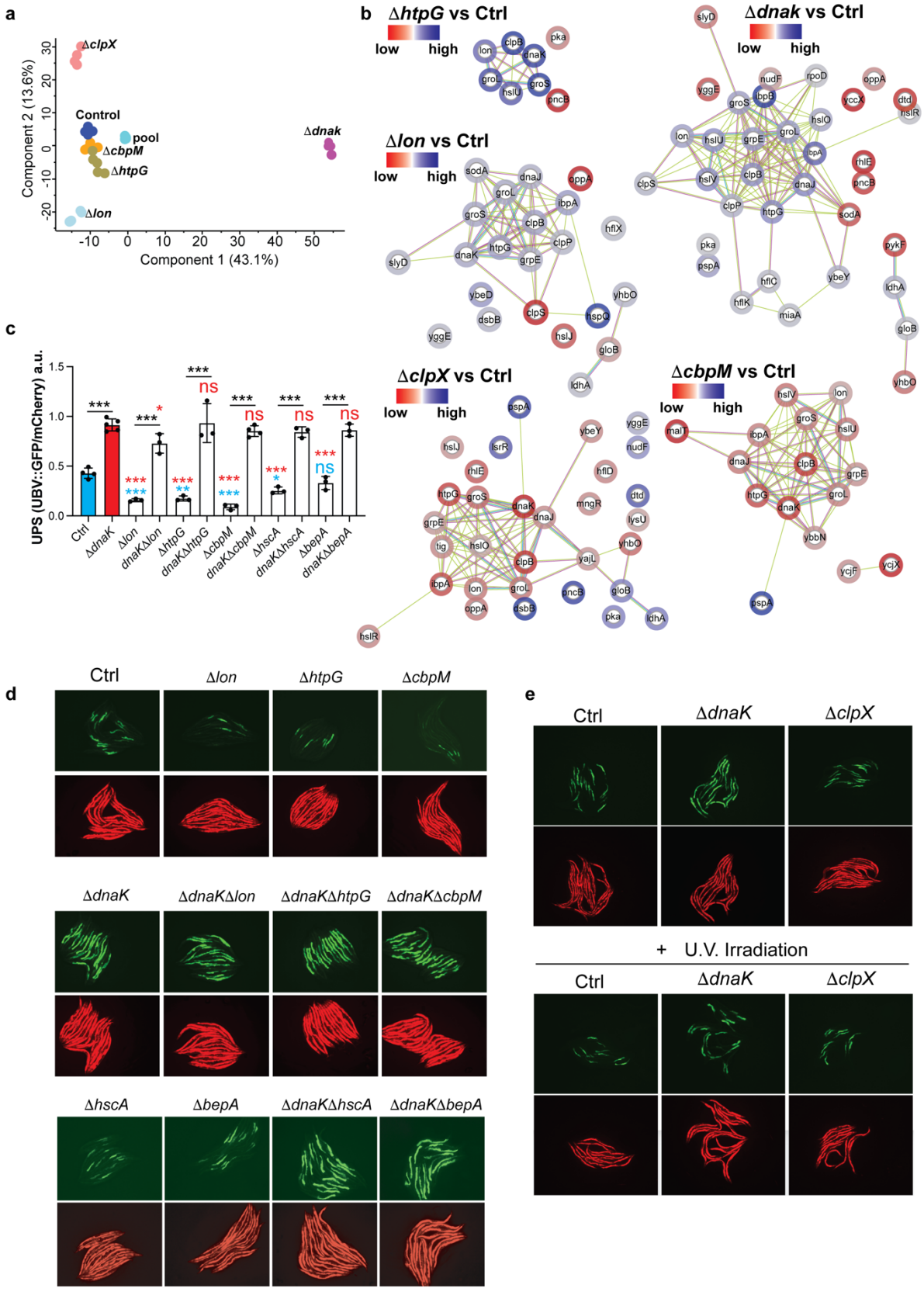


1342 **Extended Data Figure 5**



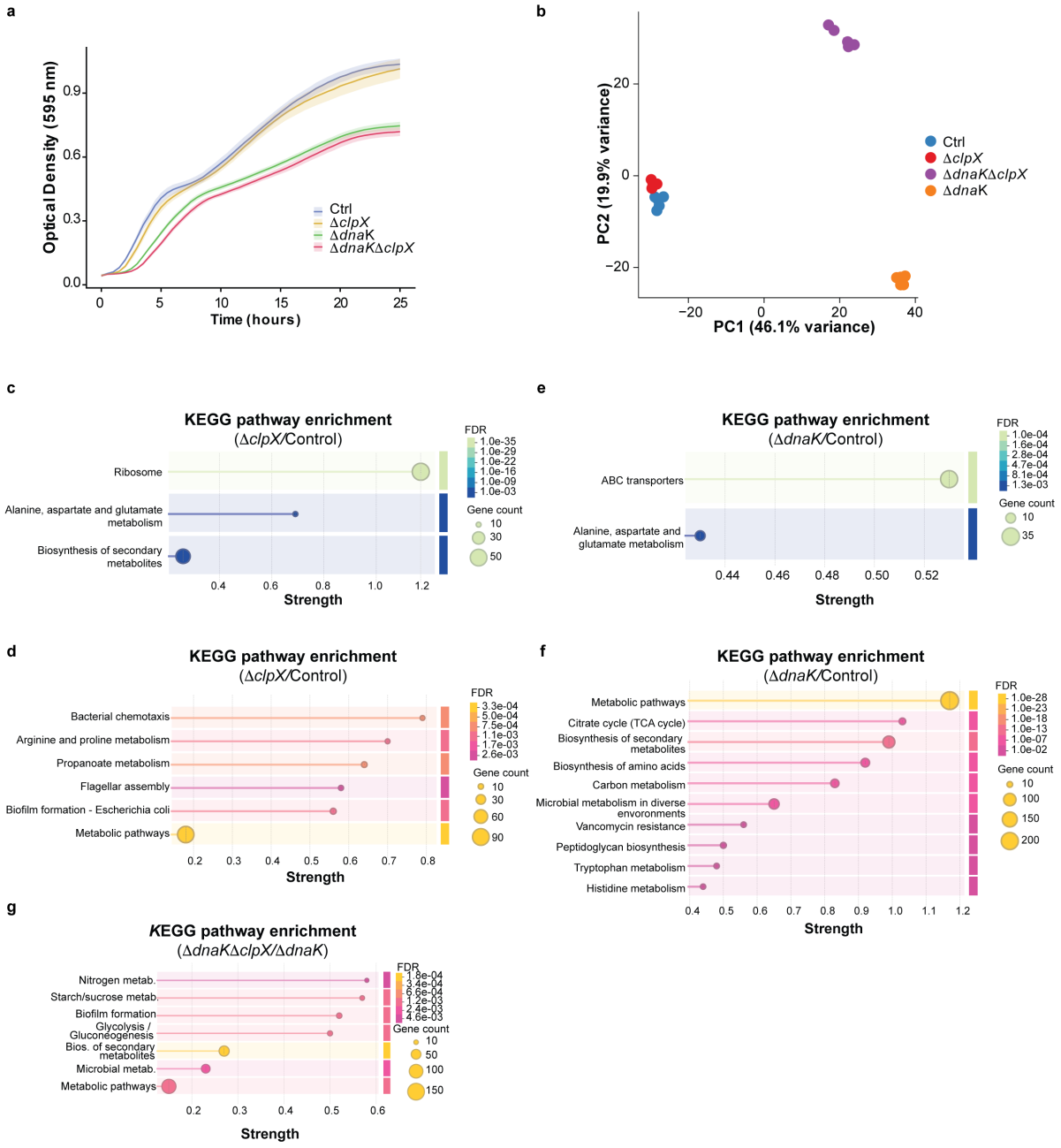
1343

1344 **Extended Data Figure 6**



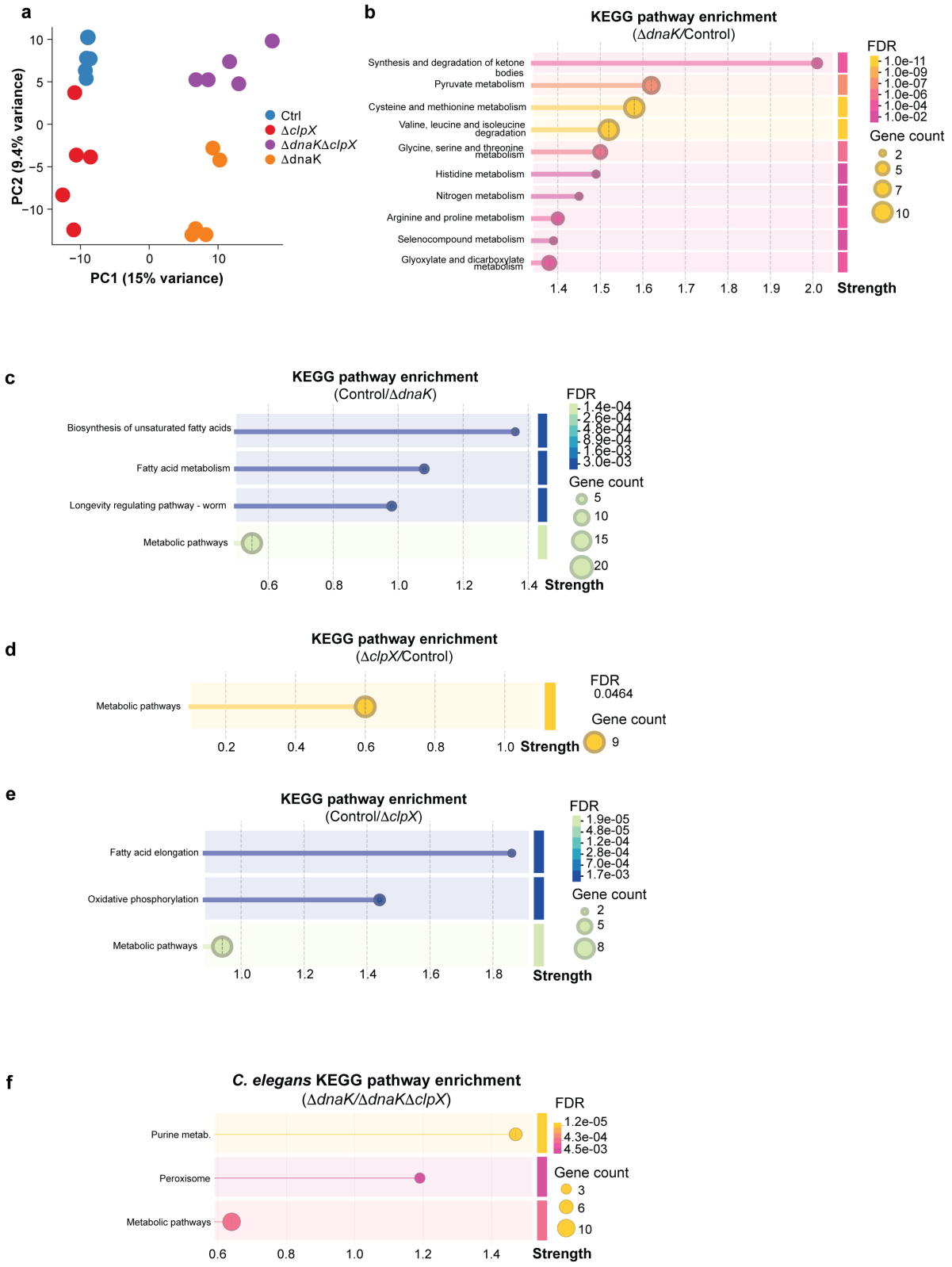
1345

1346 **Extended Data Figure 7**

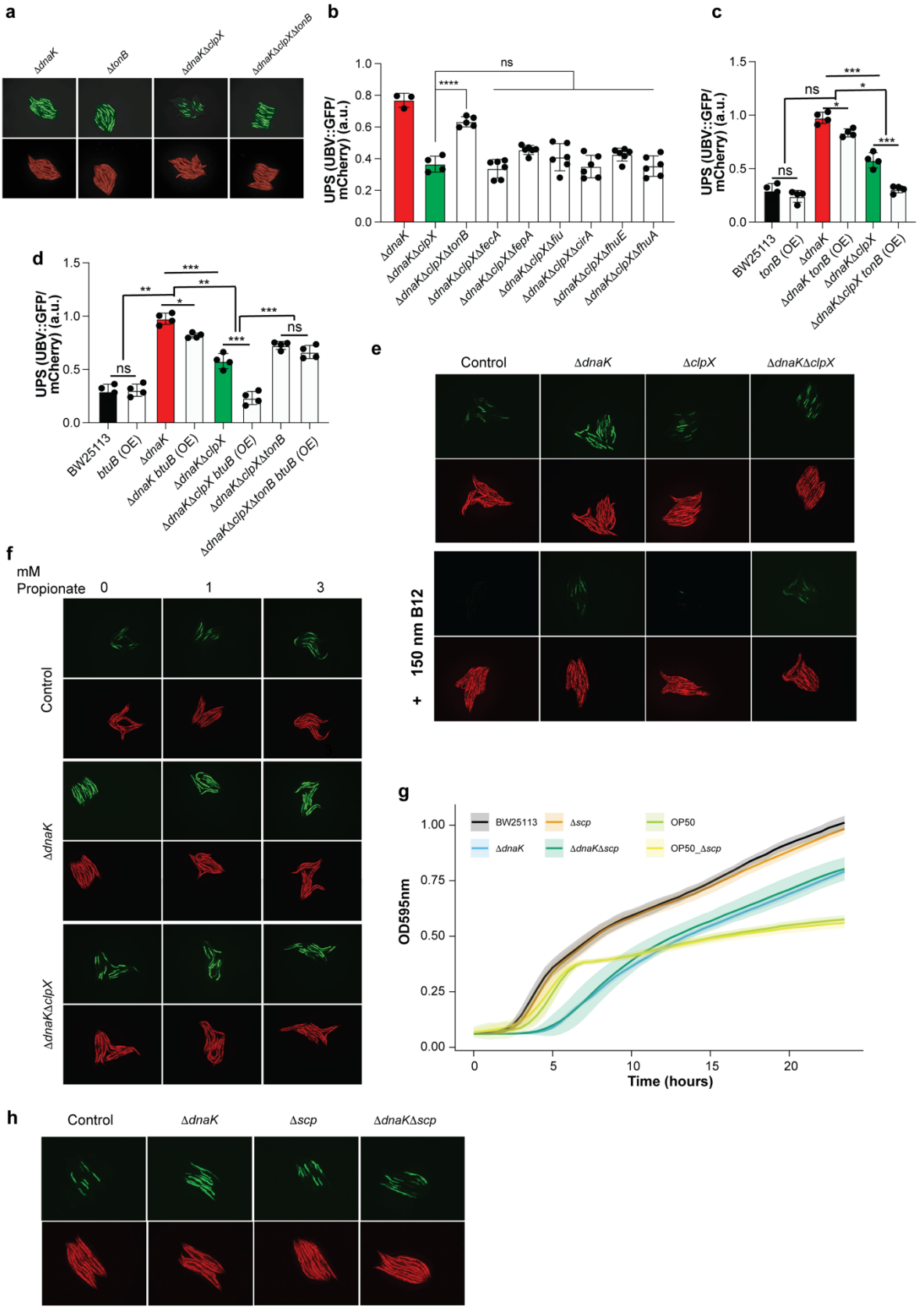


1347

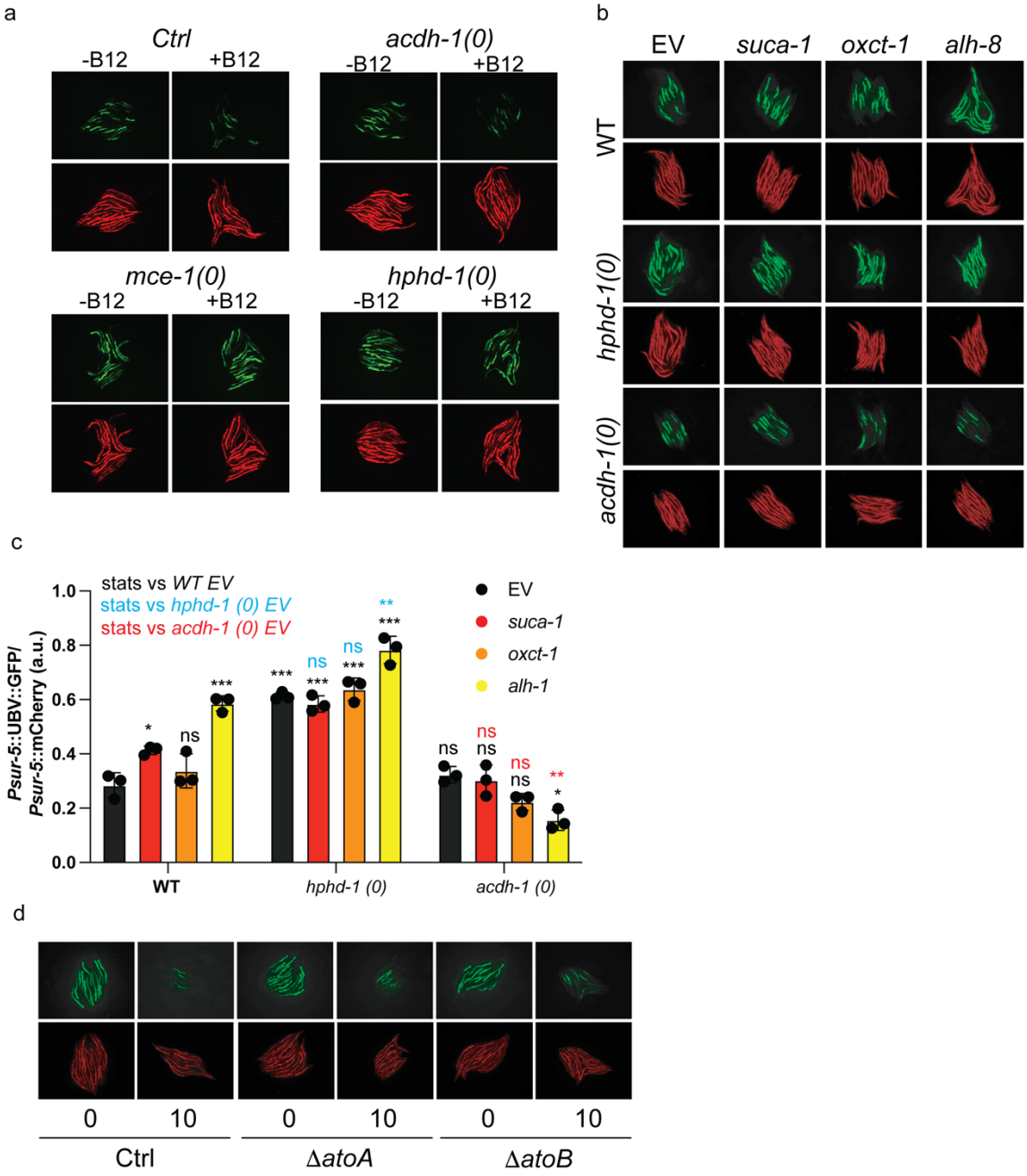
Extended Data Figure 8



Extended Data Figure 9



Extended Data Figure 10



EXTENDED DATA FIGURE/TABLE LEGENDS

Fig. S1 Strain composition and RNA-seq quality control. **a**, Number of *E. coli* strains in the panel classified as human commensal, human pathogenic, animal commensal, laboratory strains or of unknown origin. **b**, Read-count distribution per library for the high-throughput RNA-seq experiment, generated on a NextSeq 2000 P3 run (1.2×10^9 total reads). **c**, PCA of *C. elegans* transcriptomes after outlier removal, showing batch-driven separation with samples from libraries 1, 14 and 15 forming distinct clusters. **d**, PCA of the same dataset after batch correction, showing no library-driven separation. **e**, Violin plot of the number of genes detected per worm with at least 5 normalized counts. Each dot represents a worm sample; data is represented by sequencing library. **f**, Heat map of pairwise Euclidean distances based on DESeq2-normalized *C. elegans* gene expression, with hierarchical clustering of libraries with dark blue-to-light blue showing close-to-distant relationships between samples. PCA data and Euclidean distances are showing the VST normalized data from the transcriptional profiles.

Fig. S2 Functional annotation performance and embedding structure of the *E. coli* pangenome. **a**, Phylogroup frequency Pearson correlation between the laboratory *E. coli* trains and an NCBI *E. coli* collection (R = 0.82, P = 0.014). **b**, Pangenome accumulation curve for the 9,558 *E. coli* genomes. The x-axis shows the number of genomes progressively added to the analysis, and the y-axis shows the cumulative number of unique genes observed with increasing pangenome size. **c**, Distribution of the pairwise Jaccard similarity between *E. coli* strains from NCBI and the Cabreiro lab collection. Median values per collection are represented. **d**, Violin plots representing the fraction of genes annotated by different functional annotation tools: Interproscan, eggNOG-mapper, Proteinfer and GOPredSim. **e**, Pearson correlation of annotated (left) and unannotated (right) fraction per genome and per method (n=92,435). **f**, UpSet diagram of the number of genes with a GO term annotation by each one of the functional annotation methods. Dots in the lower part describe which tool or tools are being considered in each case. Genome partition is represented as a color in the bar-plot. **g**, Maximum information content for the GO terms annotated by each method per genome partition (n=1,347-60,152). **h**, PCA projection of the protein embeddings from the linear reference from the *E. coli* pangenome. Colors represent the genome fractions of core, shell and cloud (n=92,244). **i**, Principal component values of gene embeddings stratified by COG functional category and pangenome class (n=2-9,802). **j**, Distribution of genes across COG categories, partitioned into core, shell and cloud components. Bars show, for each COG

category, the number of genes assigned to the core, shell and cloud genome. **k**, PCA projections of protein embeddings split per genes classified in the different COG categories. **l**, PCA projections of the strain embeddings colored by the main *E. coli* phylogroups (n=9,558). Data shown in **g**, **i**, **h** and **l** were tested with two-tail pairwise T-test and significance is expressed as *P < 0.05, **P < 0.01, ***P < 0.001, NS P > 0.05.

Figure S3. Construction and structure of *E. coli* strain embeddings. **a**, PCA of strain embeddings for the 9,558 *E. coli* strains, built using three aggregation strategies: (left) addition of gene embeddings, (middle) average of gene embeddings and (right) weighted average of gene embeddings (see Methods for a detailed description). Points are colored by phylogroup. **b**, Violin plots showing the distributions of PC1 (left) and PC2 (right) coordinates of strain embeddings, calculated as the average of gene embeddings, across phylogroups. **c**, UMAP projection of strain embeddings for all 9,558 *E. coli* strains, revealing distinct clusters colored by phylogroup. **d**, Bubble plot of the significant Spearman correlation coefficients (ρ) between PC2 of the *E. coli* strain embeddings and WormCat functional scores across all significant worm functional categories.

Figure S4. Associations between strain embeddings, WormCat functional responses and gene reporter phenotypes. **a**, Scheme showing the high-throughput experimental design to analyze the *C. elegans* gene reporters. **b**, Pearson correlation representation of two biological replicates per gene reporter in *C. elegans* (n=589). **c**, Phylogenetic tree of the *E. coli* strain panel annotated with ecological niche and *C. elegans* fluorescent reporter responses. The innermost layer indicates the origin of each strain (human commensal, human pathogenic, animal commensal, laboratory strain or unknown). Outer layers show median normalized fluorescence ratios for each *C. elegans* gene reporter, mapped onto the corresponding *E. coli* strain tips. **d**, Pairwise Pearson correlation coefficients between *E. coli* strains separated by phylogroup (n=589). **e**, Heat map representing the percentage of positive and negative correlations within *E. coli* phylogroups given a range of correlation thresholds (x-axis) (n=589). **f**, Chord plots representing the within and between strain correlations between the main *E. coli* phylogroups for the positive (left) and positive (right) correlations (n=8-13).

Figure S5. Proteostasis at the host level is regulated by bacterial chaperones. **a**, Left, normalized brightness of the worm reporters UBV::GFP over mCherry for worms fed on

several bacterial species, where color represents each bacterial phylum. Right, representative fluorescence images from the worm fed on each bacterial species, measuring UBV::GFP and mCherry worm reporters. Correspondence between the two parts is done by a numeric code. (n=2-8) **b-c,** Western blot analysis of Tubulin-UBV::GFP (**b**) and mCherry (**c**) expression in *E. coli* chaperone and protease mutants. Each replicate (1-4) shows protein expression in various *E. coli* mutants: 1- BW25113 (control), 2- Δ*dnaK*, 3- Δ*dnaJ*, 4- Δ*lon,* 5- Δ*clpX,* 6- Δ*htpG*, 7- Δ*hscA*, 8- Δ*bepA*, and 9- Δ*cbpM*. Tubulin serves as a loading control. **d**, Ratio of the quantification of UBV::GFP over mCherry expression (n=4) (*P < 0.05, **P < 0.01, ***P < 0.001, NS P > 0.05, one-way ANOVA).

Figure S6. Bacterial chaperones and proteases drive proteostasis regulation in the host. a, Principal Component Analysis of the protein expression profile from *E. coli* BW25113 control and Δ*dnaK*, Δ*clpX*, Δ*lon*, Δ*htpG* and Δ*cbpM* mutants. **b**, Network representation from differentially expressed proteins from Δ*dnaK*, Δ*clpX*, Δ*lon*, Δ*htpG* and Δ*cbpM* mutants, protein-protein interactions extracted from STRING database. Colours represent high (blue) and low (red) expression compared to the control strain BW25113. **c**, Normalised brightness of the worm reporters UBV::GFP over mCherry for worms fed on several bacterial chaperone and protease mutants. Stats are represented as coloured stars, black for the double Δ*dnak* mutant vs Δ*dnaK* single mutant, blue for the comparison against the control strain BW25113, and red for the comparison against Δ*dnaK* mutant. (n=3-6, *P < 0.05, **P < 0.01, ***P < 0.001, NS P > 0.05, One-way ANOVA). **d**, Representative fluorescence images from the worm fed on each bacterial mutant tested in **a** and **c**. **e**, Representative fluorescence images from the worm fed on BW25113 (Control) and Δ*dnaK* and Δ*clpX* mutants living cells (top) and UV-irradiated cells (bottom). Fluorescence was measured for the UBV::GFP and mCherry worm reporters.

Figure S7 Growth curves and enrichment of transcriptional responses in *E. coli* chaperone mutants. **a**, Growth curves showing optical density at 595 nm plotted over time (hours) of wild-type BW25113 (control), Δ*clpX*, Δ*dnaK* and Δ*dnaK*Δ*clpX* double-mutant *E. coli* strains. **b**, PCA of the protein expression of *E. coli* strains BW25113 control and Δ*clpX*, Δ*dnaK* and Δ*dnaK*Δ*clpX* mutant strains. **c-d**, STRING-based KEGG pathway enrichment for genes differentially expressed in the *E. coli* Δ*clpX* mutant versus control, highlighting significantly upregulated (**c**) and downregulated (**d**) enriched pathways. Colour represents the FDR values and circle size the number of genes per category. **e-f**, STRING-based KEGG pathway enrichment for genes differentially expressed in the Δ*dnaK* mutant versus control, highlighting significantly upregulated (**e**) and downregulated (**f**) enriched pathways. Colour represents the FDR values and circle size the number of genes per category. **g**, STRING-based KEGG pathway enrichment for genes differentially expressed in the Δ*dnaK*Δ*clpX* double mutant

compared with the Δ*dnaK* single mutant, highlighting significantly upregulated enriched pathways.

Figure S8. Worm proteomics show an increase in metabolic pathways. a, Principal Component Analysis of the protein profiles of *C. elegans* fed on *E. coli* BW25113 (Control) and mutants Δ*dnaK*, Δ*clpX* and Δ*dnaK*Δ*clpX*. **b-c**, STRING-based KEGG pathway enrichment for genes differentially expressed in *C. elegans* fed with *E. coli* Δ*dnaK* mutant versus control, highlighting significantly upregulated (**b**) and downregulated (**c**) enriched pathways. Colour represents the FDR values and circle size the number of genes per category. **d-e**, STRING-based KEGG pathway enrichment for genes differentially expressed in *C. elegans* fed with *E. coli* Δ*dnaK* mutant versus control, highlighting significantly upregulated (**d**) and downregulated (**e**) enriched pathways. Colour represents the FDR values and circle size the number of genes per category. **f**, STRING-based KEGG pathway enrichment for genes differentially expressed in *C. elegans* fed with *E. coli* Δ*dnaK* versus Δ*dnaK*Δ*clpX.* Colour represents the FDR values and circle size the number of genes per category.

Figure S9. Propionate and Vitamin B12 impact bacterial proteostasis. **a**, Representative fluorescence images of the UBV::GFP and mCherry reporters from the worm fed on *E. coli* chaperone mutants Δ*dnaK,* Δ*tonB*, Δ*dnaK*Δ*clpX* and Δ*dnaK*Δ*clpX*Δ*tonB*. **b**, Fluorescence quantification of the reporters UBV::GFP and mCherry ratio in worms when fed on *E. coli* KO mutant strains (n=3-5). **c**, Fluorescence quantification of the reporters UBV::GFP and mCherry ratio in worms when fed on control *E. coli* BW25113 and mutants Δ*dnaK,* Δ*dnaK*Δ*clpX*. Bacterial strains were supplemented with an over-expression (OE) plasmid in all conditions to test for fluorescence differences (n=4, *P < 0.05, **P < 0.01, ***P < 0.001, NS P > 0.05, Two-way ANOVA). **d**, Fluorescence quantification of the reporters UBV::GFP and mCherry ratio in worms when fed on control *E. coli* BW25113 and mutants Δ*dnaK,* Δ*dnaK*Δ*clpX,* Δ*dnaK*Δ*clpX* Δ*tonB.* Bacterial strains were supplemented with an over-expression (OE) plasmid in all conditions to test for fluorescence differences (n=4, *P < 0.05, **P < 0.01, ***P < 0.001, NS P > 0.05, Two-way ANOVA). **e-f**, Representative fluorescence images of the UBV::GFP and mCherry reporters from the worm fed on control *E. coli* BW25113 and mutants Δ*dnaK,* Δ*clpX*, Δ*dnaK*Δ*clpX* in control conditions and when supplemented with 150nM of vitamin B12 (**e**) and 1-3mM of propionate (**f**). **g**, Growth curves showing optical density at 595 nm plotted over time (hours) of wild-type BW25113 (control), Δ*dnaK*, Δ*scp*, Δ*dnaK*Δ*scp* double-mutant. *E. coli* OP50 and the OP50 mutant Δ*scp* correspond to the green and yellow lines. **h**, Representative

fluorescence images of the UBV::GFP and mCherry reporters from the worm fed on control *E. coli* BW25113 and mutants Δ*dnaK*, Δ*scp* and Δ*dnaK*Δ*scp*.

Figure S10. *C. elegans* B12 shunt pathway drives propionate metabolism and UPS impairment. a, Representative fluorescence images of the UBV::GFP and mCherry reporters from *C. elegans* N2 strain (control) and worm mutants *acdh-1*(0), *mce-1*(0) and *hphd-1*(0) fed on *E. coli* BW25113 in control conditions and when 150nM B12 was supplemented. **b**, Representative fluorescence images of the UBV::GFP and mCherry reporters from *C. elegans* N2 strain (Control), *acdh-1*(0), *hphd-1*(0) in combination with KO Empty Vector (EV), *suca-1*, *oxct-1*, *alh-8* strains fed on *E. coli* BW25113. **c**, Fluorescence quantification of the reporters UBV::GFP and mCherry ratio in the worm strains from **b**. Stars describe the significance, the colour describes to what control have they been tested (n=4, *P < 0.05, **P < 0.01, ***P < 0.001, NS P > 0.05, one-way ANOVA). **d**, Representative fluorescence images of the UBV::GFP and mCherry reporters from C. elegans N2 strain fed with control E. coli BW25113, Δ*atoA* and Δ*atoB* mutants in control conditions and when 10 mM of acetoacetate were supplemented.