



Sveučilište u Zagrebu

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY

Anja Tušar

**EVOLUTIONARY PATTERNS IN THE BIOFILMS OF
PATHOGENIC *Escherichia coli* UTI89 STRAIN**

DOCTORAL THESIS

Zagreb, 2026



University of Zagreb

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY

Anja Tušar

**EVOLUTIONARY PATTERNS IN THE BIOFILMS OF
PATHOGENIC *Escherichia coli* UTI89 STRAIN**

DOCTORAL THESIS

Supervisors:

dr. rer. nat. Tomislav Domazet-Lošo, Professor

dr. sc. Momir Futo, Assistant Professor

Zagreb, 2026



Sveučilište u Zagrebu

PRIRODOSLOVNO-MATEMATIČKI FAKULTET
BIOLOŠKI ODSJEK

Anja Tušar

**EVOLUCIJSKA OBILJEŽJA BIOFILMOVA
PATOGENOG BAKTERIJSKOG SOJA *Escherichia coli*
UTI89**

DOKTORSKI RAD

Mentori:

prof. dr. sc. Tomislav Domazet-Lošo

doc. dr. sc. Momir Futo

Zagreb, 2026

This doctoral thesis was produced by the author in the Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, under the supervision of dr. rer. nat. Tomislav Domazet-Lošo, Professor, and dr. sc. Momir Futo, Assistant Professor, as part of the university postgraduate doctoral studies in Biology at the Division of Biology, Faculty of Sciences, University of Zagreb. This work has received funding from European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 955626.

Ovaj je doktorski rad izrađen u Laboratoriju za evolucijsku genetiku Zavoda za molekularnu biologiju Instituta Ruđer Bošković, pod vodstvom prof. dr. sc. Tomislava Domazeta-Loše i doc. dr. sc. Momira Fute, u sklopu Sveučilišnog poslijediplomskog dokorskog studija Biologije pri Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu. Projekt u sklopu kojeg je doktorski rad izrađen financiran je programom Europske unije za istraživanje i inovacije Horizon 2020 prema ugovoru o dodjeli bespovratnih sredstava Marie Skłodowska-Curie br. 955626.

SUPERVISORS' INFORMATION

Tomislav Domazet-Lošo is a group leader at the Ruđer Bošković Institute and a Professor at the School of Medicine, School of Medicine, Catholic University of Croatia in Zagreb. In 1997, he graduated with a degree in Biology from the Faculty of Science, University of Zagreb, Croatia. He earned his doctorate in Evolutionary genetics from the University of Cologne, Germany, in 2003. Between 2008 and 2011, he held postdoctoral positions at the Max Planck Institute for Evolutionary Biology and at the Zoological Institute, University of Kiel, Germany. Tomislav Domazet-Lošo's research lies at the interface of evolutionary genetics and evolutionary developmental biology across the tree of life, with a particular focus on macroevolutionary dynamics. He combines experimental and computational approaches to investigate the origin of new genes, the evolution of development, and the emergence of biological complexity.

Momir Futo is currently an Assistant Professor at the School of Medicine, Catholic University of Croatia. He obtained his undergraduate degree in Biology from the Faculty of Science, University of Zagreb in 2009. In 2016, he earned his doctorate from the Westfälische Wilhelms-Universität Münster, Germany, as a DAAD scholar. His doctoral research was conducted at the Institute for Evolution and Biodiversity under the mentorship of Prof. Dr. Joachim Kurtz. Following his doctoral research, Dr. Futo held a postdoctoral position in the Domazet-Lošo Group at the Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute in Zagreb, Croatia. His research interests are diverse and interdisciplinary, from data analysis with a focus on the genetic basis of biofilm formation across bacterial species to the causative agents of Lyme disease (*Borrelia* spp.). In addition, he is engaged in research on biocontrol strategies and the breeding and production-scale rearing of beneficial insects, including the black soldier fly. Overall, Dr. Futo's work bridges microbiology, evolutionary biology, entomology, and ecology.

ACKNOWLEDGEMENTS

This doctoral thesis was made possible through the help, support and guidance of many individuals, and I am sincerely grateful to each of them.

First, I would like to express my gratitude to my supervisor, Professor Tomislav Domazet-Lošo, for the opportunity to pursue this research topic and for his continuous guidance and support throughout the course of this work. His insights and expertise shaped the project at every stage.

I am equally grateful to my co-supervisor, Assistant Professor Momir Futo, for his encouragement, thoughtful advice, and invaluable assistance. His contributions have been crucial to the progress and completion of this project.

My thanks go to Dr. Sara Koska, who introduced me to bioinformatics and helped me build both confidence and perspective in the field. Her support has been instrumental in shaping my understanding and approach to the subject.

I am especially grateful to fellow doctoral student Niko Kasalo for his support, careful feedback, and help with the finishing touches of this project.

I am thankful to have been part of the Laboratory for Evolutionary Genetics at the Ruđer Bošković Institute. I express my gratitude to all current and former members of the group for their shared knowledge, experience, and generous advice. Each of their contributions has been vital to the making and completion of my doctoral thesis.

Special thanks to our collaborators, Professor Ivan Mijaković and doctoral student Ema Svetličić at the Danish Technical University, for their significant help with the facilitation and analysis of proteome data. Their expertise greatly enriched this research.

Finally, I thank my family and friends for their love, care, and encouragement throughout my doctoral studies. Their support carried me through the challenging moments and made the good ones possible.

EVOLUCIJSKA OBILJEŽJA BIOFILMOVA PATOGENOG BAKTERIJSKOG SOJA *Escherichia coli* UTI89

ANJA TUŠAR

Institut Ruđer Bošković

Formiranje biofilma u bakterijskom soju *Escherichia coli* UTI89 ima ključnu ulogu u njegovoj patogenosti i pojavi recidivnih infekcija mokraćnog sustava. Evolucijska i razvojna dinamika skrivena u osnovi formiranja biofilma *E. coli* do sada nije istražena. Kako bismo istražili korelaciju između ontogeneze i filogeneze u in vitro biofilmovima *E. coli*, proveli smo filotranskriptomске i filoproteomske analize različitih razvojnih stadija. Rezultati pokazuju kako razvoj biofilma *E. coli* oslikava ekspresijske obrasce i evolucijska obilježja slična onima prethodno opisanima kod gram-pozitivne vrste *Bacillus subtilis*. Kako razvojni proces biofilma kod gram-negativne *E. coli* također sadrži duboke makroevolucijske otiske, zaključujemo da su evolucijska obilježja sveprisutna značajka razvoja bakterijskih biofilmova. Nadalje, prostorna analiza koncentričnih regija biofilma *E. coli* u tri razvojna stadija otkrila je da su ekspresijski obrasci te evolucijske značajke oblikovani i vremenskom i prostornom razvojnom komponentom. Naši rezultati ističu duboko ukorijenjena evolucijska obilježja u organizaciji biofilma te pružaju nove uvide u razvojnu složenost uropatogenih biofilmova.

(115 stranica, 27 slika, 5 tablica, 149 literaturnih navoda, jezik izvornika: engleski)

Ključne riječi: *biofilm, razvoj, filotranskriptomika, filoproteomika, uropatogena Escherichia coli, višestaničnost*

- Mentori: prof. dr. sc. Tomislav Domazet-Lošo, Institut Ruđer Bošković, Hrvatska
doc. dr. sc. Momir Futo, Medicinski fakultet, Hrvatsko katoličko sveučilište, Zagreb, Hrvatska
- Ocjenjivači: prof. dr. sc. Damjan Franjević, Prirodoslovni-matematički fakultet, Sveučilište u Zagrebu, Hrvatska
prof. dr. sc. Jasna Hrenović, Prirodoslovni-matematički fakultet, Sveučilište u Zagrebu, Hrvatska
prof. dr. sc. Ivan Mijaković, Chalmers University of Technology, Švedska
- Zamjena: prof. dr. sc. Boris Maček, University of Tübingen, Njemačka

EVOLUTIONARY PATTERNS IN THE BIOFILMS OF PATHOGENIC *Escherichia coli* UTI89 STRAIN

ANJA TUŠAR

Ruđer Bošković Institute

Biofilm formation by *Escherichia coli* UTI89 plays a critical role in its pathogenicity and the recurrence of urinary tract infections. However, the evolutionary and developmental dynamics underlying biofilm growth in *E. coli* UTI89 remain poorly characterized. To investigate ontogeny-phylogeny correlations in *in vitro* *E. coli* biofilms, we applied phylotranscriptomic and phyloproteomic analyses across developmental stages. Our findings reveal that biofilm development in *E. coli* exhibits expression patterns and evolutionary imprints, analogous to those previously observed in gram-positive *Bacillus subtilis*. These results suggest that biofilm development in gram-negative *E. coli* is not macroevolutionary naïve and indicates that evolutionary patterns are a ubiquitous feature of biofilm development. Furthermore, spatial analysis of biofilm regions at three developmental stages revealed that expression profiles, and evolutionary patterns are shaped by temporal progression and spatial positioning. These findings highlight the evolutionary pattern of biofilm organization and provide new insights into the developmental complexity of uropathogenic biofilms.

(115 pages, 27 figures, 5 tables, 149 references, original in: English)

Keywords: *biofilm, development, multicellularity, phylotranscriptomics, phyloproteomics, uropathogenic Escherichia coli*

Supervisors: dr. rer. nat. Tomislav Domazet-Lošo, Professor, Ruđer Bošković Institute, Croatia

dr. sc. Momir Futo, Assistant Professor, School of Medicine, Catholic University of Croatia, Croatia

Reviewers: dr. sc. Damjan Franjević, Professor, Faculty of Science, University of Zagreb, Croatia

dr. sc. Jasna Hrenović, Professor, Faculty of Science, University of Zagreb, Croatia

dr. sc. Ivan Mijaković, Professor, Chalmers University of Technology, Sweden

Replacement: dr. rer. nat. Boris Maček, Professor, University of Tübingen, Germany

TABLE OF CONTENTS

1	INTRODUCTION.....	1
2	LITERATURE REVIEW.....	4
2.1	Biofilms.....	4
2.1.1	Bacterial biofilms	4
2.1.2	Medical relevance of uropathogenic <i>E. coli</i> biofilms	5
2.1.3	Biofilm as multicellular organism.....	6
2.1.4	Biofilm formation as developmental process.....	8
2.2	Evolutionary patterns	10
2.2.1	Phylogeny-ontogeny correlation in eukaryotes.....	10
2.2.2	Phylogeny-ontogeny correlation in bacteria	11
3	MATERIAL AND METHODOLOGY	13
3.1	Cultivation of <i>Escherichia coli</i> biofilms	13
3.2	Transcriptome data	13
3.2.1	RNA extraction	13
3.2.2	RNA sequencing	15
3.2.3	Transcriptome data analyses	15
3.3	Proteome data.....	17
3.3.1	Protein extraction	17
3.3.2	Mass spectrometry.....	18
3.3.3	Proteome data analyses	19
3.4	Phylostratigraphic analysis.....	20
3.5	Enrichment analysis and functional annotation	26
3.6	Data and code availability	27
4	RESULTS.....	28
4.1	Biofilm development is a stage-organized process	28

4.2	Phylogeny-ontogeny correlation reveals recapitulation pattern.....	35
4.3	Stage resolved functions distribution in <i>E. coli</i> biofilm development.	45
4.4	Spatially resolved transcription and translation in <i>E. coli</i> biofilm.....	62
4.5	Evolutionary imprints in the <i>E. coli</i> biofilm regions.....	74
5	DISCUSSION	76
5.1	Stage-organized transcription and translation during <i>E. coli</i> biofilm development.	77
5.2	Phylogeny-ontogeny correlation in <i>E. coli</i> reveals recapitulation pattern	79
5.3	Evolutionary imprints in the <i>E. coli</i> biofilm regions.....	80
5.4	Future directions.....	82
6	CONCLUSION	84
7	REFERENCES.....	86
8	APPENDICES.....	104
8.1	Figures.....	104
8.2	Tables	108
9	CURRICULUM VITAE	114

1 INTRODUCTION

Rapid advances in transcriptomic and proteomic technologies, together with improved bioinformatic workflows, have significantly expanded the scope of questions that can be addressed in biofilm research. Despite these developments, a synchronized time-resolved transcriptome and proteome analysis of uropathogenic *E. coli* biofilm development, has not yet been performed, leaving a gap in understanding of molecular and evolutionary biology of uropathogenic *Escherichia coli* (UPEC) biofilm.

Urinary tract infections (UTIs) are among the most prevalent bacterial infections, and uropathogenic *Escherichia coli* account for around 85% of cases (Hadjifrangiskou et al., 2012). The UPEC strain *E. coli* UTI89 is a widely used model for UTI pathogenesis and is capable of transitioning between extracellular and intracellular lifestyles, including biofilm-like intracellular communities in bladder epithelial cells (Zagaglia et al., 2022). Biofilms are communities of genetically identical bacterial cells that display phenotypic heterogeneity, division of labor, and coordinated production of extracellular matrix, thereby enhancing stress tolerance and protection from host immune defenses (Vlamakis et al., 2013). In UPEC, biofilm formation is closely associated with persistence and recurrence of urinary tract infection (Reigstad et al., 2007; Rijavec et al., 2008; Tajbakhsh et al., 2016; Zagaglia et al., 2022). A deeper understanding of molecular logic of UPEC biofilm development is therefore both clinically relevant and conceptually important for microbial biology.

Beyond their medical significance, biofilms provide a suitable system to research general principles of multicellular organization. While traditionally viewed as simple, unicellular organisms, when forming biofilms bacteria have increasingly been conceptualized as multicellular systems over the last several decades. This paradigm shift is driven by evidence of bacterial communication, which enables coordinated decision-making, division of labor, collective motility, and metabolic synchronization (Shapiro, 1998). These cooperative behaviors afford distinct advantages to respective populations, from metabolic efficiency and resource acquisition to collective defense. Although in nature biofilms are predominantly multi-species communities (Røder et al., 2016), single-species models have proved to be indispensable in dissecting the core mechanisms of biofilm biology (Vlamakis et al., 2013). Multicellular behaviors in biofilms are widespread and constitute an important

part of the bacterial life cycle. Biofilms represent the most common form of bacterial existence in nature (Flemming & Wuertz, 2019), and thus exhibit cellular heterogeneity similar to that of multicellular eukaryotes (Claessen et al., 2014; Penesyan et al., 2021; Jo et al., 2022; Sauer et al., 2022).

Although multicellularity arose independently across diverse eukaryotic lineages (Rensing, 2016), it was found that the correlation between phylogeny and ontogeny remains remarkably uniform across the eukaryotic tree of life (Domazet-Lošo & Tautz, 2010b; Quint et al., 2012; Cheng et al., 2015; Lotharukpong et al., 2024; Koska et al., 2025). This shared evolutionary hourglass pattern is reflected at the molecular level, where evolutionarily younger and more divergent genes are expressed during early and late development, whereas evolutionarily older and more conserved genes are predominantly expressed during mid-development (Domazet-Lošo & Tautz, 2010b; Quint et al., 2012; Cheng et al., 2015; Lotharukpong et al., 2024; Koska et al., 2025). In Futo et al. (2021) study, *Bacillus subtilis* biofilm was approached as a multicellular organism by applying phylotranscriptomics and phyloproteomics as it would be used for studying ontogeny of an embryo-forming eukaryote (Domazet-Lošo & Tautz, 2010a). This study showed that phylostratigraphy, linking evolutionary and developmental data, can be used in simple multicellular formations, such as biofilms (Futo et al., 2021). Phylotranscriptomic and phyloproteomic analyses demonstrated that biofilm development proceeds through distinct ontogenetic stages and shows a recapitulation evolutionary pattern: evolutionarily younger and more diverged genes contribute increasingly at later stages of biofilm development (Futo et al., 2021). These results proposed that biofilm formation is a *bona fide* developmental process whose organization can be examined using tools originally developed for eukaryotic evolutionary developmental biology (Domazet-Lošo & Tautz, 2010a; Futo et al., 2021). Given that biofilm development is thought to involve broadly conserved strategies across bacterial species (O'Toole et al., 2000; Stoodley et al., 2002; Lappin-Scott et al., 2014; Sauer et al., 2022), it is important to determine whether ontogeny–phylogeny correlations also occur in gram-negative pathogens such as UPEC, rather than being limited to non-pathogenic gram-positive bacteria such as *B. subtilis*.

The evolutionary patterns identified in *B. subtilis* were derived from bulk biofilm expression profiles (Futo et al., 2021), even though spatial heterogeneity is a key determinant of biofilm physiology (Serra et al., 2013a, 2013b; Dragoš et al., 2018; Silva et al., 2016). Biofilm spatial heterogeneity is known to play a critical role in community

organization and function (Dragoš et al., 2018; Serra et al., 2013a, 2013b), influencing key processes such as antibiotic resistance and virulence (Silva et al., 2016). Although several studies have characterized region-specific transcription in *E. coli* biofilms (Beebout et al., 2019; Díaz-Pascual et al., 2021; Wang et al., 2023), these analyses typically focus on a single developmental time point and do not connect spatial expression patterns to macroevolutionary imprints.

In this doctoral thesis, we aim to determine whether uropathogenic *E. coli* UTI89 biofilm development is stage-organized process and whether this process exhibits evolutionary and functional patterns across time and biofilm regions. To accomplish this aim, we defined these objectives:

O1: Determine whether *E. coli* UTI89 biofilm ontogeny is stage-organized at the transcriptome and proteome levels by profiling biofilms across one month developmental time period.

O2: Analyze phylogeny–ontogeny correlations by integrating expression profiles with phylostratigraphic gene age and divergence based indices throughout biofilm development of *E. coli* UTI89.

O3: Evaluate how biological functions are deployed across biofilm ontogeny by performing functional enrichment analyses and defining stage-specific functional programs, including the contribution of poorly annotated genes.

O4: Assess whether developmental, functional, and evolutionary organization is spatially stratified within biofilms by comparing transcriptome and proteome profiles between inner, middle, and outer biofilm regions at key developmental stages.

To address these gaps, we generated a spatio-temporal transcriptomic and proteomic data from *E. coli* UTI89 biofilm development on agar plates, spanning a one month developmental trajectory sampled from eleven timepoints.

2 LITERATURE REVIEW

2.1 Biofilms

2.1.1 Bacterial biofilms

Bacterial biofilms are complex multicellular systems comprising surface-adhered microbial communities enveloped in a self-produced extracellular matrix. Characterized by their ubiquitous distribution, bacterial biofilms colonize diverse biotic and abiotic niches that offer sufficient moisture and minimal nutrient thresholds (Singh et al., 2006). This multicellular mode of bacterial growth provides robust protection against environmental fluctuations, including extremes in temperature, pH shifts, UV exposure, and nutrient scarcity (Hall-Stoodley et al., 2004). Bacterial biofilms are considered problematic in many medical and industrial settings, where biofilm-mediated resistance to antimicrobials remains a critical challenge (Hall-Stoodley et al., 2004). Biofilm resistance to antibiotics and antimicrobial agents is driven by the physical diffusion barrier imposed by the extracellular polymeric matrix, the presence of localized microenvironments harboring dormant, stationary-phase cells, and the emergence of highly resistant phenotypes (Hall-Stoodley et al., 2004). For example, biofilms are associated with persistent infections of the oral cavity, skin, and urinary tract (López et al., 2010). Despite their pathological detriments, the unique properties of biofilms have been successfully harnessed for industrial and agricultural utility. For instance, their metabolic versatility is exploited in microbial fuel cells to convert organic waste into green energy (Logan, 2009), as well as in eco-efficient bioremediation pipelines for effluent treatment (Singh et al., 2006). Additionally, in sustainable agriculture, rhizosphere-dwelling biofilms act as effective biocontrol barriers, actively shielding plant root systems from devastating fungal, bacterial, and nematode pathogens (Vlamakis et al., 2013). Furthermore, in industrial biotechnology, productive biofilms are increasingly recognized as robust, immobilized biocatalysts capable of driving continuous bioprocesses for the synthesis of platform chemicals, organic acids, and diverse fermented products (Morikawa et al., 2006). Their structural stability and metabolic resilience also render them highly effective in environmental biotechnology, particularly in marine bioremediation systems. Here, biofilms play a critical role in the metabolic degradation of complex hydrocarbons and the mitigation of toxic oil spills in aquatic ecosystems (Head et al., 2006; Gertler et al., 2009).

In laboratory settings, bacterial biofilms can be cultivated under diverse physical conditions, yielding structurally distinct formations: macrocolony biofilms (air-agar interface), pellicles - floating biofilms (air-liquid interface), and submerged, surface-adhered biofilms (liquid-solid interface) (Vlamakis et al., 2013). Although biofilms in nature are predominantly multi-species communities (Røder et al., 2016), clinical relevance historically prompted the use of monospecies models. Consequently, gram-negative pathogens such as *Escherichia coli* and *Pseudomonas aeruginosa* became the primary models for biofilm research (Vlamakis et al., 2013). Bridging clinical pathology with developmental molecular biology, this study leverages the uropathogenic *Escherichia coli* (UPEC). This clinical isolate UTI89 was originally harvested from an acute cystitis patient (Chen et al., 2006). UTI89 strain serves as a model for investigating host-pathogen dynamics, intracellular bacterial community (IBC) development, and complex biofilm developmental programs both *in vitro* and *in vivo* (Reigstad et al., 2007; Hadjifrangiskou et al., 2012; Beebout et al., 2019; Zagaglia et al., 2022).

2.1.2 Medical relevance of uropathogenic *E. coli* biofilms

Urinary tract infections (UTIs) are among the most prevalent bacterial infections worldwide, with uropathogenic *Escherichia coli* (UPEC) acting as the primary etiological agent responsible for up to 85% of community-acquired and 25% of nosocomial UTIs (Foxman, 2010). It is thought that pathogenic UPEC strains have evolved from commensal *E. coli* strains through acquiring virulence-associated genes via horizontal gene transfer of transposons, plasmids and bacteriophages (Leimbach et al., 2013). They express a variety of virulence factors including adhesins, invasins, capsules, iron-acquisition systems, and toxins, which are interconnected in the development of acute or persistent urinary tract infections (Zagaglia et al., 2022). Biofilm production by *E. coli* is considered to be an important virulence factor that plays a key role in the persistence of UPEC and the recurrence of UTI in the human body (Rijavec et al., 2008). Typically, UPEC strains reside as commensals within the human gastrointestinal tract, from which they can opportunistically colonize the periurethral area and ascend into the urinary tract and cause disease (Kaper et al., 2004; Foxman, 2010). During urinary tract infection, *E. coli* UTI89 can transition between extracellular and intracellular lifestyles, including biofilm-like intracellular bacterial communities (IBCs) within bladder epithelial cells (Zagaglia et al., 2022). Upon entering the bladder, UPEC utilizes type 1 fimbriae to bind to superficial

umbrella cells of the bladder epithelium (Foxman & Brown, 2003). This binding triggers host-cell invasion, allowing the bacteria to escape extracellular immune surveillance and antibiotic exposure by replicating within the bladder cells' cytoplasm. Inside host cells, UPEC can assemble into highly organized, biofilm-like structures known as intracellular bacterial communities (IBCs) (Zagaglia et al., 2022). The clinical relevance of these intracellular structures is supported by the detection of intact IBCs in exfoliated urothelial cells shed in the urine of patients suffering from acute cystitis (Rosen et al., 2007). In both extracellular and intracellular environments, the capacity of UPEC to assemble biofilms is a critical determinant of pathogenesis. The structured community lifestyle of biofilms enables them to persist under intense environmental and host-derived stress (Flemming & Wingender, 2010).

2.1.3 Biofilm as multicellular organism

Biofilms, previously considered as collections of individual bacterial cells, are now being studied as multicellular aggregates (Futo et al., 2021, Claessen et al., 2014, Hung et al., 2013). When observing the morphology of a macrocolony biofilm on agar plate, it immediately becomes apparent that biofilms are not simply random clusters of cells and display behaviours similar to multicellular organisms (Futo et al., 2022; Serra et al., 2013b). Although cells within a single-species biofilm are genetically identical as they arise from a single cell, they express different genetic programs that produce subpopulations of functionally and phenotypically distinct cell types (López et al., 2010; Vlamakis et al., 2013). *In vitro* biofilm formation on agar plate reveals morphological rearrangements that occur at specific stages of biofilm development (Futo et al., 2022; Serra et al., 2013b). The development of macrocolony biofilm is a complex process involving cellular differentiation, that forms elaborate three-dimensional structures, such as radial ridges, concentric rings, labyrinth networks and branching wrinkles (Serra et al., 2013a). Different morphological forms of macrocolony biofilms arise from the interplay of transcriptional and environmental signals, leading to the specialization of subpopulations within the biofilm across spatial and temporal scales (Dragoš et al., 2018; Stoodley et al., 2002).

Spatio-temporal gene expression regulation is one of the important conditions to be met for the development of multicellular organism, and *E. coli* fits this developmental criterion by generating distinct, spatially organized subpopulations of cells (López et al., 2010; Serra et al., 2013a, 2013b). Linked to this spatio-temporal organization are gene regulatory

networks, which serve as driver for transitions through biofilm developmental stages (Monds & O'Toole, 2009). In *E. coli*, this network is centred around the regulator CsgD, which acts as a molecular switch. Under the control of the stationary-phase sigma factor sigma S (*rpoS*) and the second messenger cyclic-di-GMP, CsgD regulates motility and sessility by repressing flagellar genes while activating the synthesis of curli fimbriae and cellulose (Pesavento et al., 2008). Because this system integrates numerous environmental and physiological inputs, such as osmolarity, temperature, and nutrient status through regulators like OmpR and CpxR, it enables *E. coli* to execute coordinated and region-specific decisions (Jubelin et al., 2005; Ogasawara et al., 2010). In non-pathogenic *E. coli* biofilms, arginine metabolism is spatially organized across the biofilm regions. Specifically, biosynthesis genes (*arg* operon) are highly enriched at the nutrient-accessible biofilm periphery, whereas catabolic and utilization genes (*ast* operon) are predominantly expressed within the biofilm interior (Wang et al., 2023). This expression regulation ensures that cells at the biofilm's periphery and those deep within its core express distinct, specialized transcriptomes and proteomes, reinforcing the developmental characteristic of multicellular organisms.

In *E. coli* biofilms, the production of extracellular matrix components (such as curli fibers), outer membrane vesicles (OMVs), and autoinducers represent a form of intercellular cooperation characteristic for multicellular behavior. These components function as public goods, while their synthesis is metabolically costly for an individual cell, they provide structural integrity and chemical protection to the entire community (Nadell et al., 2016). Phenotypic heterogeneity dictates that only a specific subpopulation of *E. coli* cells actively expresses regulator *csgD* to produce curli fibers, while neighboring non-producing cells benefit from the shared matrix (Serra et al., 2013a). This division of labor is heavily coordinated by autoinducer-2 (AI-2) mediated quorum sensing, where self-produced signaling molecules orchestrate collective behavior (Laganenka et al., 2016). When specialized sub-populations secrete signals or metabolic byproducts that direct the behavior of distinct, non-producing neighboring cells without stimulating the emitter, it establishes a signaling network similar to that observed in multicellular eukaryotes (López et al., 2009). Furthermore, coordinated survival in *E. coli* biofilms extends to metabolic signaling. Under localized nutrient limitation, *E. coli* communities balance the metabolic demands of the peripheral and interior cells to prevent the starvation and death of the cells from the biofilm interior, maintaining overall biofilm viability (Cole et al., 2015). Nutrient gradients within

the biofilm lead to cell differentiation that reflects two fundamentally different physiological strategies of bacteria: growth and survival (Stewart & Franklin, 2008). Bacteria at the outer edges and in the lower layer of *E. coli* macrocolonies, close to the nutrient-providing agar surface, remain actively growing, dividing, and flagellated, whereas cells in the upper and central regions enter a stationary-phase-like state and become embedded in an extracellular matrix composed mainly of curli fibers and cellulose (Serra et al., 2013a, 2013b). This metabolic coordination is dynamically integrated with membrane potential dynamics; shifts in metabolic states correlate with changes in potassium ion-channel activity and electrical signaling across the biofilm, which synchronizes metabolic activity and stress tolerance between distant cells (Humphries et al., 2017; Martinez-Coronado et al., 2020). Through these chemical cues, *E. coli* communities dynamically regulate their architecture and stress responses, exhibiting a primitive form of collective decision-making analogous to multicellular organisms (Shapiro, 1998). Microscale chemical gradients, adaptation to local environmental conditions (Dragoš et al., 2018), stochastic gene expression and genotypic variation through mutation and selection (Besharova et al., 2016; Stewart & Franklin, 2008) are all contributing to genetic and physiological heterogeneity within biofilm.

2.1.4 Biofilm formation as developmental process

The biofilm field remains divided over a fundamental question: is biofilm maturation a reproducible, highly regulated developmental process, or is it an emergent phenomenon shaped entirely by local microenvironmental cues? Bonner (2001) views polarity and heritable pattern formation, including cell differentiation with division of labor, as fundamental properties of multicellular development.

Based on this concept, bacterial biofilms fit the criteria of a multicellular organism (Futo et al., 2021). *Bacillus subtilis* biofilm growth has been determined to be a highly regulated multicellular developmental process organized into discrete ontogenetic stages (Futo et al., 2021). Gene expression profiles were found to be temporally organized during *B. subtilis* biofilm development (Futo et al., 2021), mimicking the progression of early developmental stages in animals (Tomancak et al., 2007). However, while some aspects of biofilm development point to strictly encoded genetic programs, the extent to which these systems rely instead on flexible, local physiological responses to environmental cues remains to be determined (Monds and O'Toole, 2009). Classical developmental concepts from eukaryotic

systems, such as irreversible cellular commitment, checkpoint-mediated progression, and strict hierarchical gene regulation, are foundational to embryogenesis and stem cell differentiation, but they do not map onto the inherent plasticity of biofilms, at least not in a strict sense. Biofilm development is more flexible and environmentally responsive than, for example, stem cell differentiation in embryogenesis. Instead, the stringency of a bacterial developmental program appears to scale directly with environmental risk. In high-stakes survival scenarios, such as *Bacillus subtilis* sporulation (Higgins and Dworkin, 2012) and *Myxococcus xanthus* fruiting-body formation (Zusman et al., 2007), bacteria rely on tightly regulated developmental programs. By contrast, biofilm development under less restrictive conditions can involve more flexible and environmentally responsive differentiation, allowing cells to adjust to local microenvironmental gradients without compromising community-level function (Jo et al., 2022).

Recent research demonstrates that rather than aggregating randomly, *E. coli* cells assemble into rosettes to initiate multicellular development process that leads to the biofilm formation (Puri et al., 2022; Puri et al., 2024). Rosettes are self-organizing, circular multicellular communities that were first discovered to initiate developmental processes, like organogenesis and embryogenesis, in complex organisms (Puri et al., 2024). This discovery suggests that *E. coli* possesses a multicellular life cycle rooted in rosette-initiated development, an adaptation previously thought to be exclusive to multicellular eukaryotes (Puri et al., 2022; Puri et al., 2024). Rosette formation is far from a passive aggregation and starts with active mechanism of repositioning of adhered sister cells and was thought to distinguish multicellular organisms from unicellular ones (Puri et al., 2024). Upon clonal division at the 2- and 4-cell stages, cells actively reposition and reorient themselves relative to one another. Rather than separating, they pivot and sweep through an angular random walk until they align radially around a shared, matrix-filled central focus, generating rosette with a characteristic quatrefoil configuration (Puri et al., 2024). Genetic analyses demonstrate that this complex multicellular behavior is strictly regulated by specific genes. The *flu* gene, which encodes the autotransporter adhesin Antigen 43 (Ag43), acts as the intercellular glue required for self-recognition, ensuring sister cells remain physically tethered to maintain the clonality of the emerging community (Puri et al., 2022; Puri et al., 2024). This rosette folding process also requires active flagellar propulsion, where *de novo* synthesis of the flagellar filament via *fliC* and functional flagellar motor rotation via *motA* provide the mechanical torque needed for cells to physically push and pivot into the radial

rosette shape (Puri et al., 2024). Once this active folding process is complete, structural matrix genes like *fimA* for Type 1 fimbriae and *csgA* for curli fibers become essential to anchor, stabilize, and lock the final three-dimensional geometry of the rosette community before it transitions into elongated multicellular chains (Puri et al., 2022). Furthermore, the stress-response regulator *rpoS* acts as a downstream developmental gatekeeper, as proper rosette formation is required to trigger subsequent *rpoS* expression, which coordinates the maturation of these rosette-derived chains into tissue-like biofilms (Puri et al., 2024). This highly ordered, genetically programmed formation appears to be a widespread multicellular phenomenon across the *E. coli* lineage, extending beyond commensal laboratory strains to include highly virulent uropathogenic isolates (Puri et al., 2022; Justice et al., 2004). These findings indicate that the rosette phase serves as a crucial, reproducible checkpoint in *E. coli* development, providing the necessary architectural foundation for subsequent biofilm formation.

2.2 Evolutionary patterns

2.2.1 Phylogeny-ontogeny correlation in eukaryotes

The relationship between phylogeny and ontogeny has been a central question in evolutionary developmental biology since the nineteenth century. Early interpretations proposed a close parallel between organismal development and evolutionary history, most prominently Haeckel's recapitulation theory, summarized as "ontogeny recapitulates phylogeny" (Haeckel, 1868; Abzhanov, 2013). In contrast, von Baer's laws emphasized that embryos develop from general toward more specialized forms and do not pass through the adult stages of other species (Abzhanov, 2013). These ideas gave rise to two major models of developmental conservation: the early conservation or "funnel" model, which predicts strongest conservation at early developmental stages, and the developmental "hourglass" model, which predicts maximal conservation during mid-development, often corresponding to the phylotypic period when the basic body plan is established (Slack et al., 1993; Duboule, 1994; Kalinka & Tomancak, 2012; Abzhanov, 2013). With the development of transcriptomics and phylostratigraphy, these classical morphological concepts were extended to the molecular level by linking gene expression patterns with gene evolutionary age and sequence divergence (Domazet-Lošo et al., 2007; Domazet-Lošo & Tautz, 2010a; Quint et al., 2012). Although some comparative transcriptome studies supported early conservation or reverse-hourglass patterns, phylostratigraphic

analyses across animals, plants, fungi, and algae repeatedly revealed a molecular hourglass pattern, in which older and more conserved genes are preferentially expressed during mid-development, whereas younger and more divergent genes are more strongly expressed during early and late development (Domazet-Lošo and Tautz, 2010a; Quint et al., 2012; Cheng et al., 2015; Lotharukpong et al., 2024; Koska et al., 2025). This repeated evolutionary pattern is particularly notable because animals, plants, and fungi evolved multicellularity independently, suggesting that similar system-level principles may shape developmental programs across distant eukaryotic lineages (Rensing, 2016). Extending this framework beyond eukaryotes, Futo et al. (2021) applied phylotranscriptomic and phyloproteomic approaches to bacterial biofilm development, providing the basis for examining whether comparable phylogeny–ontogeny correlations also occur in one of the simplest forms of multicellular development, as discussed in the next chapter.

2.2.2 Phylogeny-ontogeny correlation in bacteria

Macrocolony biofilms exhibit a remarkably high level of structural organization during development, making them a suitable model for examining whether phylogeny–ontogeny correlations also occur outside eukaryotic development (Futo et al. 2021). Agar-grown macrocolony biofilms are especially useful in this context because they develop reproducible architecture over time and show pronounced spatial and temporal organization (van Gestel et al., 2015; Ackermann et al., 2015; Dragoš et al., 2018; Serra et al., 2013a). Building on phylotranscriptomic and phyloproteomic approaches originally developed for eukaryotic developmental systems, Futo et al. (2021) demonstrated that *Bacillus subtilis* biofilm ontogeny is organized into discrete developmental stages and exhibits a recapitulation-like evolutionary pattern. In contrast to the molecular hourglass pattern commonly observed in eukaryotes, *B. subtilis* biofilms showed progressive expression of evolutionarily younger and more divergent genes during later stages of biofilm growth, suggesting that bacterial biofilm development also carries macroevolutionary patterns (Futo et al., 2021). This finding indicates that phylogeny–ontogeny correlations may reflect broader system-level principles of multicellular organization rather than being restricted to animal, plant, or fungal development. The usefulness of phylostratigraphic approaches in bacteria is further supported by the identification of previously uncharacterized sporulation genes in *B. subtilis*, whose function was inferred from their shared evolutionary origin with known sporulation genes rather than from sequence homology alone (Shi et al., 2020).

However, the macroevolutionary patterns described in *B. subtilis* were recovered from temporal samples of bulk biofilms, leaving open question of how such evolutionary patterns are organized within spatially heterogeneous biofilm communities. Because spatial organization plays an important role in biofilm structure, metabolism, antibiotic resistance, and virulence (Dragoš et al., 2018; Silva et al., 2016; Andersson et al., 2019; El-Halfawy et al., 2015), examining expression profiles across distinct biofilm regions may reveal an additional layer of developmental and evolutionary organization. Although spatial transcriptome studies have already shown regional expression differences within *E. coli* biofilms, these analyses have mostly focused on single developmental time points rather than spatio-temporal changes during biofilm formation (Wang et al., 2023; Díaz-Pascual et al., 2021; Beebout et al., 2019). Therefore, combining temporal and spatial transcriptome and proteome analyses provides an opportunity to test whether evolutionary imprints are present not only across biofilm ontogeny, but also within distinct regions of developing bacterial biofilms.

3 MATERIAL AND METHODOLOGY

3.1 Cultivation of *Escherichia coli* biofilms

All experiments in this study were conducted using the uropathogenic *Escherichia coli* (UPEC) strain UTI89, a widely recognized *in vitro* model for investigating biofilm development and urinary tract pathogenesis. (Hadjifrangiskou et al., 2012; Beebout et al., 2019; Garcia et al., 2021). This clinical isolate, originally derived from a patient with an acute bladder infection (Chen et al., 2006), was obtained from the Technical University of Denmark, Biosustain.

The starting liquid culture (LC) was prepared by cultivating *E. coli* UTI89 in liquid LB medium (1% Bacto tryptone, 0.5% Bacto yeast extract, 0.5% NaCl) at 37 °C for 16 h. Biofilms were grown using a modified version of the protocol described in Serra and Henнге (2017). The original NaCl-free LB agar medium was replaced with YESCA agar medium (1% casamino acids, 0.1% yeast extract, 2.0% agar), which supported more robust biofilm formation. For biofilm cultivation, YESCA agar plates were inoculated by pipetting 10 µL of the starting liquid culture onto the agar surface. The plates were then closed, sealed around the edges with Parafilm, and incubated at 27 °C. Bulk biofilm samples were harvested at multiple developmental time points: 6 and 12 hours, and at 1, 1.5, 2, 3, 5, 7, 14, 21, and 30 days post-inoculation (Figure 1a). Biofilm samples were labelled as 6H, 12H, 1D, 1.5D, 2D, 3D, 5D, 7D, 14D, 21D, and 1M. All samples for RNA and protein extraction were collected in three biological replicates per time point.

To assess spatial dynamics during biofilm development, additional samples were collected at three key developmental stages: 1D (early stage), 3D (mid stage), and 14D (late stage). We dissected biofilms at 1D and 3D stages into two parts (inner, outer) and at 14D stage into three parts (inner, middle, outer) using a sterile scalpel (Figure 17a). For each spatial region, samples from multiple biofilms were pooled to ensure adequate biomass for RNA and protein extraction. Three biological replicates were collected for each spatial region to support reproducibility and statistical analysis.

3.2 Transcriptome data

3.2.1 RNA extraction

E. coli UTI89 liquid culture (LC) sample grown in LB medium was centrifuged at 5,000 ×

g for 10 minutes prior to RNA extraction. For spatial sampling of biofilms, the inner, middle, and outer regions were carefully dissected using a sterile scalpel and the upper parts of pipette tips. Both bulk biofilms and their sectioned spatial regions were collected from agar plates using plastic spreaders and placed into 2 mL tubes with 1.5 mL of stabilization buffer, RNeasy Protect Bacteria Reagent (Qiagen) diluted with PBS at a 2:1 volume ratio.

Total RNA was extracted using a modified protocol based on the RNeasy Protect Mini Kit (Qiagen). Biofilm samples were disrupted and homogenized using a Tissue Ruptor rotor-stator homogenizer (Qiagen) equipped with a 7 mm probe, operating at level 3 for 10 s. Homogenized samples were then vortexed for 10 s and incubated at room temperature (RT) for 5 min. Following incubation, the samples were divided equally into three sterile 2.0 mL tubes for lysis. Subsequent centrifugation at $5,000 \times g$ for 10 min at RT resulted in pellet formation, and the supernatants were discarded. Each pellet was resuspended in 220 μ L of a lysis solution consisting of 200 μ L of TE buffer (30 mM TrisHCl, 1 mM EDTA) containing lysozyme (15 mg/mL), and 20 μ L of Proteinase K (200 mg/mL). The lysis process was carried out by incubating the tubes in a shaker at 25 °C for 20 min and 550 rpm.

Following the lysis incubation, three volumes of RLT buffer (Qiagen) were added to each tube, and the mixtures were vortexed thoroughly. Samples were then centrifuged at $8,000 \times g$ for 10 min, after which the supernatants were transferred into fresh 2.0 mL tubes. To each sample, 0.5 volumes of chloroform were added, followed by a 10 min incubation at RT. Phase separation was achieved by centrifugation at $13,000 \times g$ for 15 min at 4 °C. The upper phase was pipetted into new tubes, and an equal volume of 80% ethanol was added. Samples were gently mixed by pipetting and were transferred into mini spin columns (Qiagen). After centrifugation at $10,000 \times g$ for 15 s, the flow-through was discarded. The step was repeated until the entire volume of each sample was processed through the column. Columns were then washed with 700 μ L of RW1 buffer (Qiagen) and centrifuged at $10,000 \times g$ for 15 s. After discarding the flow-through, the column was washed twice with RPE buffer (Qiagen), each wash followed by centrifugation at $10,000 \times g$ for 1 min. Collection tubes were replaced with new 1.5 mL tubes, and RNA was eluted by adding 50 μ L RNase-free water directly to the column membrane. After a 1 min incubation at RT, total RNA was recovered from the membrane by centrifuging for 1 min at $10,000 \times g$.

In the following step, DNase I stock solution (Qiagen) was mixed with RDD buffer (Qiagen) according to manufacturer's instructions. The DNase I mix was then added to

each RNA sample and incubated for 10 min at RT to digest residual genomic DNA. Afterward, LiCl was added to each sample in a final concentration 3.75 M. The mixtures were incubated over-night at -20 °C to facilitate RNA precipitation. Following an over-night incubation, samples were centrifuged for 15 min at 16,000 x g at 4 °C, with an in-between washing using 80% ethanol. Supernatants were discarded, and the pellets were dried before adding 30 µL RNase-free water. Samples were maintained on ice during quantification using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific) and during assessment of RNA integrity via agarose gel electrophoresis. Purified RNA samples were stored at -80 °C until sequencing.

3.2.2 RNA sequencing

For total RNA sequencing, ribosomal RNA (rRNA) was first depleted, and the remaining RNA was used to prepare libraries with the TruSeq Stranded kit combined with NEB rRNA depletion kit (Illumina). Sequencing was performed bi-directionally on the Illumina NovaSeq 6000 S4 platform at MacroGen Europe BV (Maastricht, the Netherlands). This approach yielded between 65 and 90 million reads per run for both entire and partial biofilm samples.

3.2.3 Transcriptome data analyses

We evaluated sequence quality using FastQC v0.11.9 (Andrews, 2010). Paired-end sequences (100 bp read length) were then mapped to the *E. coli* UTI89 reference genome (NCBI Assembly: ASM1326v1; GCA_000013265.1) using BBMap v38.39 (Bushnell, 2014). Across samples, an average of 86,846,330 read pairs per sample were generated, with an average mapping of 98.3% reads per biofilm sample. One biological replicate from the 21D time point was excluded due to failure during RNA sequencing library preparation. In addition, one biological replicate from the liquid culture (LC) time point was removed from downstream analysis due to a low number of counts. Consequently, analyses for both LC and 21D were performed using two biological replicates rather than three. For spatial biofilm samples, paired-end sequences (100 bp read length) were mapped to the same reference genome (NCBI Assembly: ASM1326v1; GCA_000013265.1) with an average of 76,080,157 read pairs per sample, achieving an average of 99.6% mapped reads per sample. We applied default settings for mapping, with the option enabled to trim read names following the first whitespace.

SAMtools package v1.9 (Li et al. 2009) was used to generate, sort, and index BAM files. We processed RNAseq data in R v4.2.2 (R Development Core Team 2008) using custom scripts published by Futo et al. (2021). Read quantification for each *E. coli* UTI89 open reading frame (ORF) was performed using the rsamtools package v2.14.0 (Morgan et al., 2017). Raw count data were obtained for 5,087 ORFs from entire biofilm data, and for 5,081 ORFs from spatial biofilm data, using the GenomicAlignments package v1.34.1 (Lawrence et al., 2013) (see Data and code availability).

To assess differential expression across time points or biofilm regions and biological replicates, Principal Component Analysis (PCA) was performed using the DESeq2 V1.38.3 R package (Love et al. 2014), with visualizations generated by the R ggplot2 package V3.5.0 (Wickham, 2016) (Figure 1bd, Figure 17bd). Raw read counts from the entire biofilm and spatial biofilm datasets were normalized by calculating the fraction of transcripts (τ) (Li et al., 2010). Following the methodology outlined in the Futo et al. (2021) paper, the multiplication step was omitted, and τ values were used directly for calculating evolutionary measures. Normalized expression values for biological replicates were resolved by calculating their median. Replicates with zero raw counts were excluded from further analysis. These normalized transcript expression values were used as input for phylotranscriptomic analyses (see Phylostratigraphic analysis).

For RNA expression profile correlation and visualization, genes with zero expression values across multiple samples were removed from both datasets, yielding 4,961 genes for the entire biofilm and 4,989 genes for the spatial biofilm dataset. When a gene showed zero expression at only a single stage, the missing value interpolated using the mean of the two adjacent stages. For zero-expression values occurring at the first or last stages of biofilm ontogeny, interpolation was performed by substituting the missing value with that of the directly adjacent stage (Futo et al., 2021). To normalize gene expression profiles across samples, each gene was scaled to its median expression level and subjected to $\log_2$ transformation, yielding per-gene standardized expression values. Visualization of expression dynamics was performed using the ggplot2 package v3.5.0 in R. Transcriptome similarity across biofilm developmental stages and spatial regions was assessed by calculating Pearson's correlation coefficients (R) for all-against-all comparisons, which were then represented as heatmaps (Figure 1ce, Figure 17ce). To delineate distinct developmental phases across biofilm stages (Figure 2a), we quantified stage-to-stage dissimilarity as $1 - R$, where R is the Pearson correlation coefficient. These dissimilarities

were then used for hierarchical clustering using the `hclust()` function in R.

We estimated overall differential gene expression across biofilm developmental stages and spatial regions within biofilm using the likelihood ratio test (LRT) implemented in the DESeq2 package v1.38.3 (Love et al., 2014).

Pairwise differential expression analysis between successive biofilm samples as well as between inner, middle, and outer biofilm region of 1D, 3D, and 14D biofilms was also performed using DESeq2 package v1.38.3 (Love et al. 2014). In all these analyses, a gene was considered differentially expressed if the adjusted p-value, corrected for multiple comparisons using the Benjamini–Hochberg procedure, was below 0.05.

3.3 Proteome data

3.3.1 Protein extraction

E. coli UTI89 liquid culture (LC) sample grown in liquid LB medium was centrifuged for 10 min at 5,000 x *g* prior to protein extraction. Entire and spatial biofilm samples grown on agar plates were harvested using sterile plastic spreaders, following the same procedure as for RNA extraction. The collected biomass was transferred into 2.0 mL tubes containing 1 mL of cell lysis buffer composed of 4% w/v SDS, 100 mM Tris pH 8.5, 5 mM beta-glycerophosphate, 5 mM NaF, 5 mM Na₃VO₄, 10mM EDTA, and 1/10 tablet of Mini EDTA-free Protease Inhibitor Cocktail from Sigma Aldrich. Samples were then resuspended in the cell lysis buffer and incubated at 95 °C for 10 min to facilitate initial cell disruption. To ensure complete lysis, samples were sonicated on ice using an Ultrasonic processor (Cole Parmer) set at 40% amplitude, applying 10 pulses per cycle and repeating three times.

Cellular debris was removed by centrifugation at 10,000 x *g* for 30 min at 4 °C. Soluble proteins present in the resulting supernatant were precipitated using a chloroform-methanol precipitation. Specifically, four volumes of 99.99% methanol, one volume of chloroform, and three volumes of milliQ water were added to each sample, followed by vortexing. Samples were then centrifuged at 5,000 x *g* for 10 min at 4 °C. The upper aqueous phase was discarded to preserve the protein-containing interphase. To further purify the protein fraction, an additional four volumes of methanol were added to each tube. After vortexing, samples were centrifuged again at 5,000 × *g* for 10 minutes at 4 °C. The resulting supernatant was discarded, and the protein pellet was air-dried for 10 min. The dried protein

pellet was dissolved in denaturation buffer (6 M urea, 2 M thiourea, 10 mM TrisHCl pH 8.0).

Protein concentrations were determined using the Bradford protein assay. For each sample, 50 µg of total protein was reduced with 20 mM dithiothreitol (DTT) for 30 min at room temperature (RT). Following reduction, protein samples were alkylated with 40 mM iodoacetamide (IAA) for 30 min at RT protected from light. To quench the alkylation reaction, DTT was added to a final concentration of 10 mM. Samples were then diluted in 50 mM ammonium bicarbonate (ABC) buffer (pH 7.5) to reduce the urea concentration to below 1 M. Proteins were digested in-solution by adding 1 µg of Pierce™ Trypsin/Lys-C Protease (Thermo Scientific) per sample and incubating over-night at 37 °C.

Prior to mass spectrometry analysis, peptides were purified using SOLAµ™ SPE cartridges (Thermo Scientific). Between each load, the column was spun through at 1,500 rpm for 1 min. Columns were first activated with 200 µL of 100% methanol, followed by equilibration with 200 µL of 80% acetonitrile (ACN) containing 0.1% formic acid (FA), and then two sequential loads of 200 µL of 3% ACN with 1% trifluoroacetic acid (TFA). Peptide samples were then loaded onto the columns and washed with two volumes of 200 µL of 0.1% FA to remove non-specifically bound material. Elution was performed in two steps using 30 µL of 40% ACN with 0.1% FA. The eluates were dried in a vacuum centrifuge and analyzed using mass spectrometry.

3.3.2 Mass spectrometry

The mass spectrometry analyses for the biofilm samples were conducted at The Proteomics Core Facility at the Danish Technical University, Denmark. Dried peptides were resuspended in 0.1% formic acid (FA) and loaded onto a 2 cm C18 trap column (Thermo Fisher Scientific, 164946), which was connected in-line to a 15 cm C18 reverse-phase analytical column (Thermo EasySpray, ES904). Peptide separation was performed using the Thermo EasyLC 1200 HPLC system with 100% 0.1% FA in water at a pressure of 750 bar, and the column oven maintained at 30 °C. Peptides were eluted over a 140-minute gradient from 6% to 60% of 80% ACN in 0.1% FA at a flow rate of 250 nL/min. Eluted peptides were analyzed using an Exploris 480 mass spectrometer (Thermo Fisher Scientific) equipped with a FAIMS Pro™ Interface. The FAIMS interface alternated between compensation voltages of –50 V and –70 V, cycling through 38 and 28 scans, respectively. Full MS1 spectra were acquired at a resolution of 60,000, with an automatic

gain control (AGC) target of 300%, a scan range of m/z 375 – 1,500, and maximum injection time set to "auto." Precursors with an intensity above 5×10^3 and charge states of 2 to 6 were selected for MS2 analysis. Dynamic exclusion was enabled for 60 s, with the exclusion list shared across compensation voltage (CV) settings. Advanced Peak Determination was activated. MS2 precursor ions were isolated using the quadrupole with a 1.6 m/z window. Fragmentation was carried out with a normalized AGC target of 75%, and the maximum injection time was set to "auto."

Fragmentation was executed using higher-energy collisional dissociation (HCD) with a normalized collision energy of 28%, and MS2 spectra were recorded at a resolution of 15,000. The acquired MS spectra were processed using the MaxQuant software suite v2.4.2.0 (Cox & Mann, 2008; Tyanova et al., 2016), integrated with the Andromeda search engine (Cox et al., 2011). Database searches were performed against a target-decoy database of *E. coli* UTI89 (NCBI Assembly accession: ASM1326v1; GCA_000013265.1), which includes 5,211 protein entries. The protease setting was specified as Endoprotease Trypsin/P, allowing for a maximum of two missed cleavages. Carbamidomethylation (Cys) was specified as a fixed modification. Label-free quantification was applied, with a minimum ratio count of two. A false discovery rate (FDR) of 1% was enforced at both the peptide and protein levels to ensure accurate identification, with initial mass tolerance set to 20 ppm. The intensity-based absolute quantitation (iBAQ) option was enabled, with the log fit feature turned off. All other parameters were left at their default settings. These analyses provided iBAQ values for 2,525 protein entries in the entire biofilm data and 2,355 protein entries in the spatial biofilm data. In the entire biofilm dataset, the third biological replicate at the 12H time point was excluded from analysis due to a low number of predicted proteins. Similarly, in the proteome dataset for spatial biofilm samples, the third biological replicates for the 3D and 14D time points of the inner biofilm region were excluded due to an uncharacteristically low number of predicted proteins.

3.3.3 Proteome data analyses

Expression similarity across time points and biological replicates for 2,518 proteins in the entire biofilm dataset, and across biofilm regions and replicates for 2,355 proteins in the spatial dataset, was assessed using Principal Component Analysis (PCA) implemented via the stats package in R v4.2.2 (R Development Core Team, 2008). PCA plots were generated using the ggplot2 package v3.5.0 (Wickham, 2016), based on raw iBAQ values (Figure 1d

and Figure 17d; see Data and code availability). Proteomic data were normalized by dividing each iBAQ value by the total sum of iBAQ values within the corresponding sample, thereby yielding partial concentration values. Median values were calculated across replicates, excluding any with zero iBAQ values. The resulting normalized protein expression data were used for downstream evolutionary analyses (see Phylostratigraphic analysis).

To prepare the proteome data for protein expression profiling, transcriptome–proteome correlation, and visualization, we excluded genes with zero-expression values at more than one stage. This filtering step resulted in a final set of 1,910 proteins for the entire biofilm dataset and 2,057 proteins for the spatial biofilm dataset. For proteins with a zero-expression value in only one stage, missing values were interpolated by averaging the values of the two adjacent stages. In cases where the zero-expression value occurred at the first or last stage of biofilm ontogeny, the missing value was imputed using the value of the only neighboring stage. To bring protein expression profiles to the same scale, each gene's expression was normalized to its median value and subsequently subjected to $\log_2$ transformation, resulting in standardized per-gene expression values. To evaluate transcriptome–proteome correlation, we calculated Pearson's correlation coefficients (R) for 1,858 matched genes/proteins, comparing normalized expression values across biofilm time points (Figure 2b, Appendix 1).

To estimate the overall differential expression of each protein across all biofilm developmental stages and spatial regions within biofilm, we employed open question of how such evolutionary signatures are organized within spatially heterogeneous biofilm communities. The likelihood ratio test (LRT) using the DESeq2 v1.38.3 (Love et al., 2014). Pairwise differential expression analysis between successive biofilm samples as well as between inner, middle, and outer biofilm region of 1D, 3D, and 14D biofilms was also performed using DESeq2 package v1.38.3 (Love et al. 2014). A protein was considered differentially expressed if the adjusted p-value, corrected for multiple comparisons using the Benjamini–Hochberg procedure, was below 0.05.

3.4 Phylostratigraphic analysis

The phylostratigraphic analysis conducted in this study builds upon previously established methodologies (Domazet-Lošo et al., 2007; Domazet-Lošo & Tautz, 2010a; Domazet-Lošo

& Tautz, 2010b; Domazet-Lošo et al., 2017; Futo et al., 2021). Phylotranscriptomic and phyloproteomic approaches aim to trace the evolutionary origin of genes and proteins through similarity searches across genomes and proteomes representing the phylogeny of an organism of interest. This enables the construction of a phylogenetic scale in which each gene and protein from the focal organism is assigned to a specific phylostratum, corresponding to its most ancient detectable homolog in the phylogeny tree (Domazet-Lošo et al., 2007; Domazet-Lošo & Tautz, 2010a). Following the relevant phylogenetic literature (Zaremba-Niedzwiedzka et al., 2017; Mukherjee et al., 2017; Hug et al., 2016; Wu et al., 2009; Brandis, 2021; Brown & Volker, 2004; Sharma et al., 2022; Williams et al., 2010; Adélou et al., 2016; Janda et al., 2021; Dunne et al., 2017; Denamur et al., 2020; Yu et al., 2021), we constructed a consensus phylogeny covering the lineage from the last common ancestor of cellular organisms to our focal species, *E. coli* UTI89 (Figure 4). Clade selection was guided by their representation in the literature, their roles in key evolutionary transitions, and the availability of high-quality reference genomes and proteomes.

We obtained the complete set of protein sequences for 922 terminal taxa represented in our consensus phylogeny from the NCBI database. These proteomes collectively served as the reference database for phylostratigraphic analysis. To construct the *E. coli* UTI89 phylostratigraphic map of *E. coli* UTI89 (Domazet-Lošo et al., 2007; Domazet-Lošo & Tautz, 2010a), we performed a comparative analysis of 5,211 *E. coli* protein coding genes against the reference database. This comparison utilized the blastp algorithm v2.9.0 (Altschul et al., 1990) with an e-value threshold of 10^{-3} . Following the established phylostratigraphic methodology (Domazet-Lošo et al., 2007; Domazet-Lošo & Tautz, 2010a; Futo et al., 2021), we assigned 5,206 protein sequences that met the threshold criteria to one of the nine phylostrata defined in the consensus phylogeny. Each protein was placed in the oldest phylostratum in which a significant homologous hit was detected, representing its earliest traceable evolutionary origin (Domazet-Lošo et al., 2007; Domazet-Lošo & Tautz, 2010b).

Transcriptome age index (TAI) represents the weighted mean of phylogenetic ranks (phylostrata), whose statistical properties and the biological interpretation have been previously described (Domazet-Lošo & Tautz, 2010a):

$$\text{TAI} = \frac{\sum_{i=1}^n ps_i e_i}{\sum_{i=1}^n e_i},$$

where ps_i is an integer indicating the phylostratum of the protein i , e_i is the normalized

transcript expression value of the gene i , and n is the total number of genes analyzed. Using expression values for 4,984 protein coding genes, we calculated the TAI for each developmental stage (entire biofilm dataset) using myTAI package v0.9.3. (Drost et al., 2018) (Figure 5a). In the dataset for spatial biofilms, TAI was calculated using expression values of 4,979 protein coding genes.

To analyze the proteome data in similarly, we introduced the Proteome age index (PAI), which is the weighted mean of phylogenetic ranks (phylostrata) using normalized protein expression values and assigned protein phylostratum, considering the total number of proteins analyzed:

$$PAI = \frac{\sum_{i=1}^n p_{si} q_i}{\sum_{i=1}^n q_i},$$

where p_{si} represents an integer-coded phylostratum assigned to protein i , q_i is the normalized protein expression value of the protein i , and n corresponds to the total number of proteins analyzed. From the bulk biofilm dataset, we used 2,518 proteins for PAI calculation and from spatial biofilms dataset, we obtained 2,355 proteins for PAI calculation (Figure 5b)

TAI and PAI profiles used to assess the phylogeny-ontogeny correlation were evaluated for statistical significance using the ‘Early conservation test’ (Drost et al., 2018). This test requires the predefinition of developmental modules corresponding to early, middle, and late biofilm developmental stages. Based on principal component analysis (PCA) and Pearson’s correlation patterns derived from both transcriptome and proteome datasets (Figure 1), we defined the developmental stages as follows: the early stage from 6H to 1D, the mid stage from 1.5D to 3D, and the late stage from 5D to 1M. These module definitions were used to test for evolutionary conservation patterns across biofilm ontogeny. The statistical significance of differences in mean TAI and PAI values between the inner and outer regions of 1D and 3D biofilms, and among pairs of regions (inner–middle, inner–outer, and middle–outer) in 14D biofilms, was assessed using the myTAI pairwise test (Figure 23).

We estimated evolutionary divergence rates of *E. coli* UTI89 proteins with regard to *Klebsiella pneumoniae* and *Yersinia pestis*. Comparison to these two bacterial species gave the best results in terms of the sufficient number of detected orthologues and the suitability of the evolutionary distances for estimating evolutionary rates. We identified orthologous

pairs by reciprocal best hits using blastp v2.13.0 with an e-value threshold of 10^{-5} . We obtained 3,137 orthologs for comparison with *Klebsiella pneumoniae* MGH78578 (NCBI Assembly accession: GCA_000016305.1; ASM1630v1) and 2,530 orthologs for comparison with *Yersinia pestis* biovar Microtus str. 91001 (NCBI Assembly accession: GCA_000007885.1; ASM788v1). We globally aligned *E. coli* UTI89 and each bacterial strain to obtain orthologous pairs using the Needleman–Wunsch algorithm, followed by codon alignments in pal2nal (Suyama et al., 2006). The non-synonymous (dN) and synonymous (dS) substitution rates were estimated for each bacterial species independently using Comeron’s method (Comeron, 1995). The entire workflow to calculate dN and dS was performed using R package orthologr v0.4.0 (Drost et al., 2015).

Using the calculated dN values, we calculated the transcriptome non-synonymous divergence index (TdNI), which represents the weighted mean of non-synonymous gene divergence:

$$\text{TdNI} = \frac{\sum_{i=1}^n dN_i e_i}{\sum_{i=1}^n e_i}$$

TdNI profiles are calculated using a real number representing the non-synonymous divergence of the gene i (dN_i), normalized transcript expression value of the gene i (e_i), and the total number of genes analyzed (n) (Futo et al., 2021).

Similarly, dS values were used to compute the transcriptome synonymous divergence indices (TdSI), which represents the weighted mean of synonymous gene divergence:

$$\text{TdSI} = \frac{\sum_{i=1}^n dS_i e_i}{\sum_{i=1}^n e_i},$$

where dS_i is a real number that represents the synonymous divergence of gene i , e_i is the normalized transcript expression value of the gene i , and n is the total number of genes analyzed (Futo et al., 2021).

For proteome analysis, we introduced the proteome non-synonymous gene divergence index (PdNI) and the proteome synonymous gene divergence index (PdSI), using normalized protein expression values instead of normalized transcript expression values, as done for TdNI and TdSI. Using dN and dS values for the above-mentioned bacterial strains, we calculated the PdNI and PdSI, representing the weighted mean of non-synonymous and synonymous divergence:

$$PdNI = \frac{\sum_{i=1}^n dN_i q_i}{\sum_{i=1}^n q_i},$$

where dN_i is a real number representing the nonsynonymous divergence of protein i , q_i denotes the normalized expression level of protein i , and n refers to the total number of proteins included in the analysis, and:

$$PdSI = \frac{\sum_{i=1}^n dS_i q_i}{\sum_{i=1}^n q_i},$$

where dS_i represents the synonymous divergence value of protein i , q_i is the normalized protein expression value of protein i that acts as weight factor and n is the total number of proteins analyzed.

Transcriptome codon usage bias index (TCBI) represents a weighted arithmetic mean of 4,984 transcriptome expression values and the measure independent of length and composition (MILC) (Supek & Vlahoviček, 2005, Futo et al., 2021):

$$TCBI = \frac{\sum_{i=1}^n MILC_i e_i}{\sum_{i=1}^n e_i},$$

where MILC is a real number which represents the codon usage bias of gene i , e_i is the normalized transcript expression value of the gene i , and n is the total number of genes analyzed. MILC is a real number representing the codon usage bias of a gene and was calculated using R package coRdon v1.16.0 (Elek et al., 2019), in relation to codon usage bias of *E. coli* UTI89 ribosomal genes. Proteome codon usage bias index (PCBI) was similarly calculated using 2,518 protein expression values and MILC values (Futo et al., 2021):

$$PCBI = \frac{\sum_{i=1}^n MILC_i q_i}{\sum_{i=1}^n q_i}.$$

PCBI is weighted arithmetic mean of codon usage bias, where MILC is a real number which represents the codon usage bias of protein i , e_i is the normalized protein expression value of the protein i , and n is the total number of proteins analyzed.

We calculated the relative expression of genes (Figure 7) for each phylostratum (ps) and developmental stage (s) as follows:

$$RE(ps)_s = \frac{\bar{f}_{ps,s} - \bar{f}_{ps,min}}{\bar{f}_{ps,max} - \bar{f}_{ps,min}},$$

where $\bar{f}_{ps,s}$ is the mean normalized transcription value of genes assigned to phylostratum

(ps) at the given stage (s), and $\bar{f}_{ps,min}$ and $\bar{f}_{ps,max}$ are the minimal and maximal mean normalized transcription values of genes from the same phylostratum (ps) across all stages (Domazet-Lošo & Tautz, 2010b).

To assess the similarity of relative expression values (RE) between phylostrata, we calculated Pearson's correlation coefficients in an all-against-all manner across phylostrata ($ps1$ – $ps9$) (Figure 7). To test whether biofilm ontogeny recapitulates gene phylogeny at the level of relative expression, we calculated the relative expression transcriptome age index ($reTAI$) for each developmental stage (s) using the equation:

$$reTAI_s = \sum_{i=1}^n ps_i \times RE_{i,s},$$

where ps_i is the integer-coded phylostratum i and $RE_{i,s}$ is the relative expression of phylostratum i at stage s and n is the total number of phylostrata. Pearson's correlation coefficients were then calculated between $reTAI_s$ and the ordinally coded biofilm developmental stages.

To characterize the relative expression of genes within each phylostratum (ps) by functional category, we assigned Gene Ontology (GO) functional categories to each gene as described below. For each GO category ($func$), we calculated the mean normalized transcription value within each phylostratum (ps) and developmental stage (s). Using these values, we computed the relative expression of functional categories for each phylostratum and stage as follows:

$$RE(func)_{ps,s} = \frac{\bar{f}_{ps,s} - \bar{f}_{ps,min}}{\bar{f}_{ps,max} - \bar{f}_{ps,min}},$$

where $\bar{f}_{ps,s}$ is the mean normalized transcription value of a given functional category ($func$), at the given phylostratum (ps) and stage (s). $\bar{f}_{ps,min}$ and $\bar{f}_{ps,max}$ represent the minimum and maximum mean normalized transcription values of the given functional category ($func$) within the same phylostratum (ps) across all stages (Domazet-Lošo & Tautz, 2010b). To identify which functional categories best correspond to the relative expression profile in a given phylostratum, we calculated Pearson correlation between the $RE(func)_{ps,s}$ and the $RE(ps)_s$ for every functional category.

To control for the potential influence of horizontal gene transfer (HGT) on our evolutionary age estimates, we recalculated the TAI and PAI indices after filtering out candidate HGT-derived genes (Figure 6, Table S1, Appendix A). Candidate HGT-derived genes for *E. coli*

UTI89 were identified using the HGTree database, utilizing both its 'standard' and 'strict' classification thresholds, which are defined by reference database size (Choi et al., 2023).

3.5 Enrichment analysis and functional annotation

Gene Ontology (GO) functional annotations for the *Escherichia coli* UTI89 strain were obtained by identifying orthologs using the eggNOG v5.0 database (Huerta-Cepas et al., 2019) through eggNOG-mapper v2.1.9 (Cantalapiedra et al., 2021). Annotations were generated using the default search parameters, with the taxonomic scope restricted to Bacteria. This resulted in GO term assignments for 3,108 genes. GO term names and definitions used for functional enrichment analyses were downloaded from the Gene Ontology Resource (<https://geneontology.org/docs/download-ontology/>, accessed February 9, 2023). Prior to enrichment analysis, synonyms and obsolete GO terms were removed to unify the terminology. Genes without annotations from eggNOG-mapper were manually labelled with a custom tag: "GO:no_ID no annotation". To identify enriched biological functions in transcriptome and proteome datasets, we performed one-tailed hypergeometric tests. P-values from these enrichment analyses were corrected for multiple testing using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

We performed functional enrichment analysis of Gene Ontology (GO) annotations at individual developmental stages for both transcriptome and proteome data (Figure 9; Figure 10). Enrichment of GO categories was assessed for genes with standardized expression values ≥ 0.5 ($\log_2$ scale) above the median and a fold change (FC) > 2 during biofilm development. For the proteome dataset, enrichment analysis was conducted on proteins with standardized expression values ≥ 0.15 ($\log_2$ scale) above the median and FC > 0 . Additionally, we investigated the distribution of functionally uncharacterized, yet expressed, genes across phylostrata during biofilm development (Figure 11; Appendix B).

We analyzed Gene Ontology (GO) enrichment profiles for differentially expressed transcripts and proteins across spatial regions of biofilms. Pairwise differential expression analyses between the inner, middle, and outer biofilm regions at 1-day (1D), 3-day (3D), and 14-day (14D) biofilm time points were performed using the DESeq2 package v1.38.3 (Love et al., 2014). GO functional enrichment was conducted on transcripts showing significant differential expression (p-value < 0.05) with FC > 2 , and on proteins with significant differential expression (p-value < 0.05) and FC > 0 (Figures 18–21).

Visualization of the enrichment results was done using custom R scripts built on the ggplot2 package v3.5.0 (Wickham, 2016).

Structural support for annotated versus unannotated *E. coli* UTI89 proteins was evaluated using AlphaFold predictions (Jumper et al., 2021). Local prediction confidence score was assessed by extracting the mean AlphaFold confidence values from the B-factor fields, while secondary structural composition (helices and sheets) was defined using the DSSP algorithm (Appendix C). Statistical differences in structural confidence and secondary structure density between annotated and unannotated protein groups were assessed via a two-sided Mann–Whitney U test with continuity correction (Figure 15, Appendix D).

3.6 Data and code availability

The transcriptome datasets generated in this study were submitted to the NCBI Gene Expression Omnibus and can be accessed under the GEO Series accession number GSE272617. The mass spectrometry-based proteomics datasets were deposited in the ProteomeXchange Consortium through the PRIDE partner repository and are available under the identifier PXD053727.

The datasets supporting the results presented in this doctoral thesis are available in the Supplementary Material of Tušar et al. (2026).

The custom code used for data processing and analysis are publicly available in the GitHub repository: <https://github.com/bacillus-biofilms/biofilm-data-analysis>

4 RESULTS

4.1 Biofilm development is a stage-organized process

To determine whether *E. coli* UTI89 biofilm development is stage-organized process, we sampled entire biofilms at 12 time points covering a full span of biofilm ontogeny, from a liquid starting culture to a 1-month old biofilm (Figure 1a). We first sampled bulk biofilms across successive developmental stages, treating biofilm as an individual organism to obtain organism-level view of biofilm growth. Unlike approaches that focus on specific biofilm subpopulations, this view allows evolutionary indices to be calculated at the entire-biofilm level, facilitating direct comparison with prior phylotranscriptomic studies in *Bacillus* (Futo et al., 2021) and in various eukaryotes (Domazet-Lošo & Tautz, 2010a; Quint et al., 2012; Cheng et al., 2015; Koska et al., 2025; Lotharukpong et al., 2024).

We recovered transcriptome data for 4,866 (93%) *E. coli* UTI89 genes, which revealed a time-resolved principal component analysis (PCA) profile with three main phases of biofilm ontogeny: early (6H – 1D), mid (1.5D – 3D), and late stage (14D – 1M), with a transitional mid-late stage (5D – 7D) (Figure 1c).

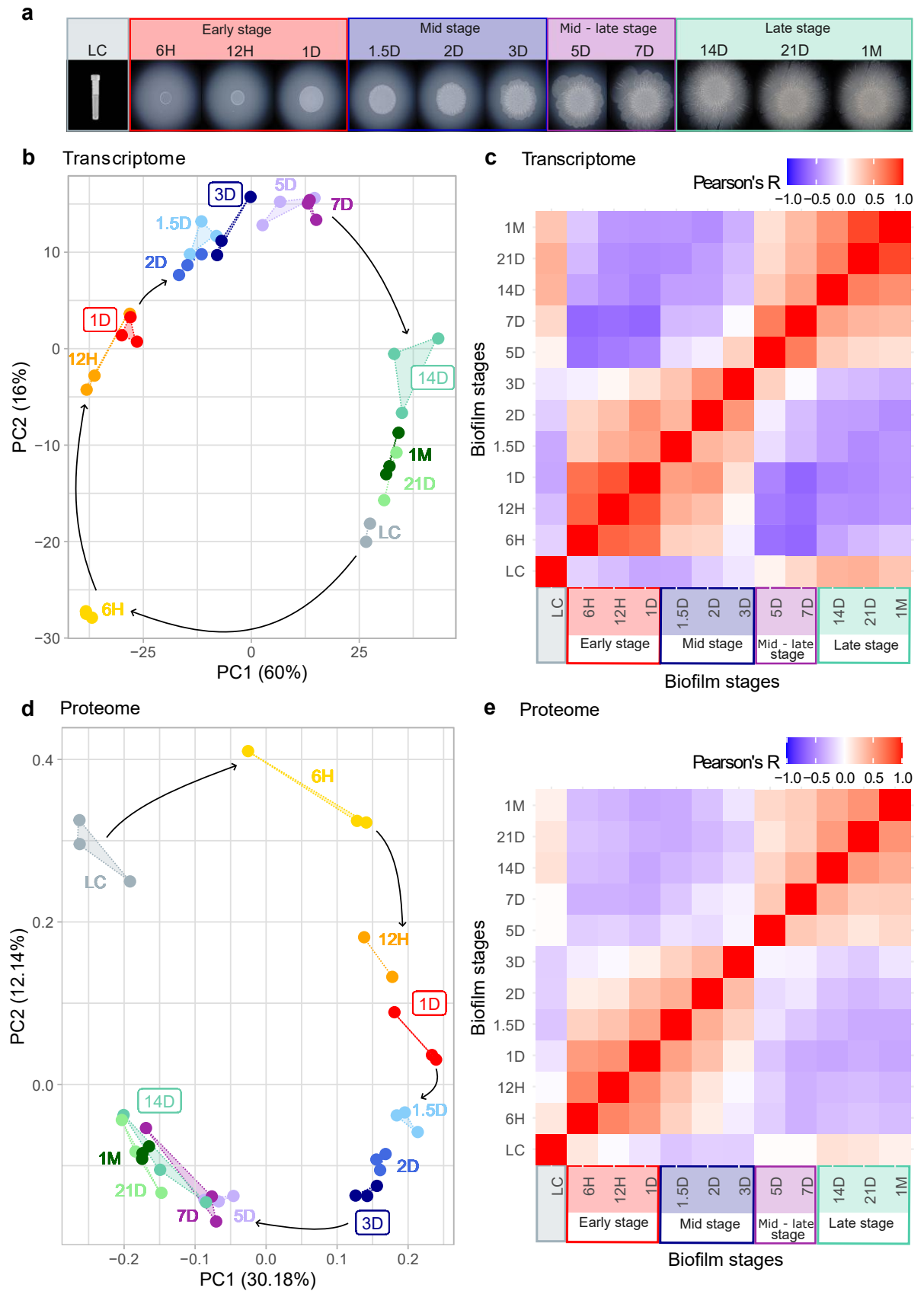


Figure 1. *E. coli* UTI89 biofilm development reveals stage-organized transcription and translation. (a) Representative morphology of *E. coli* UTI89 biofilms on solid agar plates during

biofilm growth at different time points: 6 hours (6H), 12 hours (12H), 1 day (1D), 1.5 days (1.5D), 2 days (2D), 3 days (3D), 5 days (5D), 7 days (7D), 14 days (14D), 21 days (21D), and 1 month (1M) after inoculation with liquid culture (LC). (b, d) Principal component analysis (PCA) of (b) transcriptome and (d) proteome data across time points. Colors on the PCA indicate different time points; biological replicates share the same color and are connected by dotted lines. Black arrows indicate the temporal order of sampled time points. Boxed time points (1D, 3D, and 14D) were used for spatial biofilm experiments. (c, e) Pearson's correlation coefficients between biofilm ontogeny time points were calculated all-against-all comparison for (c) transcriptome data and (e) proteome data. Early (6H–1D), mid (1.5D–3D), mid-late (5D–7D), and late (14D–1M) stages of biofilm growth are depicted in different colors on the x-axis. Supplementary data provided in Tušar et al. (2026).

Pearson's correlation analysis (Figure 1c) and hierarchical clustering of the transcriptome data (Figure 2a) further supported the presence of distinct developmental stages and identified three main phases of biofilm ontogeny. Notably, transcriptome profiles of the liquid culture (LC) sample showed higher correlation with late biofilm stages (14D–1M). This is consistent with the LC sample being collected during stationary phase planktonic growth, a non-proliferative state associated with reduced metabolic activity under nutrient limitation. Such nutrient depleted conditions resemble those experienced by late stage (14D–1M) biofilms grown on agar plates.

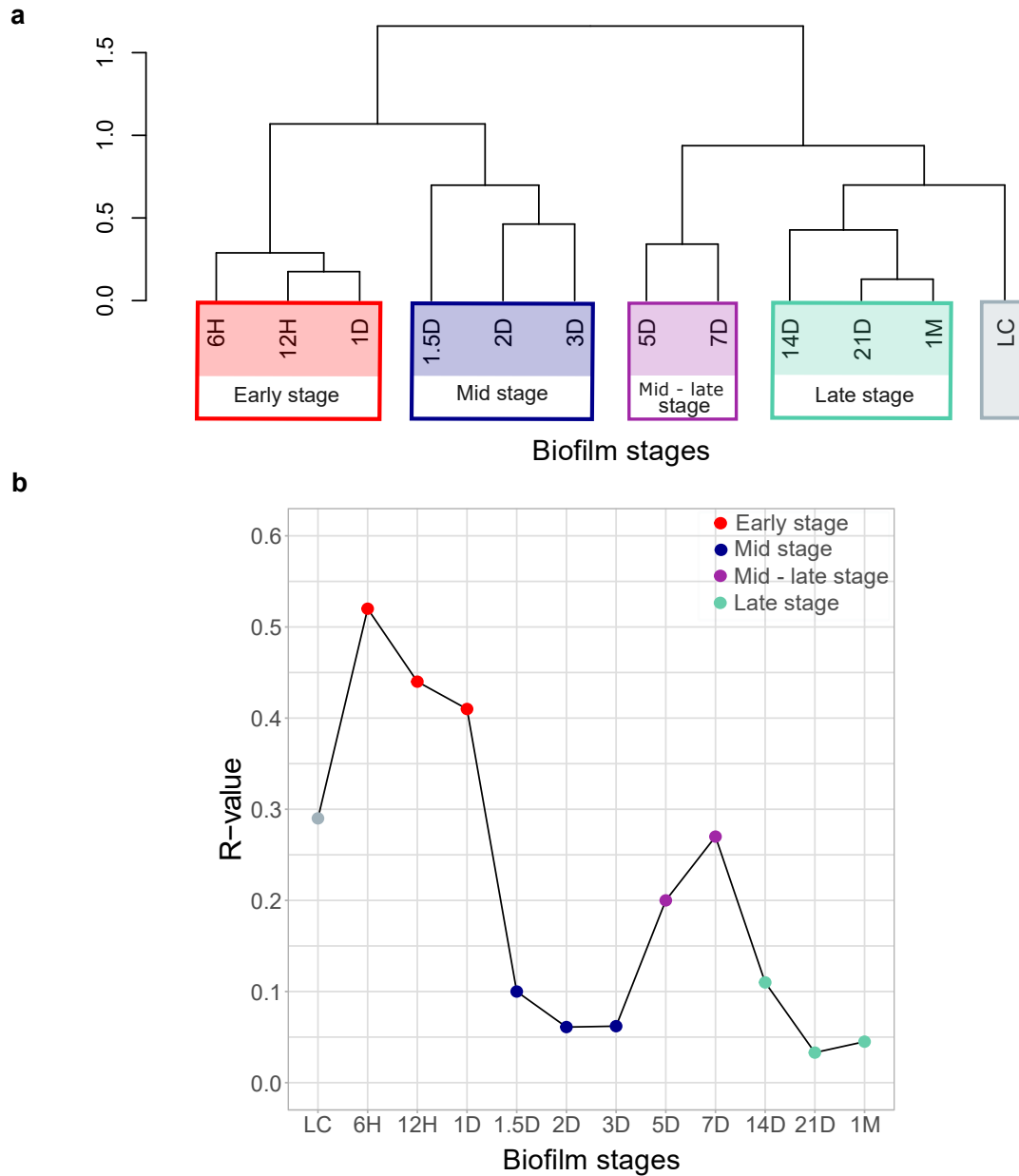


Figure 2. Stage-specific grouping of *E. coli* UT189 biofilm expression profiles. (a) Hierarchical clustering of biofilm transcriptomes based on stage-to-stage dissimilarity, defined as 1 minus R , where R is the Pearson correlation coefficient between two biofilm stages. The scale bar indicates the distance between clusters (see Methods). The clustering supports the evidence for four temporal groups corresponding to early (6H–1D), mid (1.5D–3D), mid-late (5D–7D), and late (14D–1M) biofilm stages. (b) Variation in transcriptome–proteome correlation across biofilm development, expressed as Pearson correlation coefficients (R) between matched transcriptome and proteome profiles at each stage. Correlations were moderate at early stages (6H, 12H, 1D), lower at LC, 5D, and 7D, and very low at 1.5D, 2D, 3D, 14D, 21D, and 1M. These patterns support four broad phases of biofilm development: early (6H–1D), mid (1.5D–3D), mid-late (5D–7D), and late (14D–1M).

Pearson's correlations coefficients were calculated from normalized transcript and protein expression values on a $-\log_2$ scale; corresponding p -values are provided in Appendix 1. Genes or proteins with zero expression in either transcriptome or proteome dataset were excluded from the analysis. Supplementary data provided in Tušar et al. (2026).

Additional analyses that included datasets with unresolved biological replicates supported the robustness of our results, as biological replicates showed consistently high correlation coefficients. These findings indicate strong biological reproducibility of the biofilm developmental program (Appendix 2). A comparable ontogeny progression has been reported for biofilm development of non-pathogenic bacterium *Bacillus subtilis* (Futo et al., 2021). Together with previous work, this finding is consistent with the view that key biofilm-forming strategies may be, at least in part, conserved across bacterial species (O'Toole et al., 2000; Stoodley et al., 2002; Lappin-Scott et al., 2014; Sauer et al., 2022).

Furthermore, we evaluated differential expression along all ontogeny time points which revealed 4,464 (92%) differentially expressed genes. By filtering differentially expressed genes with minimum two-fold expression change, the number of genes decreased to 1,639 (34%). To explore dynamics of expression between successive ontogeny time points, we compared transcriptomes with pairwise comparison analysis (Figure 3). On average, 7% of expressed genes were significantly upregulated by more than two-fold during transitions between successive stages (Figure 3). While the expression shifts observed here are substantial, their overall extent appears lower than reported for *Bacillus subtilis* biofilm development, where 99% of genes (without a fold-change cutoff) and 59% (with a twofold-change cutoff) were differentially expressed across time points (Futo et al., 2021). Nevertheless, these results indicate that biofilm ontogeny involves extensive transcriptional reprogramming, with dynamics that are comparable to those described in complex developmental systems such as animal embryogenesis (Domazet-Lošo & Tautz, 2010a; Yanai et al., 2018). For example, during plant somatic embryogenesis, 84.1% of genes were reported as differentially expressed, and 43.2% of these genes exhibited changes greater than two-fold (Koska et al., 2025).

Consistent with observations in *Bacillus subtilis* biofilms (Futo et al., 2021), the most pronounced transcriptional shift occurred immediately after inoculation of agar plates with liquid culture (Figure 3). This transition from liquid culture (LC) to the early biofilm stage (6H), also apparent in the PCA (Figure 1b), was associated with significant upregulation (more than two-fold) of 26% of expressed *E. coli* genes (Figure 3). In *B. subtilis*, a

comparable transition was accompanied by more than two-fold upregulation of ~30% of genes (Futo et al., 2021). These findings support the view that the shift from planktonic growth to a surface-attached community entails extensive transcriptome reprogramming in bacteria (O'Toole et al., 2000; Futo et al., 2021; Sauer et al., 2022). The second most dynamic increase in gene upregulation occurred later in development, during the 7D–14D transition, when ~18% of expressed *E. coli* genes were significantly upregulated by more than two-fold (Figure 3). Overall, the transcriptome data indicate that *E. coli* UTI89 biofilm development proceeds in discrete phases (early, mid, mid-late, and late), each marked by characteristic shifts in gene expression.

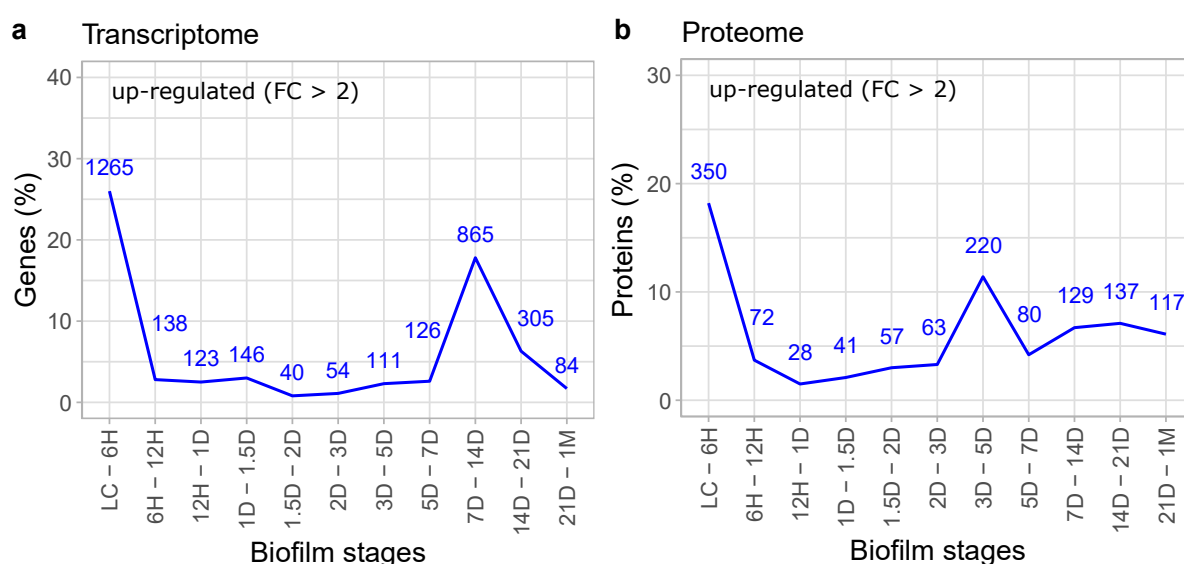


Figure 3. Differential expression in pairwise comparison between successive biofilm samples. A burst of up-regulated differentially expressed genes are observed at biofilm inoculation (LC–6H) in both (a) transcriptome and (b) proteome datasets. Numbers above each time point indicate the total number of differentially expressed genes or proteins. Pairwise differential expression was assessed using DeSeq2, with p-values adjusted for multiple comparisons testing using the Benjamini and Hochberg procedure (cut off set at 0.05). Up-regulated (a) genes and (b) proteins are defined by a fold-change greater than two (FC > 2). Supplementary dataset is provided in Tušar et al. (2026).

Transcriptome and proteome data show generally low correlation (Figure 2, Appendix 1; Liu et al., 2016; Futo et al., 2021), underscoring the value of assessing both molecular layers to capture expression dynamics. At the same time, current mass spectrometry–based proteomics typically provides lower genome-wide coverage and reduced sensitivity for low-abundance proteins than RNA sequencing (Buccitelli & Selbach, 2020). Thus, while

RNA sequencing offers a more comprehensive view of global expression changes, integrating transcriptome and proteome datasets yields a more complete picture by linking transcriptional regulation to protein-level output (Futo et al., 2021; Buccitelli & Selbach, 2020).

Therefore, we quantified proteome data from the same *E. coli* biofilm stages used for transcriptome sequencing (Figure 1a). In total, we identified 2,518 predicted proteins, corresponding to 48% of the theoretical *E. coli* UTI89 proteome. Although proteome coverage was lower than transcriptome coverage, principal component analysis (PCA) and Pearson's correlation analyses revealed similar temporal patterns across the two datasets (Figure 1d,e), including a stage-ordered trajectory in PCA (Figure 1d) and discrete ontogenetic phases (Figure 1e). These findings are consistent with the developmental progression identified in the transcriptome data (Figure 1b,c) and indicate that biofilm development follows a structured, stage-specific program at both the transcriptome and proteome levels.

Compared with the transcriptome data, the proteome data displayed fewer differentially expressed proteins across biofilm ontogeny (1,301 proteins; 68%). Nevertheless, the fraction of upregulated proteins showing at least a two-fold change (711 proteins; 37%) was similar to that observed in the transcriptome dataset. In pairwise comparisons of successive stages, an average of 8% of proteins were significantly upregulated with a minimum two-fold change (Figure 3b), which is lower than the corresponding proportion at the transcript level (Figure 3a). Consistent with the transcriptome data (Figure 3), the largest shift in protein expression occurred during the transition from liquid culture (LC) to the early biofilm stage (6H) (Figure 3).

Overall, the proteome profiles also support a structured developmental progression and delineate distinct ontogeny stages, despite generally low transcriptome–proteome correlations and the lower coverage inherent to protein quantification (Figure 2). Notably, the correspondence between mRNA and protein levels decreased at later stages of biofilm maturation (Figure 2b; Appendix 1). This pattern reflects increased contributions of post-transcriptional and post-translational regulation, differences in mRNA turnover and protein stability, and greater metabolic heterogeneity in mature biofilms. Together, these factors can weaken the relationship between transcript abundance and protein levels, thereby reducing overall transcriptome–proteome correlation (Figure 2b, Appendix 1).

4.2 Phylogeny-ontogeny correlation reveals recapitulation pattern

Evolutionary gene age estimates were derived through a phylostratigraphic procedure (Domazet-Lošo et al., 2007; Futo et al., 2021; Čorak et al., 2023), based on consensus phylogeny. We linked transcriptome and proteome profiles to evolutionary age estimates to test whether *E. coli* UTI89 biofilm development follows an evolutionary logic in the form of phylogeny-ontogeny correlations. We constructed consensus phylogenetic tree spanning from the last common ancestor of cellular organisms to the focal strain *E. coli* UTI89 (Figure 4, Supplementary dataset is provided in Tušar et al. (2026)).

This phylostratigraphic approach resulted in assignment of 5,206 (99.9% of 5,211 predicted genes) *E. coli* UTI89 genes to nine phylostrata (PS); with PS1 representing the oldest and PS9 the youngest phylostratum (Figure 4). Consistent with phylostratigraphic results reported for *Bacillus subtilis* (Futo et al., 2021), older phylostrata contained the largest fraction of *E. coli* UTI89 genes, whereas younger phylostrata were comparatively sparsely populated. Specifically, 2,871 genes (55.1%) were assigned to PS1 (Cellular organisms) (Figure 4). In contrast, only 73 genes (1.4%) mapped to PS8 (*Escherichia coli* group B2/F) and 162 genes (3.1%) to PS9 (*Escherichia coli* UTI89) (Figure 4).

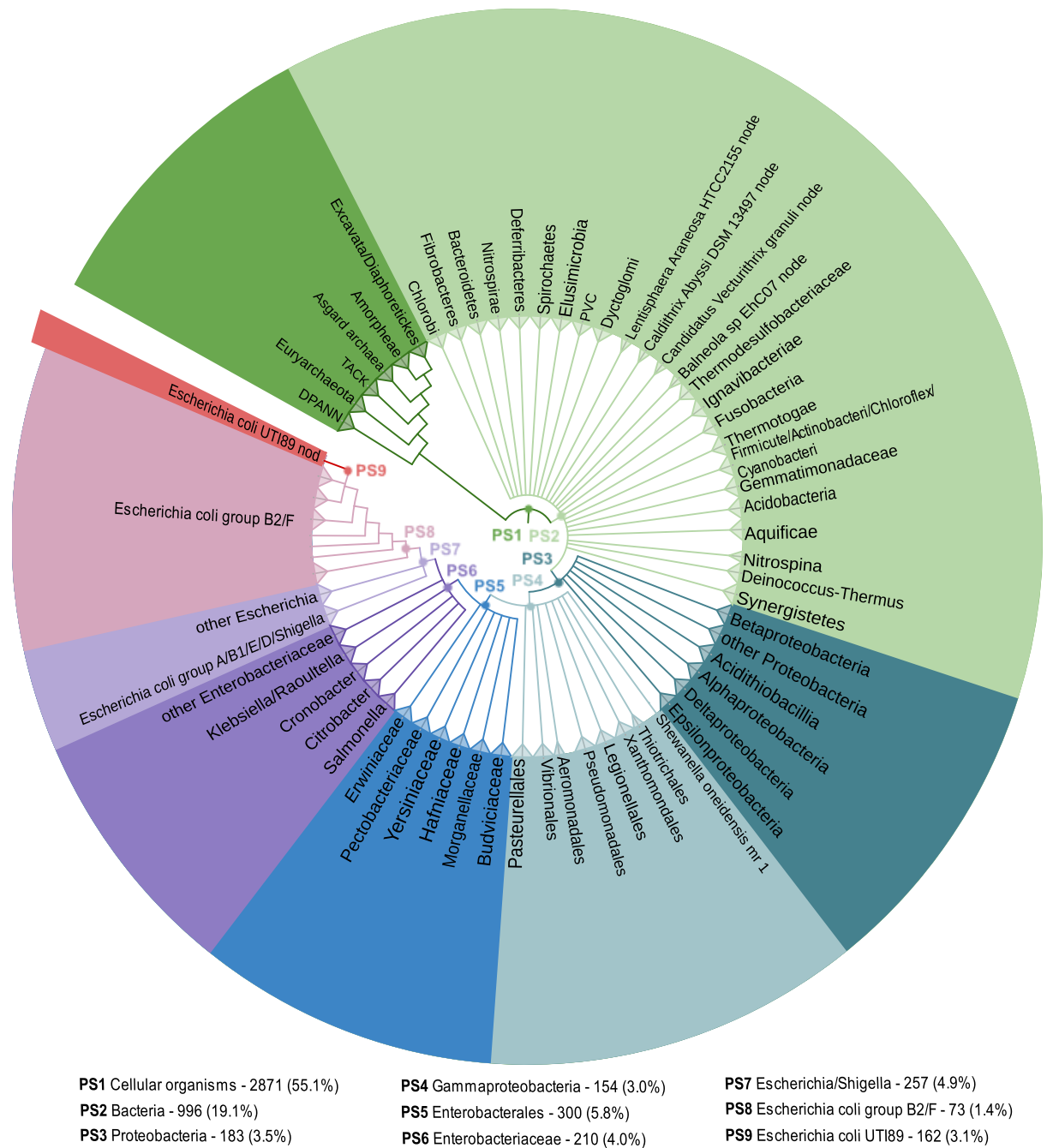


Figure 4. The consensus tree covers the evolutionary divergence from the last common ancestor of cellular organisms to the focal species *E. coli* UTI89. The tree was constructed for phylostratigraphic analysis to reflect the evolutionary trajectory of the focal strain and the availability of reference genomes. The nine internodes (phylostrata) from the root to *E. coli* UTI89) were labelled by PS1–PS9. For each phylostratum, the number and percentage of *E. coli* UTI89 genes assigned to that phylostratum are indicated following the phylostratum label. Supplementary dataset is provided in Tušar et al. (2026).

To assess whether biofilm development follows an evolutionary pattern, we integrated transcriptome and proteome expression values with gene age estimates to calculate the

transcriptome age index (TAI) and proteome age index (PAI) (Figure 5). TAI and PAI are cumulative measures for the evolutionary age of transcriptome and proteome data (Domazet-Lošo & Tautz, 2010a; Futo et al., 2021). As reported for *B. subtilis* biofilms (Futo et al., 2021), *E. coli* UTI89 biofilms showed a positive correlation between biofilm ontogeny stage and the evolutionary age of expressed genes and proteins. This was evident in both the transcriptome (Figure 5a) and proteome (Figure 5b) dataset. Evolutionarily younger genes contributed increasingly to the expression profile as biofilms progressed to later stages (Figure 5). This trend corresponds to a recapitulation pattern observed in evolutionary developmental studies of *B. subtilis* biofilm (Futo et al., 2021).

Notably, at the transcriptome level the trend reverses in the late stages of biofilm: TAI increases between 7D and 1M, indicating a shift toward relatively older transcriptomes during maturation (Figure 5a). A similar decrease in transcriptome “youthfulness” has been reported during aging in animals and plants (Domazet-Lošo & Tautz, 2010b; Koska et al., 2025).

To isolate the evolutionary dynamics unique to biofilm ontogeny, we restricted our primary analyses to solid-agar growth stages (Figure 5), excluding the initial liquid culture (LC) which serves as a pre-inoculation reference rather than a part of biofilm development. Although including the LC yields a false hourglass-like (older–younger–older) pattern at the onset of cultivation (Appendix 3), we do not interpret this as an early developmental phase. Rather than representing a surface-attached developmental program, the elevated TAI and PAI values of the LC reflect a shared physiological state between planktonic cells in stationary phase and late stage *Escherichia coli* biofilms, both of which are characterized by maximal cell density and decelerated growth rates (Appendix 3).

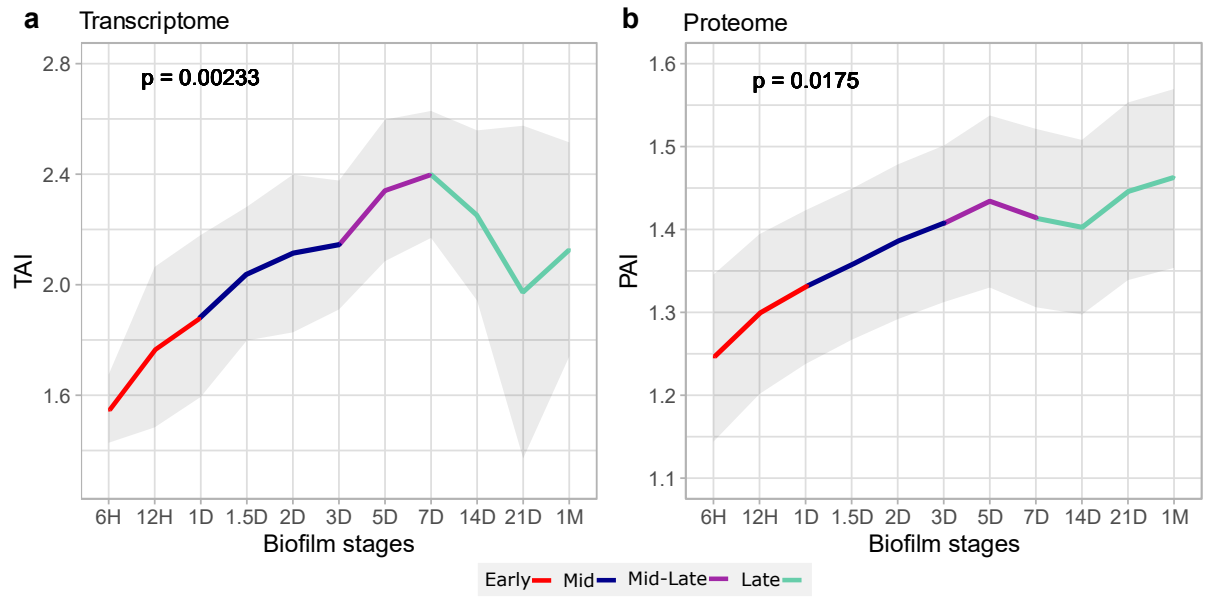


Figure 5. *Escherichia coli* biofilms exhibit positive phylogeny-ontogeny correlation. (a) Transcriptome age index (TAI) and (b) proteome age index (PAI) profiles of *E. coli* UTI89 biofilm ontogeny indicate a recapitulation pattern: early stages are enriched for expression of evolutionarily older genes, whereas later stages show relatively higher expression of younger genes (Supplementary dataset is provided in Tušar et al. (2026)). Statistical support for these evolutionary patterns was assessed using the early conservation test (p-values) (see Methods). Grey shading indicates $\pm$ one standard deviation derived from permutation analyses. Biofilm stages are color-coded as early (red), mid (blue), mid-late (purple), and late (green).

Horizontal gene transfer (HGT) represents a potential confounding factor in bacterial phylostratigraphy, as it can cause recently acquired genes to be erroneously assigned to older evolutionary strata. This typically occurs when a gene is transferred from the focal lineage to a phylogenetically distant group, whereas the reverse scenario (genes acquired via HGT from remote lineages into focal lineage) is minimized by comprehensive reference database coverage. To rigorously account the potential impact of HGT on our phylostratigraphical analysis, we systematically filtered the *E. coli* UTI89 gene set using HGT candidates predicted via phylogenetic tree reconciliation (Choi et al., 2023). HGT-derived genes accounted for only a minor fraction of the genome, 0.4% using standard criterion and 0.1% using strict criterion (Table 1, Appendix A). Recalculation of the TAI and PAI metrics following the exclusion of HGT-derived genes produced patterns highly congruent with the complete dataset (Figure 5, Figure 6), confirming that horizontal gene flow has a negligible impact on our developmental profiles.

Table 1. Distribution of all genes and HGT-derived genes (number n and percentage %) across phylostrata using strict and standard methods for HGT gene detection (Appendix A). Percentages indicate the proportion of HGT-derived genes within each phylostrata. Supplementary dataset is provided in Tušar et al. (2026).

	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	Total
All genes, n	2,871	996	183	154	300	210	257	73	162	5,206
Standard HGT method, n (%)	4 (0.1)	5 (0.5)	3 (1.6)	0 (0.0)	5 (1.7)	4 (1.9)	1 (0.4)	0 (0.0)	2 (1.2)	24 (0.5)
Strict HGT method, n (%)	0 (0.0)	1 (0.1)	1 (0.5)	0 (0.0)	1 (0.3)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (0.1)

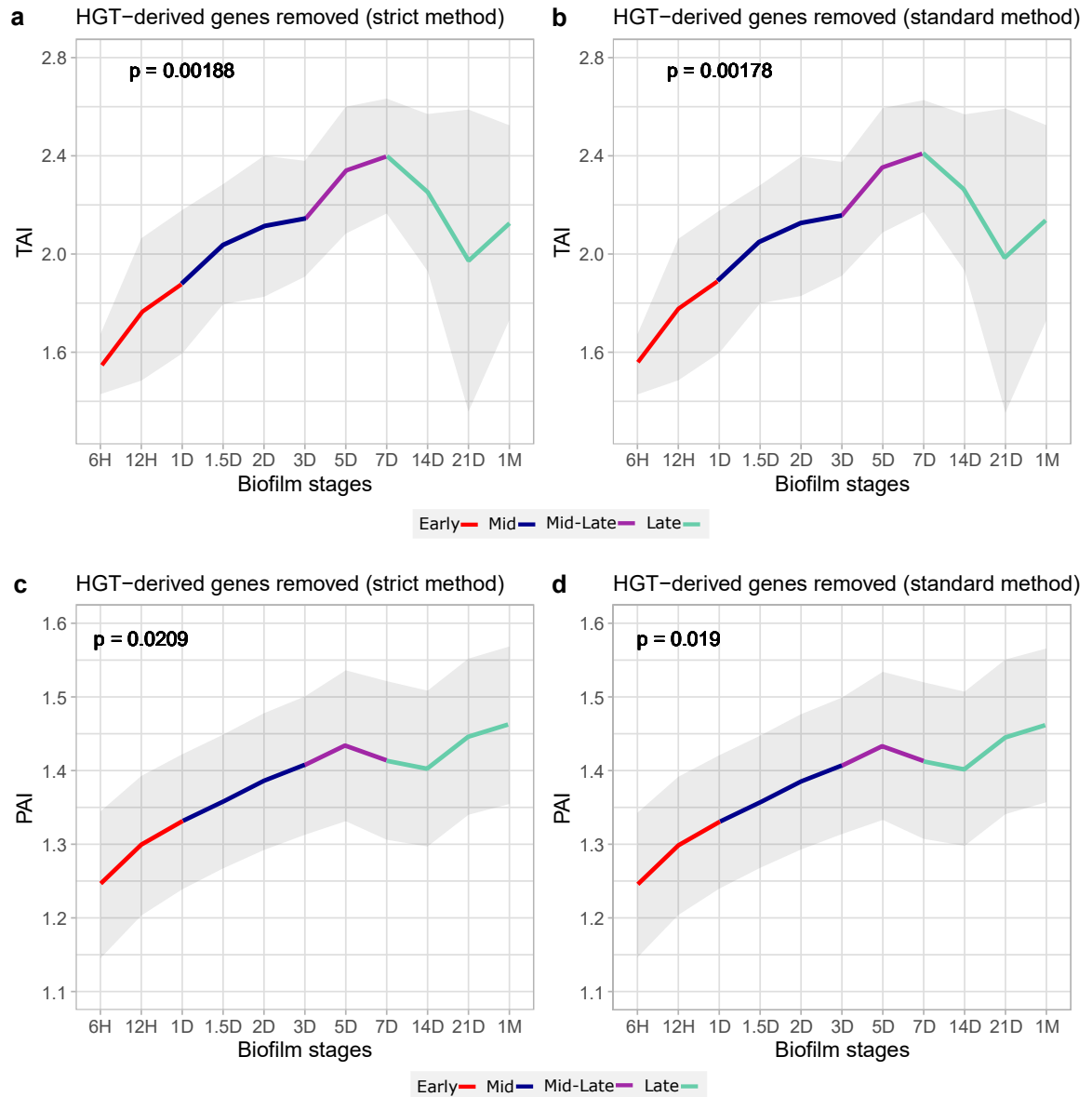


Figure 6. Phylostratigraphic profiles of biofilm development after removing HGT-derived genes. (a, b) Transcriptome age index (TAI) and (c, d) proteome age index (PAI) profiles of *E. coli* UTI89 biofilm ontogeny computed using an HGT-depleted gene set (Appendix A). Excluding candidate HGT genes yields evolutionary profiles that closely mirror the original, unfiltered profiles (Figure 5). The statistical significance (p-value) of the observed evolutionary trends was assessed using the early conservation test (see Methods). Grey shaded areas indicate $\pm$ one standard deviation based on permutation analysis. Distinct stages of biofilm development are represented by color: early (red), mid (blue), mid-late (purple), and late (green).

To identify which evolutionary intervals contribute most strongly to the recapitulation-like ontogeny–phylogeny correlations (Figure 5a), we quantified phylostratum-specific relative expression patterns from the transcriptome data (Figure 7; Domazet-Lošo & Tautz, 2010a).

These profiles clustered into three major groups of phylostrata with shared temporal dynamics (Figure 7a–d). Genes assigned to the oldest phylostratum (PS1; cellular organisms) and to the youngest stratum (PS9; *E. coli* UTI89) exhibited a U-shaped pattern, with elevated relative expression at the earliest biofilm stage (6H) and again during late stage biofilm development (21D–1M) (Figure 7a; Figure 4). In contrast, genes originating at PS2 (Bacteria), PS3 (Proteobacteria), and PS5 (Enterobacterales) reached their highest relative expression predominantly during the mid stages of biofilm development (1.5D–7D) (Figure 7b). A third set of comparatively younger phylostrata, PS4 (Gammaproteobacteria), PS6 (Enterobacteriaceae), and PS7–PS8 (*E. coli* group), showed a more progressive increase across ontogeny, with maximum during the mid-late (5D–7D) and late (14D–1M) phases (Figure 7c). To quantify the overall relationship between relative expression dynamics and biofilm development progression, we correlated the relative expression transcriptome age index (reTAI; see Methods) with ordinally coded biofilm stages (Figure 7e). This analysis revealed a strong positive correlation (Pearson: $R=0.76$, $p\text{-value}=0.007$), providing additional support for a recapitulation-like pattern in the transcriptome data.

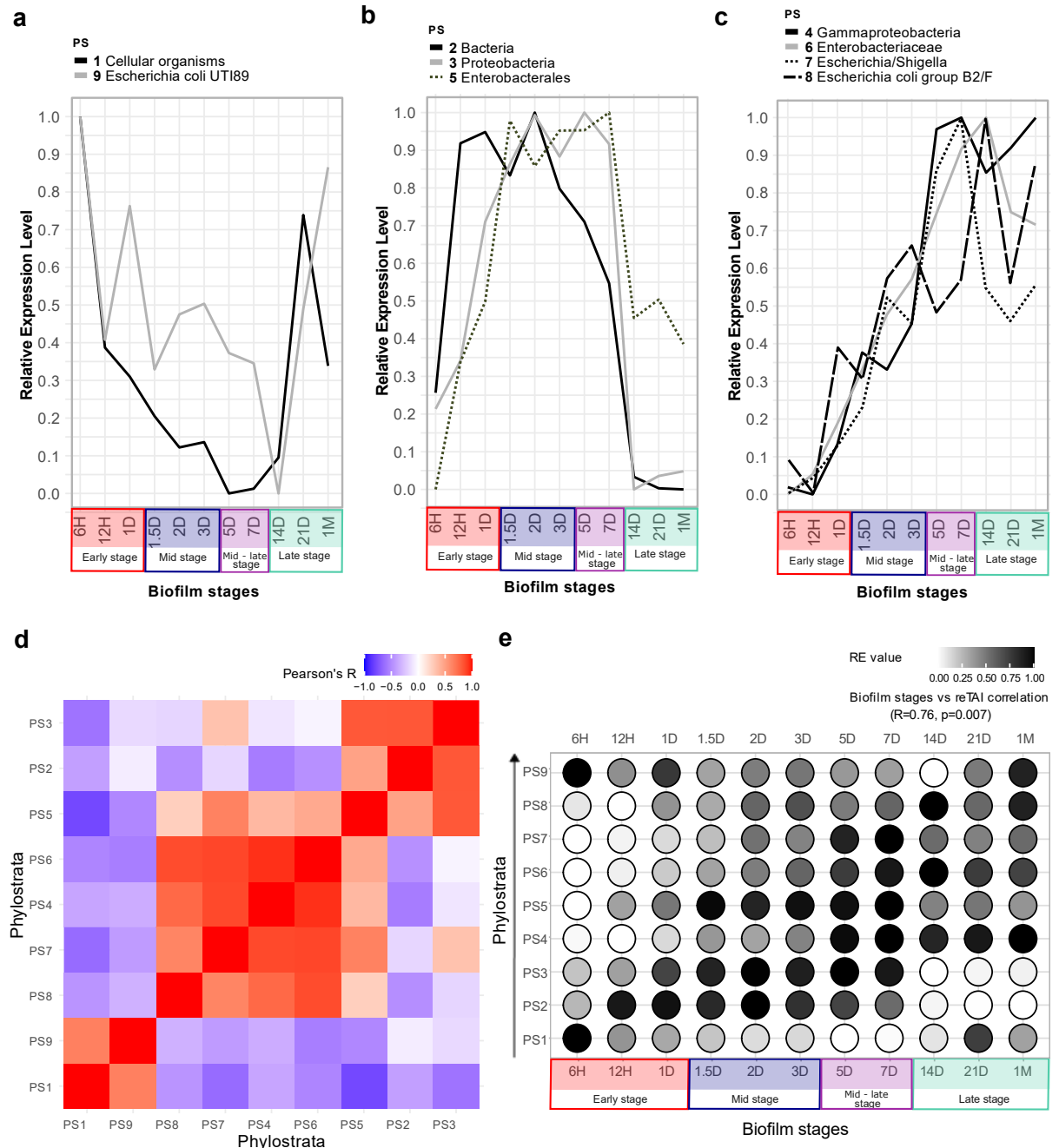


Figure 7. Relative expression per phylostratum across stages of *E. coli* biofilm ontogeny.

Relative expression values were scaled to the maximum (1) and minimum (0) expression values observed across biofilm time points (see Methods). (a) The oldest (PS1) and youngest (PS9) gene sets display similar hourglass-like relative expression profiles. (b) Genes assigned to Bacteria (PS2), Proteobacteria (PS3), and Enterobacterales (PS5) show peak relative expression during the mid stages of biofilm ontogeny. (c) Genes originating in Enterobacteriaceae and younger phylostrata (PS4, PS6–PS8) exhibit elevated relative expression at late stages. (d) Similarity of relative expression values (RE) between phylostrata was quantified by calculated all-against-all Pearson correlation coefficients among phylostrata (PS1–PS9). (e) Relative expression values across biofilm

ontogeny stages and phylostrata are visualized as a bubble plot using black–white gradient. The association between reTAI and ordinaly coded developmental stages was assessed using Pearson correlation; the corresponding *R* and *p*-values are reported on Figure (see Methods). Supplementary dataset is provided in Tušar et al. (2026).

We derived TAI and PAI from a phylostratigraphic framework to summarize the cumulative evolutionary age of transcriptomes and proteomes by mapping genes to their most ancient detectable homologs across the tree of life (Domazet-Lošo et al., 2007; Domazet-Lošo & Tautz, 2010a; Futo et al., 2021). An additional way to quantify evolutionary patterns in expression datasets is to focus on sequence divergence, estimating evolutionary rates of coding sequences rather than the age of entire genes (Quint et al., 2012; Futo et al., 2021; Koska et al., 2025). To capture these evolutionary patterns while accounting for codon usage effects, we introduced transcriptome- and proteome-based divergence indices: nonsynonymous (TdNI, PdNI) and synonymous (TdSI, PdSI) divergence indices (Futo et al., 2021). These indices were derived from alignments of *E. coli* UTI89 coding sequences with those of the closely related Enterobacteriaceae species *Klebsiella pneumoniae* MGH78578 (Figure 8). Using a closely related strain for comparison is important for robust estimation of protein-coding genes sequence conservation and for reliable inference of synonymous and nonsynonymous divergence rates. In the transcriptome, both TdNI and TdSI showed a significant recapitulation-like pattern, with divergence at nonsynonymous and synonymous sites increasing toward later stages of biofilm development (Figure 8ac). A similar trend was observed in the proteome for PdNI and PdSI (Figure 8bd). To evaluate sensitivity to the choice of bacterial species for comparison, we repeated the analysis using the more distantly related *Yersinia pestis* (within order Enterobacterales) and obtained highly consistent results (Appendix 4).

The transcriptome codon bias index (TCBI) and proteome codon bias index (PCBI) also showed a significant recapitulation pattern with codon bias decreasing toward later stages of biofilm development (Figure 8). Taken together, evolutionary indices derived from both transcriptome and proteome data indicate that biofilm ontogeny is shaped by evolutionary constraints acting at multiple depths, from gene age (TAI, PAI) to more recent sequence and codon-level divergence (TdNI, TdSI, TCBI, PdNI, PdSI, PCBI) (Figure 5; Figure 8). Across all these measures, later biofilm stages are characterized by relatively higher expression of more diverged genes, consistent with a recapitulation pattern.

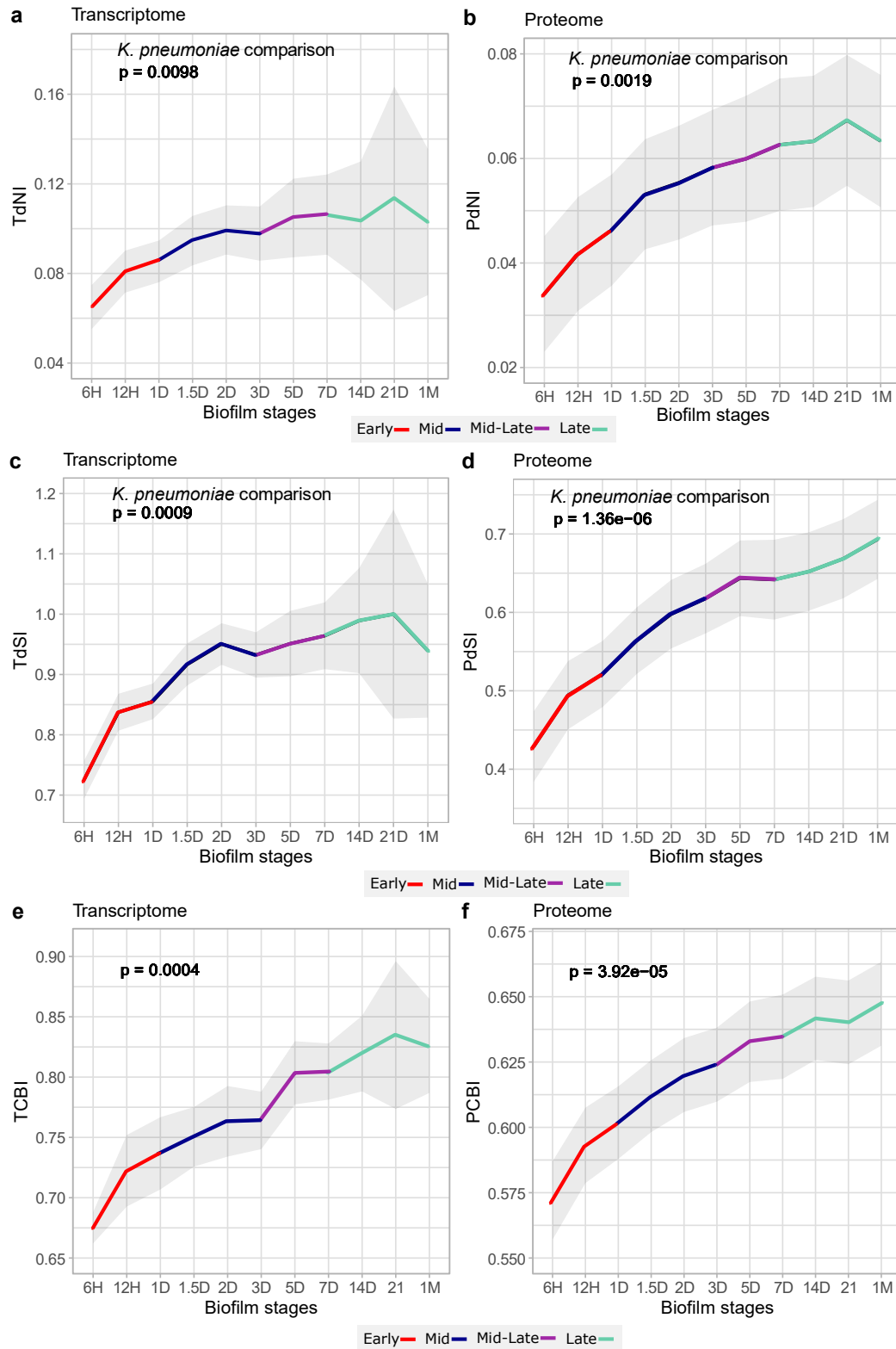


Figure 8. Ontogeny-phylogeny correlation in uropathogenic *E. coli* biofilms at (non)-synonymous sites and codon usage bias. (a) Transcriptome non-synonymous divergence index (TdNI), (b) proteome non-synonymous divergence index (PdNI), (c) transcriptome synonymous divergence index (TdSI), and (d) proteome synonymous divergence index (PdSI) show a similar

recapitulation pattern: early biofilm stages express conserved genes at synonymous and non-synonymous divergence sites, while later stages express more diverged genes. Synonymous and non-synonymous divergence rates were estimated from comparisons of *E. coli* UTI89 with *Klebsiella pneumoniae* (Enterobacteriaceae). Analyses using the more distantly related *Yersinia pestis* (Enterobacterales) yielded comparable results (Appendix 4). Codon usage bias indices for the transcriptome (TCBI) and proteome (PCBI) were calculated using MILC (see Methods) and both exhibited a significant recapitulation pattern. Statistical support (p-value) for these evolutionary patterns was assessed using the early conservation test (see Methods). Grey shading indicates $\pm$ one standard deviation derived from permutation analysis. Stages of biofilm ontogeny are color-coded as early (red), mid (blue), mid-late (purple), and late (green) stage. Supplementary dataset is provided in Tušar et al. (2026).

4.3 Stage resolved functions distribution in *E. coli* biofilm development.

To characterize functional dynamics during *E. coli* UTI89 biofilm development, we performed GO (Gene Ontology)-based functional enrichment analysis for upregulated transcripts and proteins (Figure 9, Figure 10). The resulting enrichment profiles indicate stage-dependent functional specialization, with each developmental stage showing enrichment for a distinct set of biological functions (Figure 9). These profiles parallel observations in *B. subtilis* biofilms, which likewise progress through functionally discrete stages (Futo et al., 2021).

During early biofilm development (6H–1D), enriched GO terms were dominated by biosynthetic pathways, energy generation, ribosome biogenesis/translation, iron homeostasis, and motility, consistent with rapid growth and biofilm establishment (Figure 9). In the mid phase (1.5D–3D), enrichment shifted toward locomotion, bacterial-type flagella, and energy production, suggesting continued expansion at the biofilm periphery and the contribution of flagellated cells as structural elements (Figure 9; Serra et al., 2013a, 2013b). During mid-late (5D–7D) and late (14D–1M) stages, enriched functions were associated with regulation of metabolic processes, post-replication repair, and stress responses (Figure 9). For example, enrichment of SOS response and DNA repair terms at 21D is consistent with increased environmental stress at this stage (Figure 9).

Proteome-based enrichment analyses revealed broadly comparable stage-dependent functional patterns (Figure 10), indicating that biofilm development follows a stepwise functional program at both transcriptome and proteome levels, despite generally modest transcriptome–proteome correspondence (Figure 2).



Figure 9. *E. coli* UTI89 biofilm development is temporally organized into functionally discrete stages in transcriptome data. Functional enrichment analysis of GO annotations was performed

for genes whose expression at a given stage exceeded the median of their overall expression profile by at least 0.5 (log2 scale). Enrichment significance was tested using a one-tailed hypergeometric test and corrected for multiple comparisons with the Benjamini–Hochberg procedure. Only functional enrichments with adjusted $p < 0.05$ were retained. The complete set of enriched GO terms is provided in Tušar et al. (2026), whereas the figure shows a non-redundant subset of GO functions. The magnitude of functional enrichments is shown as log-odds (circle size), and significance (p -value) is indicated by color intensity. Biofilm developmental stages are color-coded as liquid culture (grey), early (red), mid (blue), mid-late (purple), and late (green). Supplementary dataset is provided in Tušar et al. (2026).

These functional profiles indicate that *E. coli* biofilm development is organized into discrete, temporally ordered stages that correspond to the transcriptome and proteome stage structure. In this respect, *E. coli* biofilm ontogeny resembles other developmental systems—including animals (Domazet-Lošo & Tautz, 2010a; Levin et al., 2012; Yanai, 2018), plants (Quint et al., 2012; Koska et al., 2025), and *B. subtilis* biofilms (Futo et al., 2021)—in exhibiting a stage-organized developmental architecture.

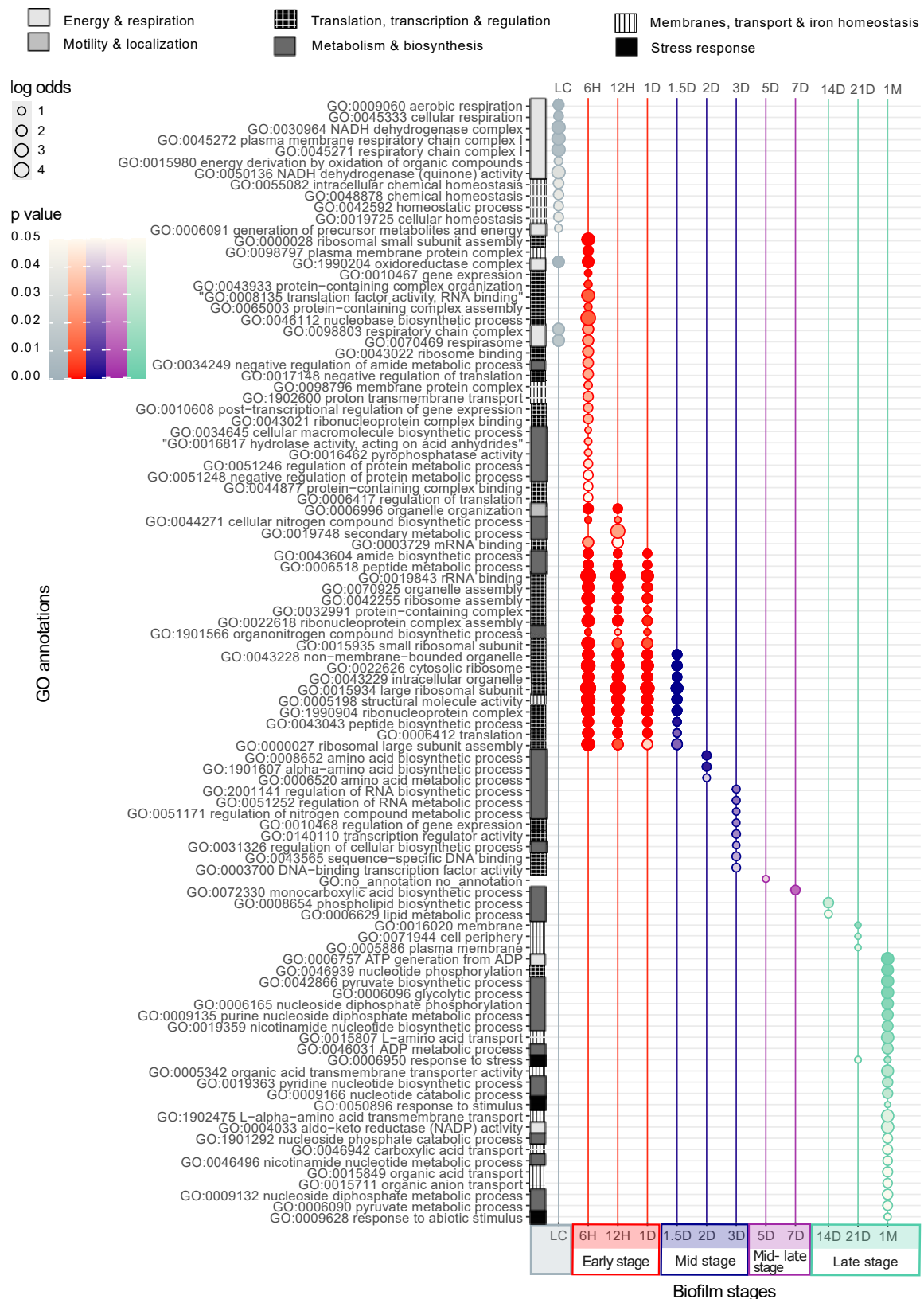


Figure 10. Temporal distribution of functional categories in the *E. coli* UTI89 biofilm proteome. Functional enrichment of GO categories was performed for upregulated proteins with expression levels at least 0.15 (log₂ scale) above the median during biofilm development.

Significance of enrichment was assessed using a one-tailed hypergeometric test and corrected for multiple comparisons using the Benjamini-Hochberg procedure (significance threshold: p -value < 0.05). Functional enrichment magnitude is visualized as log odds (circle sizes), while significance is represented by color intensity corresponding to p -values. Biofilm developmental stages are color-coded as follows: liquid culture (grey), early stage (red), mid stage (blue), mid-late stage (purple), and late stage (green). GO terms were grouped into broader functional categories, as indicated in the top legend. Supplementary dataset is provided in Tušar et al. (2026).

Transcripts annotated as “unknown function” were significantly enriched from mid to late biofilm stages (2D–1M), suggesting that important components of the *E. coli* UTI89 biofilm program remain uncharacterized (Figure 11). Mapping these unannotated genes onto the phylostratigraphic map further showed increasing enrichment from Bacteria (PS2) toward *E. coli*-specific strata (PS9), indicating that they are disproportionately represented among younger phylostrata (Figure 11). This pattern is reminiscent of animal development, where phylogenetically restricted (orphan) genes are often expressed in later ontogenetic stages and can contribute to clade-specific traits (Domazet-Lošo & Tautz, 2010a; Tautz & Domazet-Lošo, 2011; Xia et al., 2025).

a

	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	SUM
All expressed genes	2819 (58.0%)	942 (19.4%)	174 (3.6%)	151 (3.1%)	277 (5.7%)	175 (3.6%)	186 (3.8%)	58 (1.2%)	81 (1.7%)	4863
Unannotated genes	552 (30.3%)	431 (23.6%)	97 (5.3%)	78 (4.3%)	199 (10.9%)	147 (8.1%)	181 (9.9%)	57 (3.1%)	81 (4.4%)	1823

b

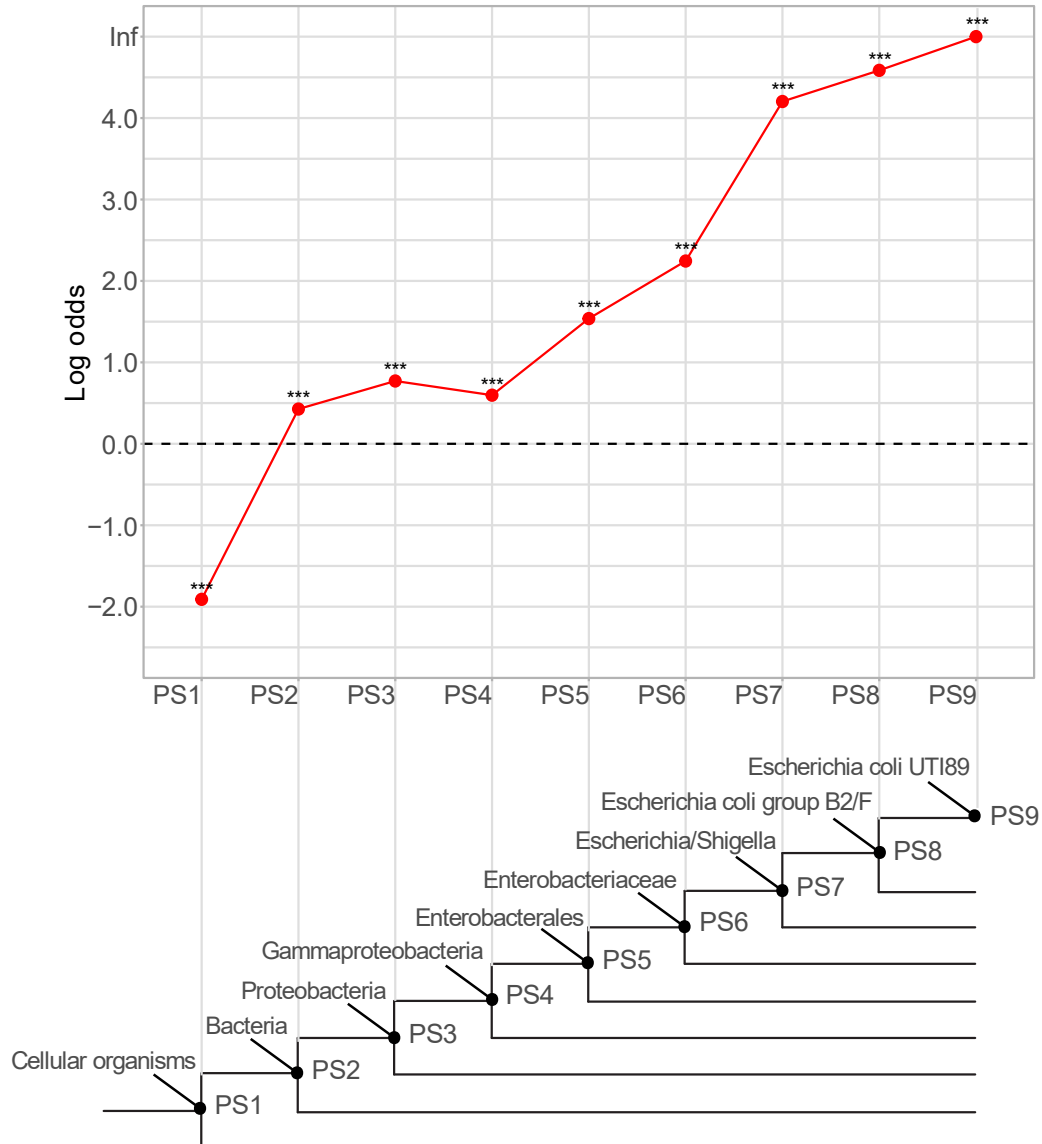


Figure 11. Phylostratigraphic distribution of functionally uncharacterized genes expressed during *E. coli* UTI89 biofilm development. (a) Numbers and percentages of expressed genes across phylostrata are summarized for two sets: all expressed genes and the subset without functional annotation (unannotated genes). (b) Distribution of unannotated genes across the phylostratigraphic map. Enrichment of unannotated genes within phylostrata was tested using a one-tailed hypergeometric test with multiple-testing correction (***) p-value < 0.001 (Appendix B). Supplementary dataset is provided in Tušar et al. (2026).

The recently discovered rosette-initiated multicellular life cycle in *Escherichia coli* is governed by a highly coordinated transcriptional program (Puri et al., 2022; Puri et al., 2024). During early morphogenesis, newly divided sister cells actively reposition and fold into a rosette configuration. This biophysical motion is driven by flagellar activity that requires *de novo* flagella synthesis, suggesting coordination between adhesion and motility during morphogenesis (Puri et al., 2024). To determine whether the molecular logic of this rosette-driven program is conserved during biofilm development, we systematically profiled the temporal expression of its key genes across our biofilm developmental time course (Figure 12, Figure 13). These genes showed a clear temporal succession during biofilm ontogeny that closely mirrored the developmental ordering observed during free-floating rosette assembly (Figure 11, Figure 12). Early stage biofilms were characterized by peak expression of genes encoding flagellar components and type 1 fimbriae (Figure 11bc, Figure 12bc), aligning with their early roles in driving sister-cell folding and stabilizing initial cellular chains. Similarly, the outer membrane adhesin Antigen 43 (Ag43), which facilitates the initial cell-to-cell adhesion required to initiate rosette formation, was highly expressed during these early biofilm stages (Figure 11a, Figure 12a). As development progressed to the mid stages, we observed a transitional shift: curli-associated genes reached their maximum expression levels, reflecting their role in reinforcing mature rosette structures and stabilizing the emerging biofilm matrix (Figure 11bc, Figure 12bc). This cascade culminated in a regulated transition toward stress-associated states, highlighted by the late stage up-regulation of the stationary phase sigma factor *rpoS* (Figure 11e, Figure 12e), a transcriptomic shift that parallel the *rpoS*, dependent maturation and chain morphogenesis phases of the rosette life cycle. Collectively, while the absolute timescales differ between our experimental system of biofilm development and the rosette life cycle (Puri et al., 2022; Puri et al., 2024), the conserved ordering of transcript and protein modules is consistent with deployment of related developmental logic under distinct environmental and temporal regimes (Figure 11, Figure 12).

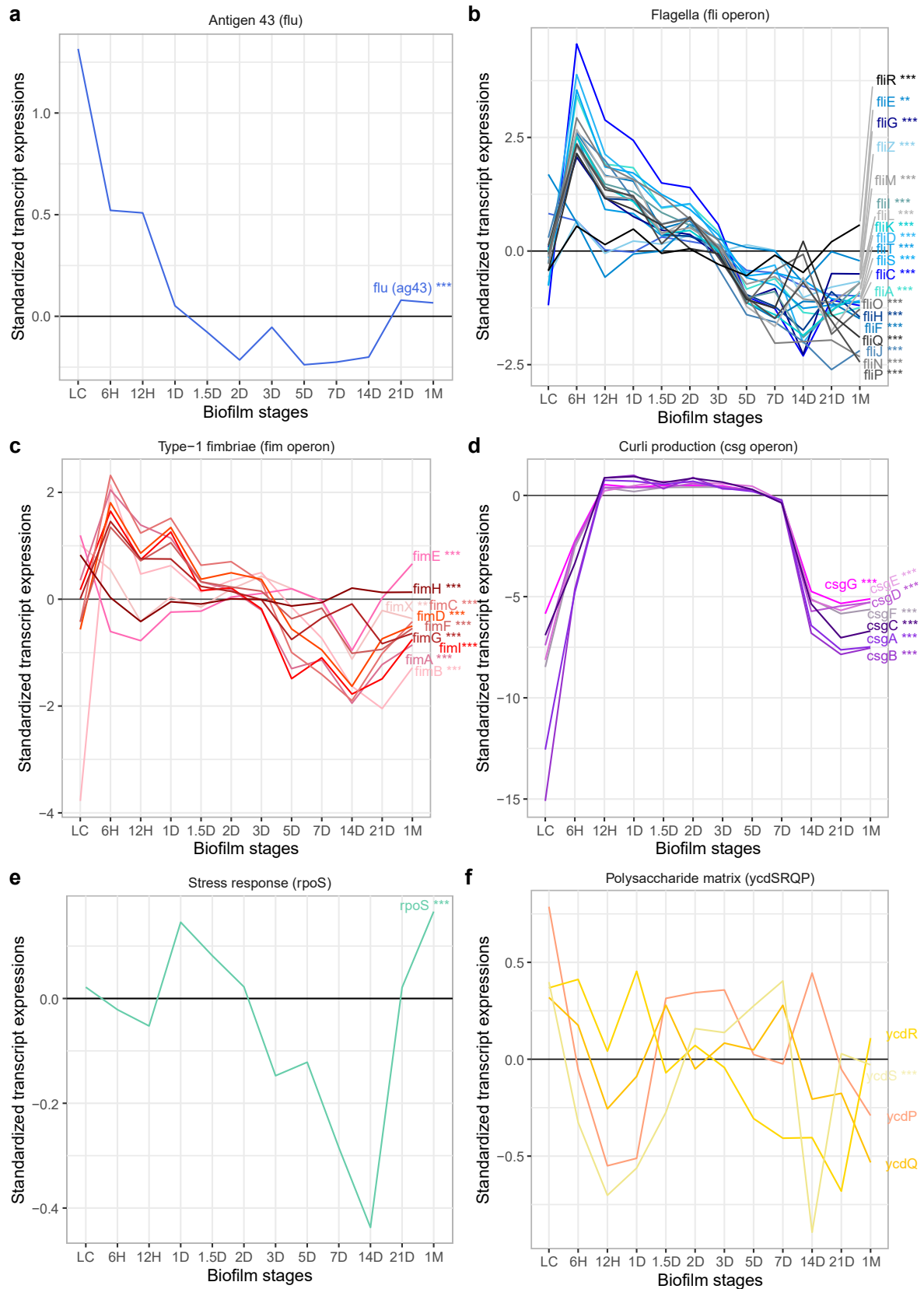


Figure 12. Transcript-level expression dynamics of genes associated with rosette-initiated biofilm development. (a) Antigen 43 (flu) (n = 1), (b) flagella (fli operon) (n = 20), (c) type-1

fimbriae (fim operon) (n = 10), (d) curli production (csg operon) (n = 7), (e) stress response (rpoS) (n = 1), and (f) polysaccharide matrix (ycdSRQP operon) (n = 4). Gene expression profiles were standardized to the same scale: expression values were normalized to the median and log2 transformed, resulting in per-gene standardized expression values. The complete gene list and underlying expression values are provided in Tušar et al. (2026). The black horizontal line denotes the median of standardized transcriptome expression values. Differential expression across all developmental stages was tested for each gene using DESeq2 v1.38.3 (Love et al., 2014) with a likelihood ratio test (LRT). Asterisks indicate statistical significance (* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001). Supplementary dataset is provided in Tušar et al. (2026).

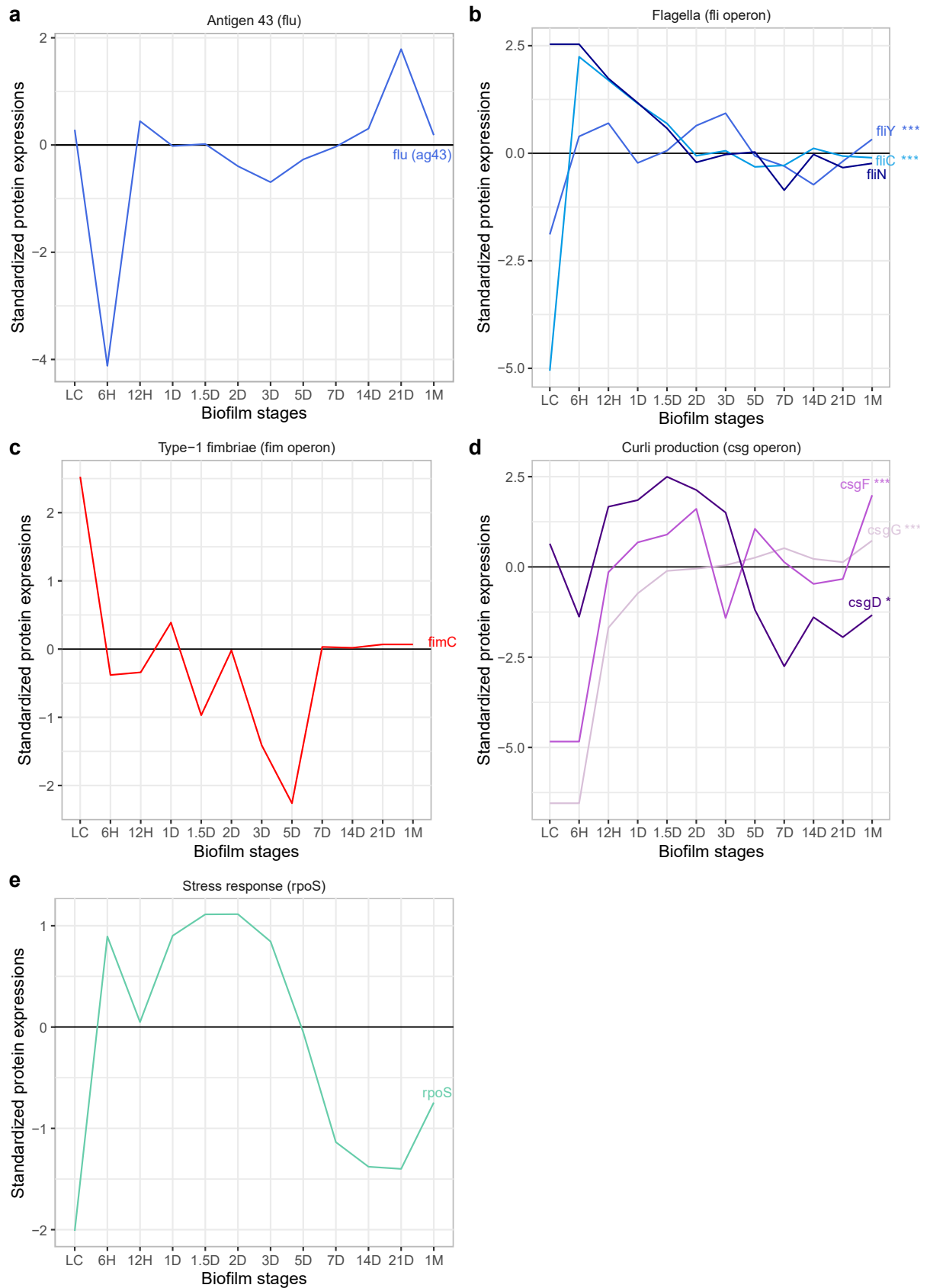


Figure 13. Protein-level expression profiles of genes associated with rosette-initiated biofilm development. (a) Antigen 43 (*flu*) (n = 1), (b) flagella (*fli* operon) (n = 3), (c) type-1 fimbriae (*fim*

operon) (n = 1), (d) curli production (*csg* operon) (n = 3), and (e) stress response (*rpoS*) (n = 1). Proteins encoded by *ycdSRQP* operon (polysaccharide matrix) were not detected in our proteome data. Protein expression profiles were standardized to the same scale: expression values were normalized to the median and log2 transformed, resulting in per-protein standardized profiles. The complete protein list and underlying protein expression values are provided in Tušar et al. (2026). The black horizontal line denotes the median of standardized proteome expression values. Differential expression across all developmental stages was assessed for each protein using DESeq2 v1.38.3 (Love et al., 2014) with a likelihood ratio test (LRT). Asterisks indicate statistical significance (* *p*-value < 0.05, ** *p*-value < 0.01, *** *p*-value < 0.001). Supplementary dataset is provided in Tušar et al. (2026)

We analyzed the expression of individual genes to select representative examples of evolutionary older genes that dominate early biofilm stages and younger genes that increase expression in late stages of biofilm development (Figure 14). In the early development of biofilm (6H – 1D), highly expressed genes that map to the oldest phylostrata (PS1-PS2) and encode functions linked to host colonization and resource acquisition. A prominent example is the *chu* heme transport system (Mey et al., 2021) which is strongly upregulated at the 6H stage (Figure 14a). The *chu* pathway is a well-characterized heme acquisition module in *E. coli*, in which the outer membrane receptor *chuA* recognizes hemoglobin-derived heme and supports iron scavenging under iron limitation (Mey et al., 2021). Notably, recent work shows that *chuA* can rapidly extract heme from hemoglobin prior to import, underscoring the efficiency of this pathway in host-associated environments (Fox et al., 2025). Within the same operon, *chuS* functions as a heme oxygenase, and *chuS*-dependent carbon monoxide production has been linked to modulation of host innate immune responses and changes in gut microbiota composition that favor expansion of *E. coli* strains carrying *chuS* (Maharshak et al., 2015). A second early stage (6H) example of highly expressed old genes is the *neu* operon (PS1–PS2), which contributes to sialic acid (polysialic acid) capsule biosynthesis (Figure 14b). Components of the *neu* gene cluster are required for sialic acid synthesis and capsule formation, and this capsule can act as an anti-phagocytic, host-mimicking barrier that promotes immune evasion (Vann et al., 2004). The strong early expression of *chu* and *neu* operons is therefore consistent with rapid deployment of host-relevant functions during initial biofilm establishment, when iron availability and immune pressures are expected to be key constraints.

Although many old genes (PS1-PS2) peak early in the biofilm development, a subset also

shows elevated expression in late biofilm stages (14D–1M), consistent with the U-shaped relative expression patterns observed for the oldest strata (Figure 7a). The gene *ygfA* (PS1) is an example (Figure 14c), as *ygfA* is involved in the formation of persister cells that may facilitate relapsing infections (Hansen et al., 2008) and is known to be highly expressed in *E. coli* biofilms (Ren et al., 2004). This interpretation aligns with the idea that late stage biofilms increasingly emphasize stress tolerance and phenotypic heterogeneity, potentially including persistence-associated states.

In contrast, the functional landscape of evolutionarily younger genes (PS7–PS9), which are preferentially expressed in late stages, is harder to summarize because majority of them remain poorly annotated (Figure 11). One young gene with a defined role, which maps to the *Escherichia/Shigella* phylostratum (PS7), is *dinQ* (UTI89_C4020; PS7; Figure 14d). A LexA/RecA-regulated SOS gene *dinQ* encodes a small, single-pass transmembrane peptide whose expression is tightly controlled by antisense small RNAs and can affect membrane-associated processes under DNA damage stress (Weel-Sneve et al., 2013).

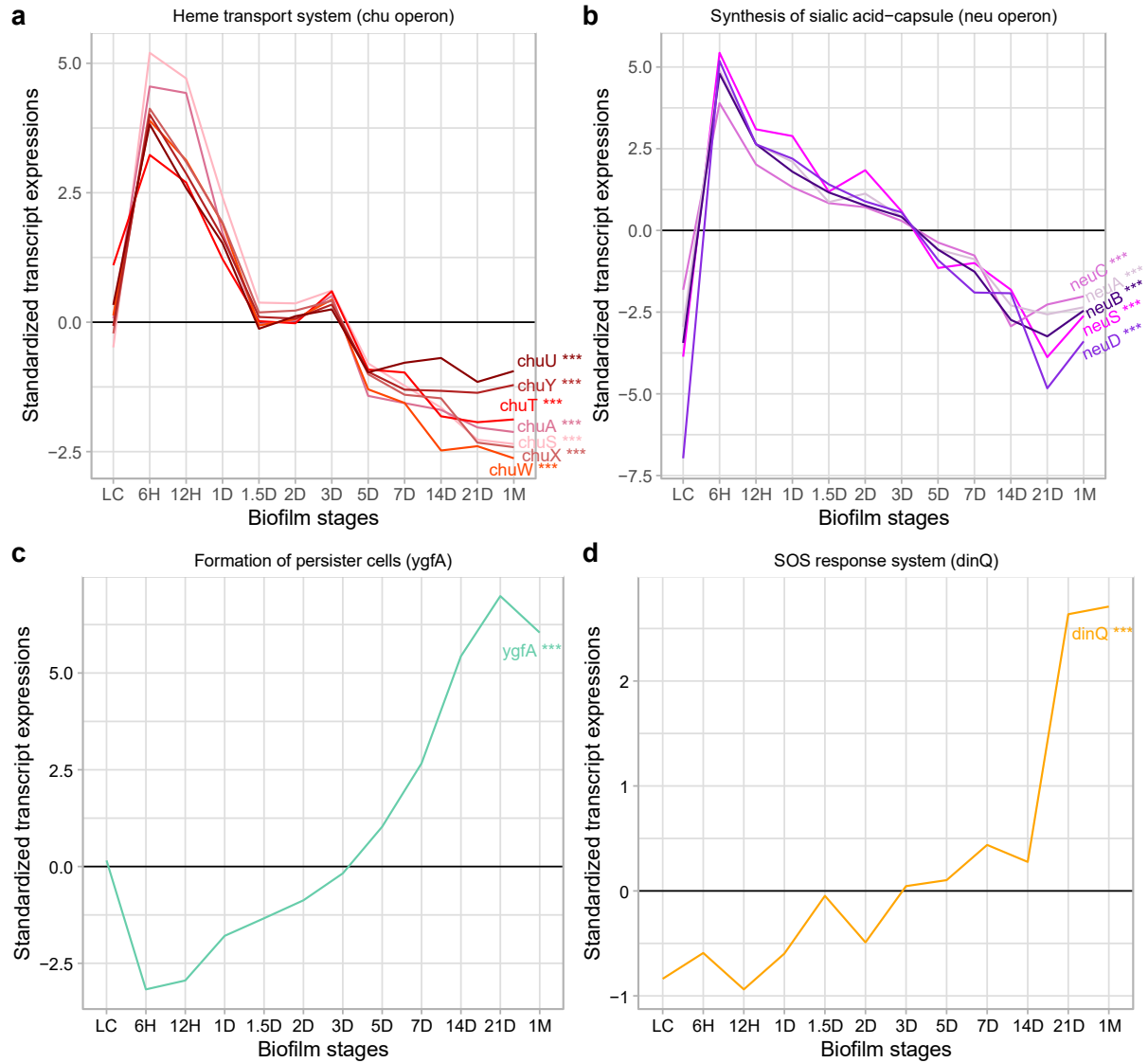


Figure 14. Representative genes with peak expression in early and late biofilm stages. Standardized expression profiles of (a) the *chu* operon, phylostratum 1–2; (b) the *neu* operon, phylostratum 1–2; (c) the *ygfA* gene, phylostratum 1; and (d) the *dinQ* gene, phylostratum 7. To identify genes that contribute strongly to early and late biofilm stages, we examined stage-resolved expression profiles of individual genes and selected representative examples with annotated functions and high expression in either early (6H) or late (21D–1M) biofilm development. Supplementary dataset is provided in Tušar et al. (2026).

To evaluate structural differences between annotated and unannotated proteins, we collected AlphaFold predictions for 4,960 *E. coli* UT189 proteins (Jumper et al., 2021). We compared protein length normalized frequencies of secondary structural elements (helices and sheets) and AlphaFold fold confidence scores between annotated and unannotated proteins (Figure 15). Annotated proteins generally displayed slightly higher frequency of

secondary structures and higher prediction confidence (Appendix C, D). When stratified by evolutionary age, both structural metrics decreased progressively toward young phylostrata in both annotated and unannotated proteins (Figure 15). This pattern indicates that evolutionarily young proteins tend to possess fewer predicted structural elements and lower fold-prediction confidence, consistent with a higher prevalence of intrinsic disordered regions (Wilson et al., 2017) that may explain the challenges surrounding their experimental and computational annotation (Figure 11).

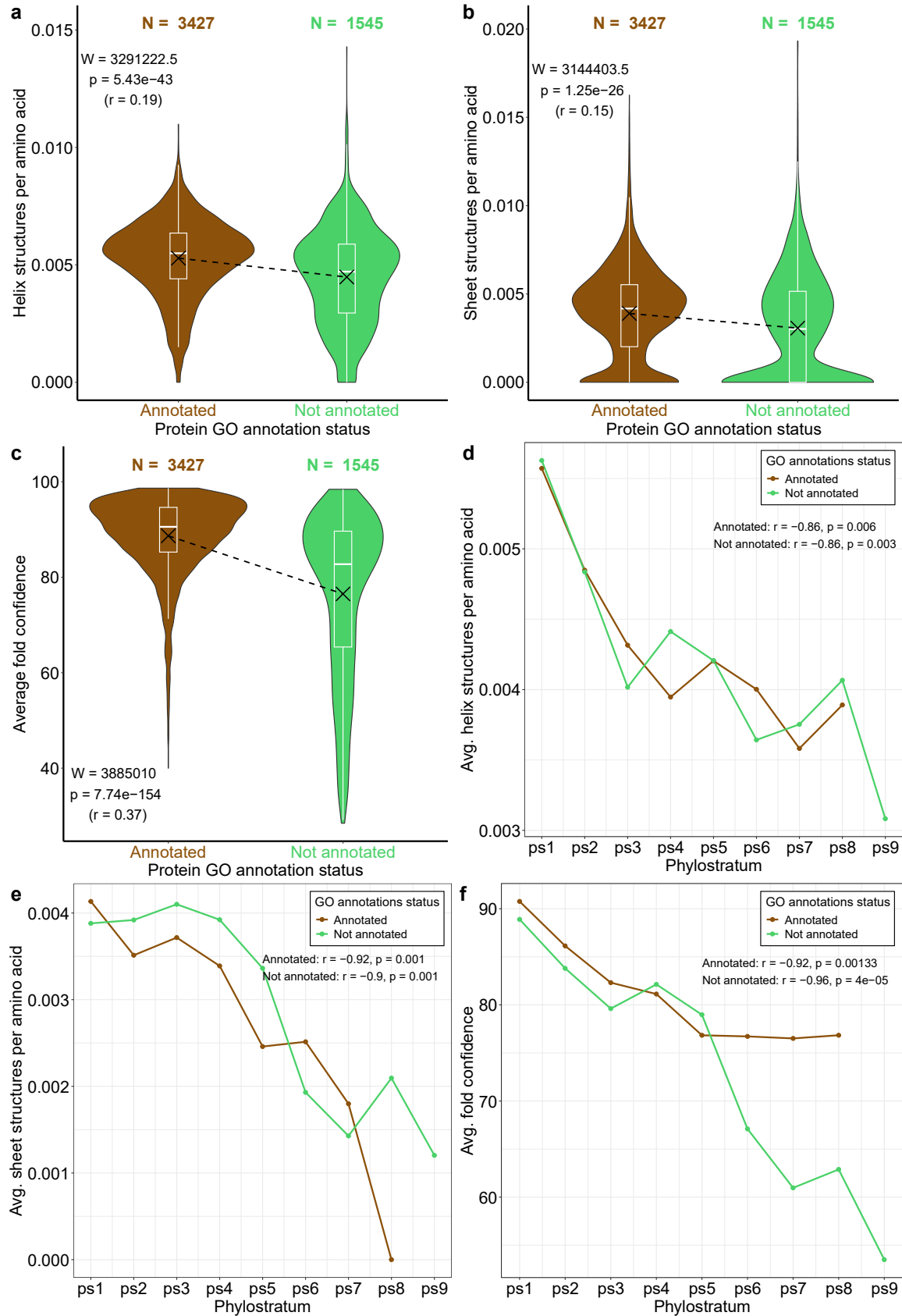


Figure 15. Comparison of structural properties and evolutionary age trends between annotated and unannotated *E. coli* UTI89 proteins. Differences between annotated and

unannotated proteins were assessed for (a) the number of helix structures per protein length, (b) the number of sheet structures per protein length, and (c) average fold confidence. Pairwise statistical significance was assessed via a two-sided Mann–Whitney U test with continuity correction (W, p-value, and effect size r are shown) (Appendix D). X symbols indicate means, bars represent medians, the lower and upper hinges correspond to the first and third quartiles, and whiskers extend to the most extreme values within 1.5 times the interquartile range from the hinges. Violin plots represent the density distribution of data points. Correlations between protein age and (d) the number of helices per protein length, (e) the number of sheets per protein length, and (f) average fold confidence were also calculated separately for annotated and unannotated proteins. Pearson correlation coefficients (d–f) were calculated using the `cor()` function in R. Functional annotations were assigned via eggNOG-mapper, where 'unannotated' designates proteins lacking any Gene Ontology (GO) terms. Supplementary dataset is provided in Tušar et al. (2026).

To better understand the functions linked to the recapitulation pattern in the transcriptome data (Figure 5a), we linked phylostratum-level relative expression profiles to expression profiles of Gene Ontology (GO) functional categories. Specifically, for each phylostratum we compared its overall relative expression profile (Figure 7a) with the relative expression profiles of GO-defined functional subsets within that phylostratum (see Methods). This analysis identified functional categories whose temporal dynamics most closely matched the corresponding phylostratum profile (Figure 16). Functions associated with the oldest phylostratum (PS1; cellular organisms) were dominated by core cellular processes, including basic metabolism, biosynthesis, and transport. In contrast, phylostrata of intermediate evolutionary origin (PS2–PS5; Bacteria to Enterobacterales) were preferentially associated with regulatory functions, homeostasis, stress responses, and biofilm-related processes, including pathways involved in colanic acid biosynthesis (Prigent-Combaret et al., 2000; Hanna et al., 2003). In the youngest phylostrata (PS7–PS9; Enterobacteriaceae to *Escherichia coli* UTI89), the phylostratum-specific relative expression profiles were driven largely by genes lacking functional annotation (Figure 16).

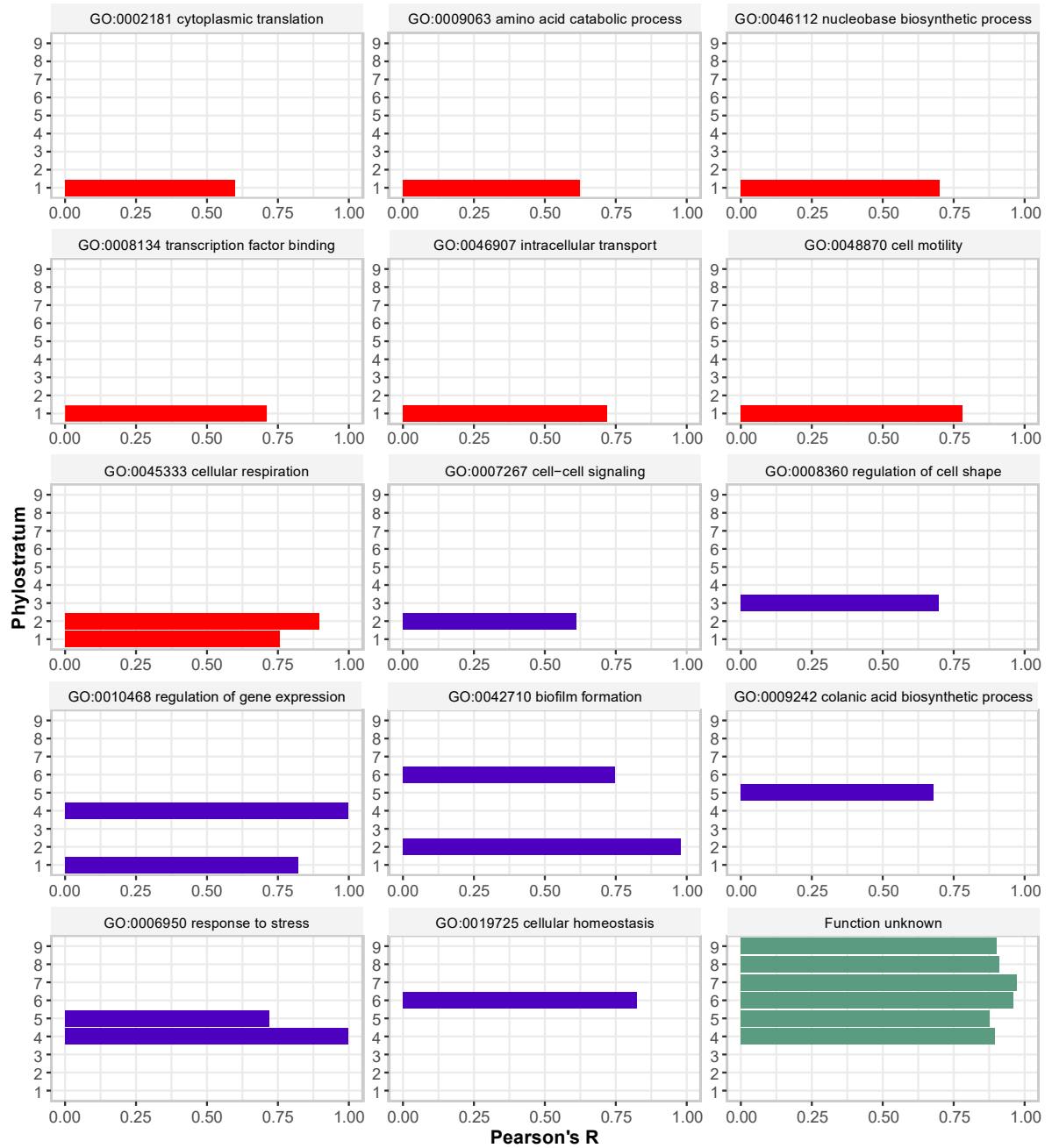


Figure 16. Functional categories associated with phylostratum-specific relative expression profiles. To identify Gene Ontology (GO) categories whose dynamics best matched the relative expression profile of each phylostratum, we calculated Pearson correlations between $RE(func)_{ps}$ and $RE(ps)$ for each functional category (see Methods). Significant Pearson correlation coefficients R (p -value < 0.05) are shown on the x-axis, and phylostrata (PS1–PS9) are shown on the y-axis. GO terms enriched in older phylostrata are highlighted in red, those predominant in intermediate phylostrata in blue, and those associated with younger phylostrata in green. In the youngest phylostrata, “unknown function” showed the strongest correspondence with the observed relative expression profiles. Supplementary dataset is provided in Tušar et al. (2026).

4.4 Spatially resolved transcription and translation in *E. coli* biofilm

In previous chapters, we described developmental regularities in expression patterns using bulk biofilm samples. However, the question remains whether resolved expression imprints, evolutionary patterns, and functional distribution also exists across spatial regions within individual biofilms. To address this question, we recovered region-resolved transcriptomes and proteomes at three developmental stages (Figure 17). Biofilm collection for spatially resolved samples was done for 1-day-old biofilm (1D, early stage), 3-days-old biofilm (3D, mid stage), and 14-days-old biofilm (14D, late stage) (Figure 17). We sampled inner and outer regions from 1D and 3D biofilms, and inner, middle, and outer regions from 14D biofilms (Figure 17a).

Transcriptome data for spatial samples was recovered for 4,989 (96%) and proteome data for 2,355 (45.2%) *E. coli* UTI89 genes. Principal component analysis (PCA) showed distinct transcriptome and proteome profiles across biofilm regions and developmental time points in *E. coli* UTI89 biofilm, with three biological replicates grouped together (Figure 17bd). However, PCA patterns together with all-against-all Pearson correlation analyses (Figure 17ce) indicated that developmental time is a stronger driver of expression variation than spatial position within the biofilm.

Among the genes detected in the spatial samples, 32% (1D), 44% (3D), and 9% (14D) showed significant differential expression between biofilm regions, pointing to spatial organization of transcription within the biofilm. Spatial differentiation between the inner and outer regions was most pronounced in 3D biofilms. Notably, the 3D outer region clustered more closely with both inner and outer regions of 1D biofilms than with the 3D inner region (Figure 17ce). To further characterize this spatial heterogeneity, we performed functional analyses of differentially expressed genes, which revealed a spatially structured distribution of functional categories in *E. coli* UTI89 biofilms (Figure 18, Figure 19, Figure 20).

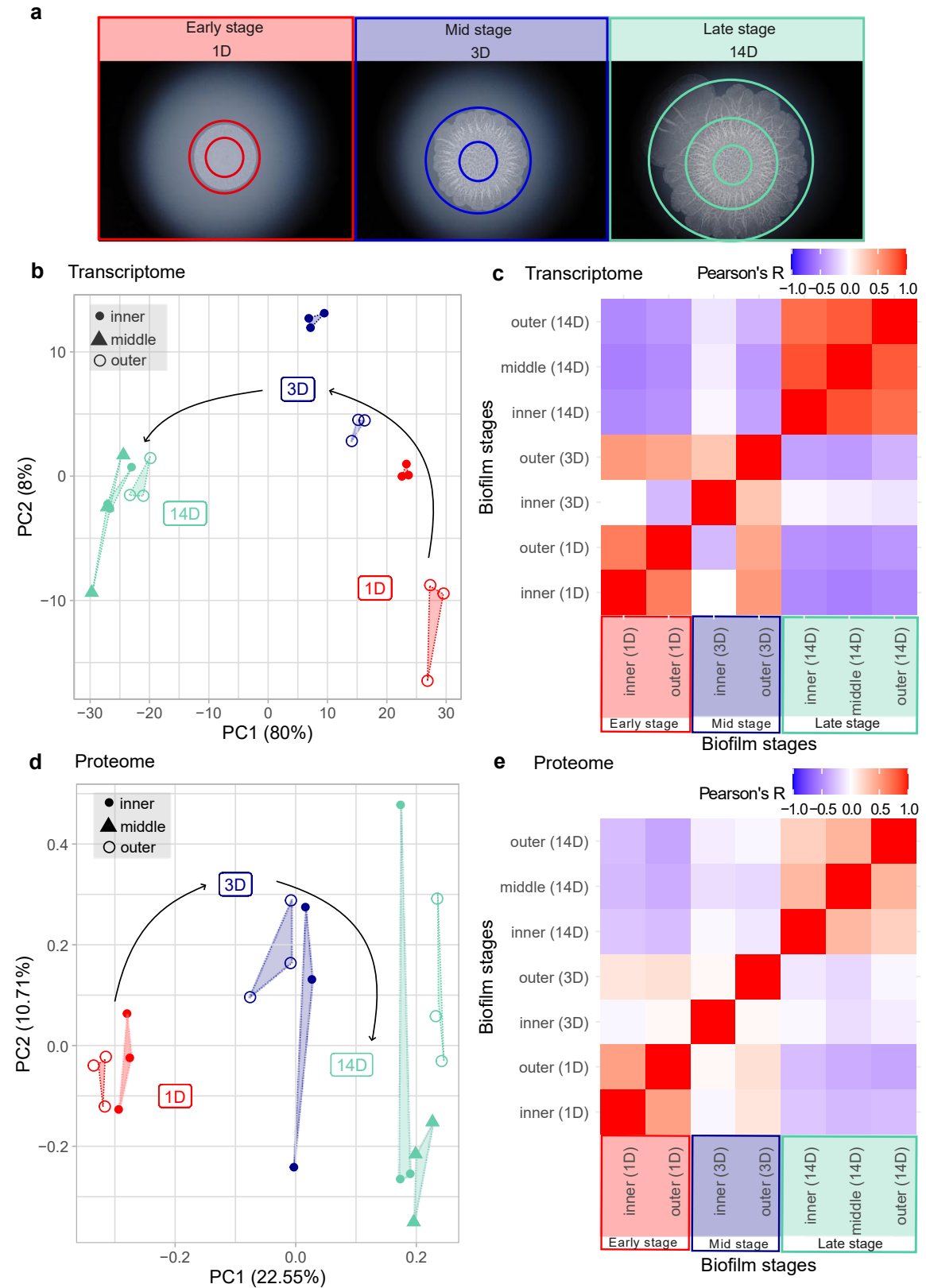


Figure 17. *E. coli* UTI89 biofilm shows spatio-temporally resolved transcription and translation. (a) Representation of biofilm sample collection for 1-day-old (early stage), 3-days-old (mid stage), and 14-days-old (late stage) biofilms. The sampling areas of the inner, middle, and

outer biofilm regions were determined based on biofilm morphology and are indicated by circles. (b, d) Principal component analysis (PCA) of (b) transcriptome data and (d) proteome. Biofilm stages are represented in different colors: 1D (red), 3D (blue), and 14D (green). Symbols denote parts of the biofilm: inner (filled circle), middle (filled triangle), and outer (empty circle). Biological replicates sharing the same color are connected with dotted lines. Black arrows show the consecutive order of biofilm samples in the experimental timeline. (c, e) Pearson's correlation coefficients calculated across biofilm regions and developmental time points in all-against-all comparison for (c) transcriptome and (e) proteome data. Early (1D), mid (3D), and late (14D) stages are marked with different colors on the x-axis. Supplementary dataset is provided in Tušar et al. (2026).

In 1D biofilms, the outer region was enriched for transcripts involved in iron acquisition, ribosome biogenesis and translation, cell motility, and carbohydrate metabolism (Figure 18). This profile is consistent with metabolically active cells at the biofilm periphery, which are likely exposed to fluctuating nutrients and oxygen availability. The prominence of iron uptake functions suggests that iron limitation is already a relevant constraint at this early stage, potentially due to rapid growth and competition for micronutrients. In contrast, the inner region of 1D biofilms showed enrichment for amino acid biosynthetic pathways and stress-response functions (Figure 18), pointing to a more resource-limited or physiologically challenged microenvironment in the biofilm interior, where cells may invest in biosynthetic self-sufficiency and protective responses.

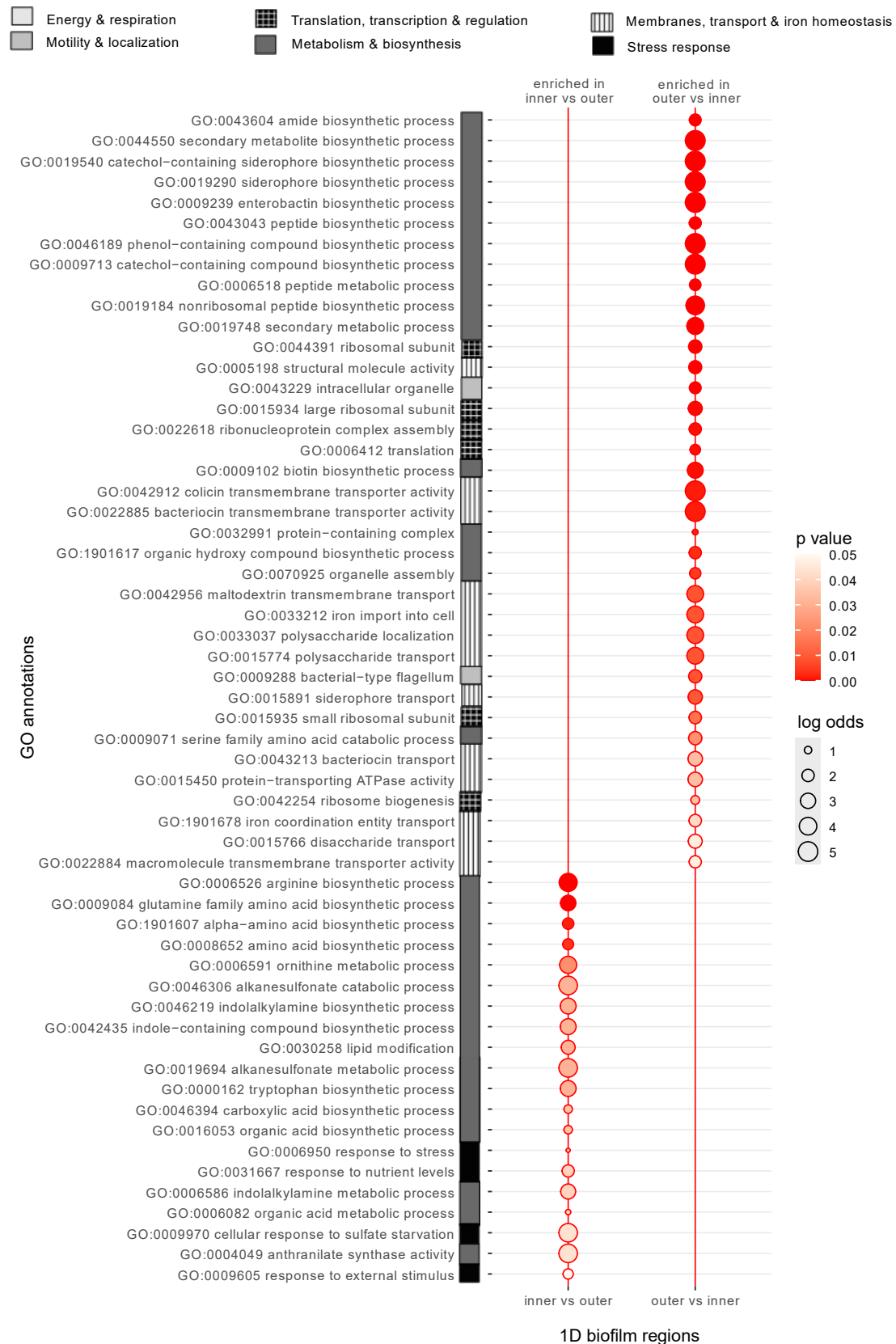


Figure 18. Spatial organization of functional categories in the early stage (1D) biofilm. Shown are enriched GO functional annotations for upregulated genes that are differentially expressed

between regions of 1-day old *E. coli* UTI89 biofilms. Region-specific differentially expressed genes were identified by pairwise DESeq2 comparisons of each 1D region against the other 1D region. Genes were considered to be differentially expressed if the shift in its expression was statistically significant (p -value < 0.05) and filtered for the magnitude of change to be at least two-fold ($FC > 2$). Significance of enrichment was tested using a one-tailed hypergeometric test for upregulated genes, corrected for multiple-testing correction. The effect of functional enrichments is depicted by real log odds (circle sizes), while their significance is indicated by color shades corresponding to p -values. GO terms were grouped into broader functional categories, as shown in the top legend. Supplementary dataset is provided in Tušar et al. (2026).

A comparable spatial division was observed in 3D biofilms (Figure 19). The outer region was enriched for functions related to nucleotide production, amino acid catabolism, and multiple transporter systems, indicating sustained metabolic activity and active uptake of diverse substrates at the biofilm edge. By comparison, the inner region of mid stage (3D) biofilms, also the most transcriptionally distinct sample (Figure 17c), was characterized by enrichment of genes with unknown function and genes implicated in social and environmental interactions (Figure 19). This suggests that, as the biofilm matures, the interior becomes a specialized niche in which intercellular signaling, competitive/cooperative behaviors, and currently uncharacterized processes gain prominence, potentially reflecting increased physiological heterogeneity and localized microenvironments within the biofilm.

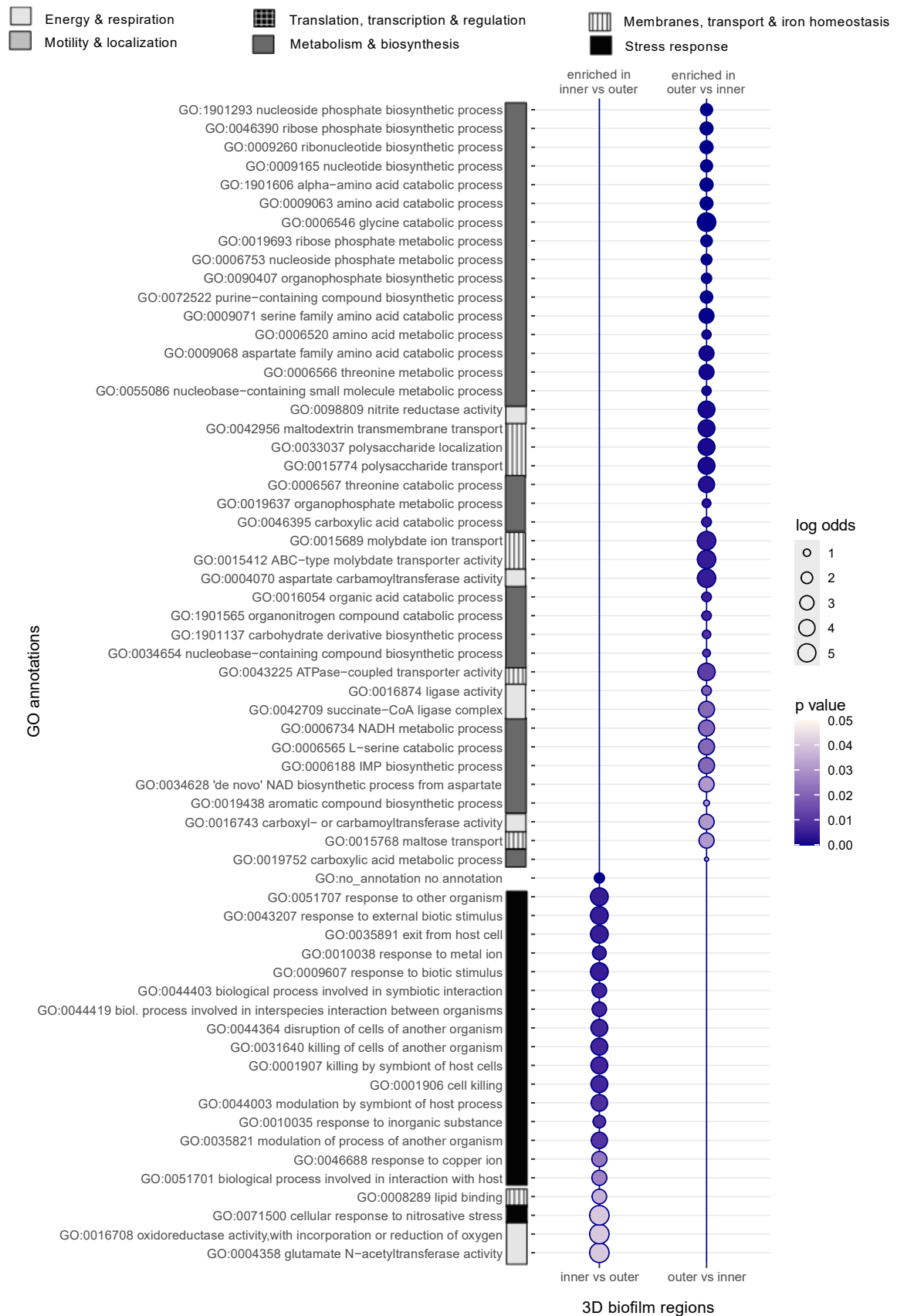


Figure 19. Spatial organization of functional categories in the mid stage (3D) biofilm. Shown are enriched GO functional annotations for upregulated genes that are differentially expressed between regions of 3-day old *E. coli* UTI89 biofilms. Region-specific differentially expressed genes were identified by pairwise DESeq2 comparisons of each 3D region against the other 3D region. Genes were considered to be differentially expressed if the shift in its expression was statistically significant (p -value < 0.05) and filtered for the magnitude of change to be at least two-fold (FC > 2). Significance of enrichment was tested using a one-tailed hypergeometric test for upregulated genes, corrected for multiple-testing correction. The effect of functional enrichments is depicted by real log odds (circle sizes), while their significance is indicated by color shades corresponding to p -values. GO terms were grouped into broader functional categories, as shown in the top legend. Supplementary dataset is provided in Tušar et al. (2026).

In late stage (14D) biofilms, the outer region was enriched for functions linked to polysaccharide metabolism, biofilm formation, cellular pH homeostasis, and protein folding, consistent with active matrix production and stress adaptation at the biofilm periphery (Figure 20). Relative to the outer region, the middle region showed higher transporter-related activity, whereas the inner region was dominated by amino acid catabolic functions, suggesting distinct metabolic strategies across biofilm regions.

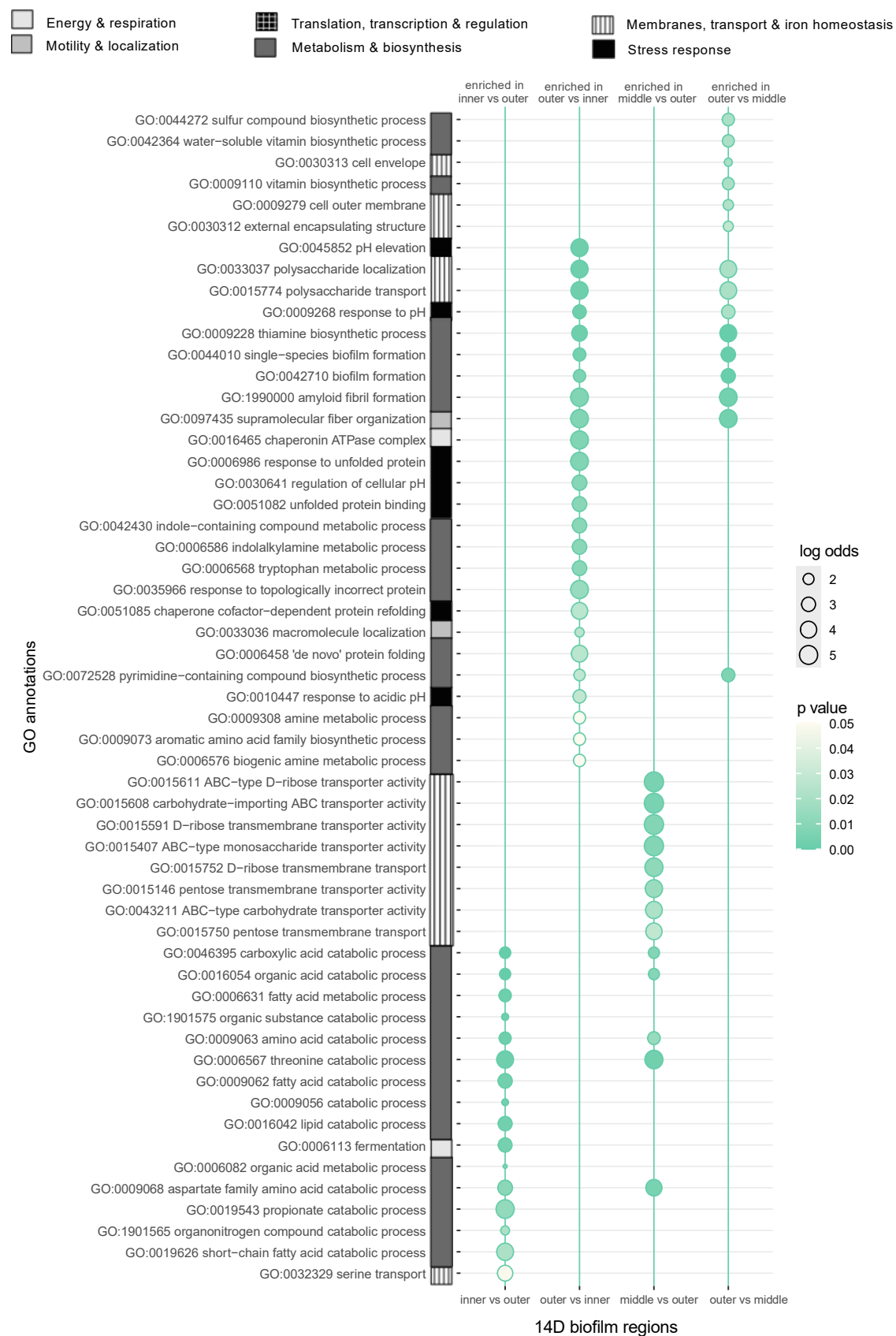


Figure 20. Spatial organization of functional categories in the late stage (14D) biofilm. Shown are enriched GO functional annotations for upregulated genes that are differentially expressed

between regions of 14-day old *E. coli* UTI89 biofilms. Region-specific differentially expressed genes were identified by pairwise DESeq2 comparisons of each 14D region against the other 14D region. Genes were considered to be differentially expressed if the shift in its expression was statistically significant ($p\text{-value} < 0.05$) and filtered for the magnitude of change to be at least two-fold ($FC > 2$). Significance of enrichment was tested using a one-tailed hypergeometric test for upregulated genes, corrected for multiple-testing correction. The effect of functional enrichments is depicted by real log odds (circle sizes), while their significance is indicated by color shades corresponding to p -values. GO terms were grouped into broader functional categories, as shown in the top legend. Supplementary dataset is provided in Tušar et al. (2026).

Proteome-based enrichment analyses broadly mirrored these spatial trends, but the fraction of proteins showing significant regional differences was substantially lower than for transcripts (6% at 1D, 7% at 3D, and 10% at 14D). This reduced signal likely reflects the lower proteome coverage and narrower dynamic range of mass spectrometry compared with RNA sequencing (Figure 21).

These results indicate pronounced functional heterogeneity between the inner and outer regions of biofilms during development. At the same time, several metabolic programs are spatially dynamic rather than fixed, with their regional enrichment shifting across biofilm ontogeny. For example, amino acid biosynthesis is enriched in the inner region at the early (1D) stage, whereas amino acid catabolism is most prominent in the outer region at the mid (3D) stage. By the late (14D) stage, enrichment of amino acid catabolic functions shifts toward the inner region.

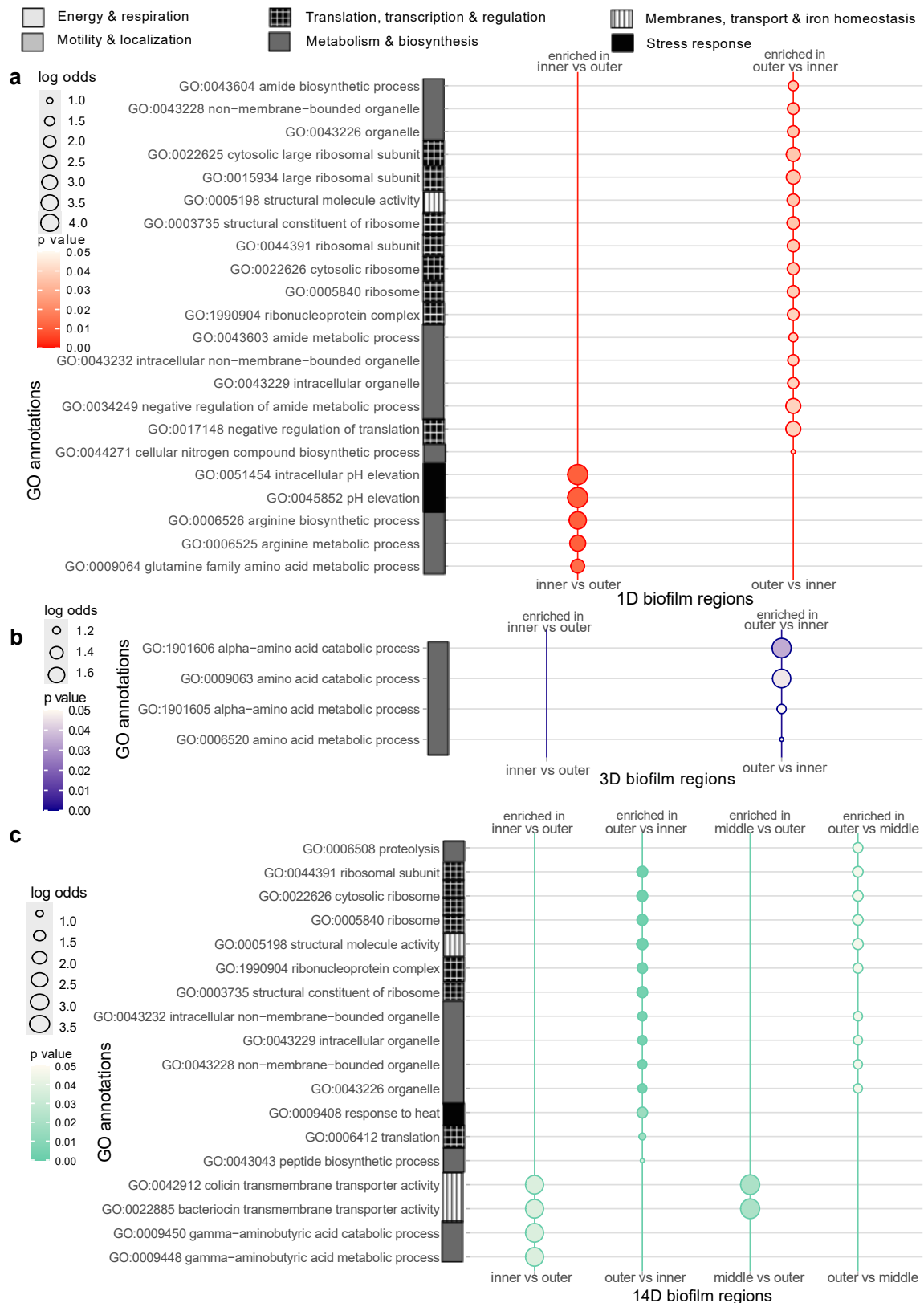


Figure 21. Spatial organization of functional categories in the proteome data of early stage (1D), mid stage (3D), and late stage (14D) *E. coli* biofilm development. Shown are enriched GO functional annotations for upregulated proteins that are differentially expressed between regions of

E. coli UTI89 biofilms. Region-specific differentially expressed genes were identified by pairwise DESeq2 comparisons of each 1D region against the other 1D region. Proteins were considered to be differentially expressed if the shift in its expression was statistically significant ($p\text{-value} < 0.05$) and filtered for the magnitude of change to be at least two-fold ($FC > 2$). Significance of enrichment was tested using a one-tailed hypergeometric test for upregulated proteins, corrected for multiple-testing correction (Supplementary dataset is provided in Tušar et al. (2026)). The effect of functional enrichments is depicted by real log odds (circle sizes), while their significance is indicated by color shades corresponding to p-values. GO terms were grouped into broader functional categories, as shown in the top legend. Supplementary dataset is provided in Tušar et al. (2026).

An unexpected difference emerged when we compared the spatial organization of arginine metabolism in our biofilms with results reported for a non-pathogenic *E. coli* strain (Wang et al., 2023). It was observed that genes involved in arginine biosynthesis were enriched at the biofilm periphery, whereas genes associated with arginine utilization were predominantly expressed in the interior (Wang et al., 2023). In contrast, we observed the opposite spatial pattern during early (1D) and mid (3D) stage development, with arginine biosynthesis and utilization showing reversed regional enrichment (Figure 22). By the late (14D) stage, this spatial separation was no longer evident, suggesting that arginine-associated functions become more spatially diffuse as the biofilm matures (Figure 22).

Given that the spatial enrichment patterns we detected for the majority of functional categories align closely with the patterns reported by Wang et al., (2023), the specific divergence in arginine metabolism is likely attributable to differences in experimental conditions (e.g., media composition, oxygen and nutrient gradients, surface properties) and strain-specific regulatory architecture.

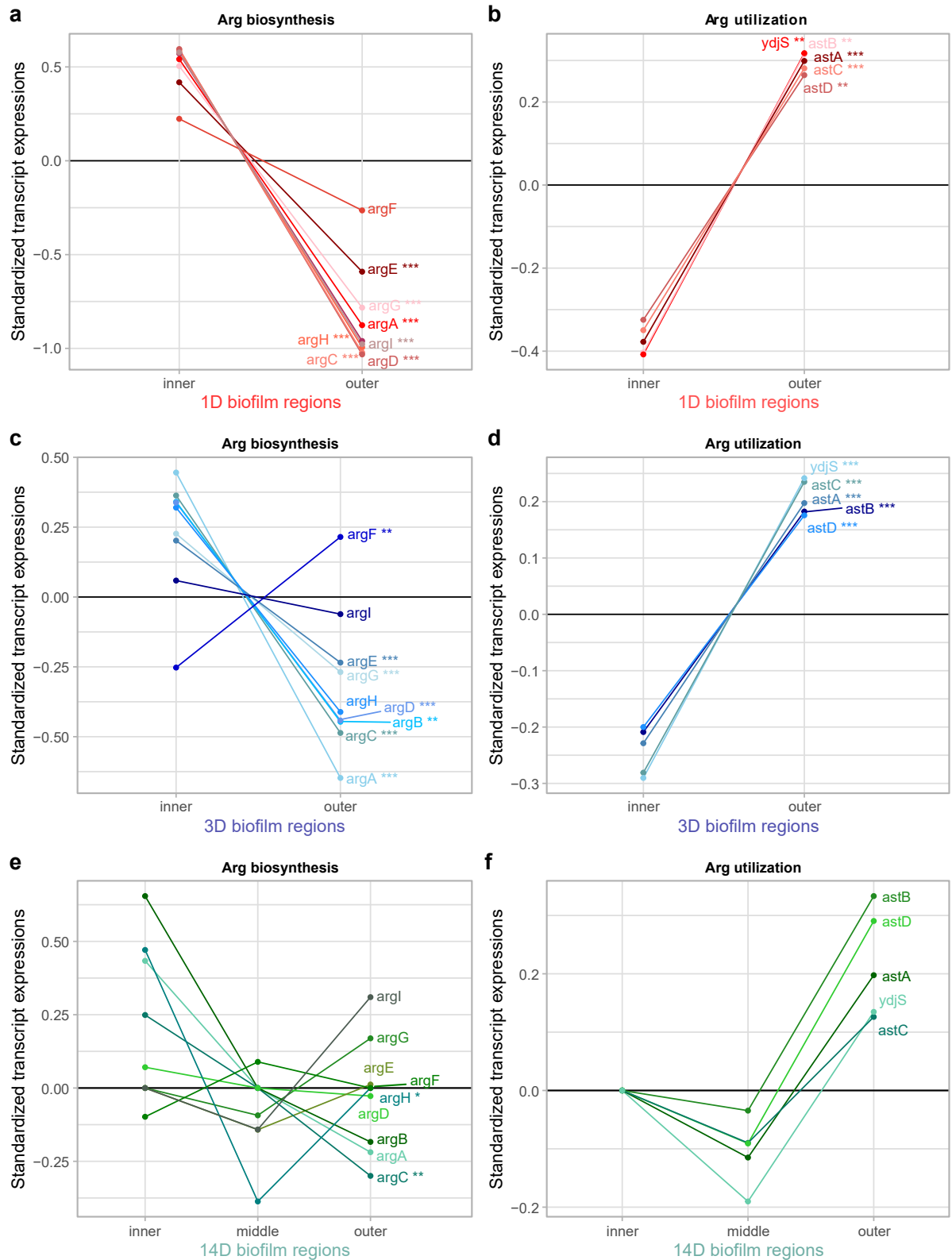


Figure 22. Standardized expression profiles for arginine biosynthesis and utilization genes in across spatial regions and developmental stages of *E. coli* UTI89 biofilms. Panels show (a,b) early stage (1D), (c,d) mid stage (3D), and (e,f) late stage (14D) biofilms. Arginine biosynthesis

genes are shown in (a,c,e), and arginine utilization genes in (b,d,f). To standardize gene expression profiles on a common scale, expression values were normalized to the median and log2 transformed, resulting in per-gene standardized expression values. Differential expression across developmental stages and spatial regions was tested using DESeq2 v1.38.3 (Love et al., 2014) with a likelihood ratio test (LRT). Statistically significant results are indicated with asterisks (* p -value < 0.05, ** p -value < 0.01, *** p -value < 0.001). Supplementary dataset is provided in Tušar et al. (2026).

4.5 Evolutionary imprints in the *E. coli* biofilm regions

Identifying a recapitulation pattern during the biofilm ontogeny of uropathogenic *E. coli*, similar to evolutionary patterns reported for non-pathogenic *B. subtilis* biofilm (Futo et al., 2021), motivated us to further investigate whether evolutionary imprinting is also spatially structured within individual biofilms. Based on the observations that unannotated genes are biased toward younger evolutionary origins (Figure 11) and display marked spatial enrichment in the inner regions of 3-day-old biofilms (Figure 19), we reasoned that the spatial compartmentalization of a biofilm reflects localized variations in macroevolutionary age profiles.

To test this, we compared TAI and PAI values across different regions of early (1D), mid (3D), and late (14D) stage biofilm (Figure 23). In 1D (early) biofilms, the inner region showed a non-significant tendency toward a younger transcriptome and proteome in comparison to the outer region (Figure 23ad). This regional contrast became stronger and statistically significant at 3D (mid stage), indicating a younger molecular profile in the inner region of more mature biofilms (Figure 23be). At 14D (late stage), the TAI pattern reversed, with the inner region exhibiting an older transcriptome and the outer region a younger transcriptome (Figure 23c). This reversal was not mirrored at the proteome level (Figure 23f), likely reflecting the lower coverage and narrower dynamic range of the proteomics dataset.

In summary, these results show that macroevolutionary patterns are both temporally structured across biofilm ontogeny and spatially partitioned within biofilm regions, with this spatial stratification manifesting most prominently in mid stage (3D) biofilms (Figure 23be).

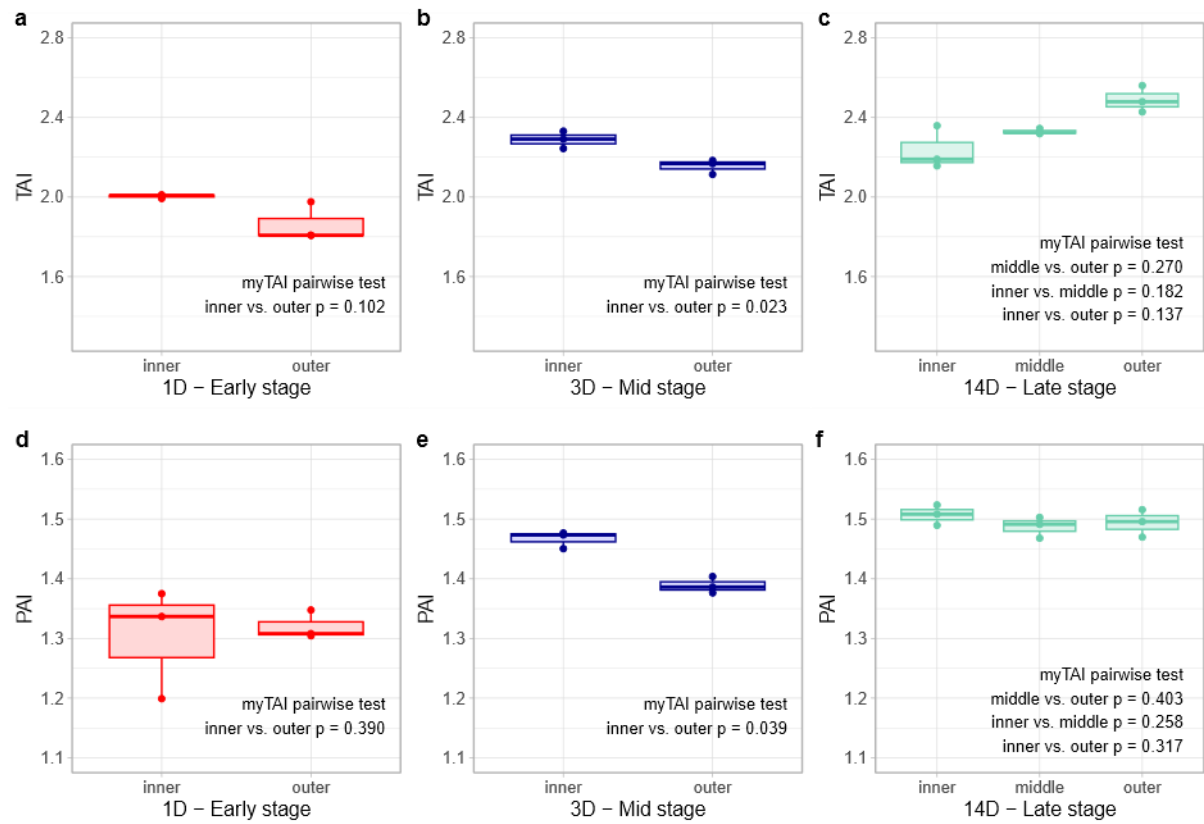


Figure 23. Evolutionary imprints in the *E. coli* biofilm regions. (a–c) Transcriptome age index (TAI) for inner versus outer regions of (a) 1D, (b) 3D, and (c) 14D biofilms. (d–f) Proteome age index (PAI) for inner versus outer regions of (d) 1D, (e) 3D, and (f) 14D biofilms. The significance of TAI and PAI differences between biofilm regions was assessed using the myTAI pairwise test. Three biological replicates per sample are depicted as individual dots in the boxplots. Three biological replicates were analyzed per sample; individual replicates are shown as points on the boxplots. Supplementary dataset is provided in Tušar et al. (2026).

5 DISCUSSION

Our findings show that biofilm development in uropathogenic *Escherichia coli* UTI89 is not an unorganized, continuous cell accumulation, but rather a structured multicellular process defined by distinct temporal and spatial expression programs. By integrating transcriptomic, proteomic, phylotranscriptomic, and phyloproteomic analyses, we revealed macroevolutionary imprints in *E. coli* biofilm that closely parallel those previously identified in *Bacillus subtilis* (Futo et al., 2021).

The presence of these shared ontogeny–phylogeny correlations across evolutionary divides implies that such patterns reflect a fundamental, generalized principle of bacterial biofilm development rather than a species-specific artifact. Furthermore, the pronounced spatio-temporal organization of *E. coli* biofilms reveals striking system-level analogies with developmental processes in multicellular eukaryotes (Domazet-Lošo & Tautz, 2010b; Quint et al., 2012; Koska et al., 2025). Despite their independent evolutionary origins and vast phylogenetic divergence across the tree of life, both bacterial biofilms and eukaryotic organisms exhibit punctuated expression dynamics, spatial functional stratification, and coordinated ontogenetic trajectories (Domazet-Lošo & Tautz, 2010a, Quint et al., 2012; Koska et al., 2025). These cross-kingdom parallels likely stem from shared ecological constraints. Diverse biological systems frequently encounter similar selective pressures that, when acting on genetic networks and cellular machinery, drive convergent adaptations. In support of this concept, the sequential functional transitions observed across *E. coli* UTI89 biofilm maturation mirror the developmental shifts documented in *B. subtilis* (Futo et al., 2021; Puri et al., 2023, 2024). This functional conservation suggests that the transition to a collective, biofilm-bound state has repeatedly evolved under common environmental drivers (O’Toole et al., 2000; Stoodley et al., 2002; Lappin-Scott et al., 2014; Sauer et al., 2022).

The correlation between transcriptome and proteome data was relatively weak across biofilm developmental stages (Figures 2b, Appendix 1). Technical differences in RNA and protein extraction from biofilm samples likely contributed to poor data correlation (see Methods and Materials). Biologically, weak correlation also reflects fundamental differences in RNA and protein molecule stability. Bacteria rapidly reconfigure their transcriptome to cope with microenvironmental shifts, whereas proteomic adjustments occur on a slower, more conservative timescale (Bathke et al., 2019). These discrepancies

did not confound our broader conclusions; phylostratigraphy patterns remained consistent across both transcriptome and proteome datasets, demonstrating that evolutionary imprints are deeply ingrained in biofilm ontogeny regardless of regulatory lags.

5.1 Stage-organized transcription and translation during *E. coli* biofilm development

Based on transcriptome and proteome data, we first established that *E. coli* biofilm growth follows a stage-organized developmental process rather than a continuous and uniform temporal progression (Figure 1). This provides an essential foundation for interpreting biofilm formation as a regulated multicellular process. Similar temporal organization was previously observed in developing *Bacillus subtilis* biofilms, suggesting that stage-structured biofilm ontogeny is not restricted to a single bacterial species or strain (Futo et al., 2021).

The strongest transcriptional shift occurred during the transition from liquid culture to early biofilm formation, when 28% of *E. coli* genes were significantly upregulated with a fold change greater than 2 (Figure 3). This transition marks the shift from a planktonic lifestyle to a surface-attached community and supports previous observations that bacteria undergo major transcriptional reprogramming during the onset of biofilm formation (O'Toole et al., 2000; Futo et al., 2021). Thus, the earliest stages of biofilm development are not a passive consequence of growth on a surface but reflect an active and coordinated reprogramming of cellular state.

Functional enrichment analysis further supports the interpretation of biofilm growth as a stage-organized process (Figure 9, Figure 10). Distinct functional categories were enriched at different developmental time points, indicating that successive phases of biofilm ontogeny are associated with specific biological programs (Figure 9, Figure 10). This stage-specific distribution of functions resembles developmental organization observed in both microbial biofilms and animal embryos (Monds & O'Toole, 2009; Levin et al., 2012; Yanai, 2018; Futo et al., 2021). During the early stages of biofilm growth, from 6 hours to 1 day, enriched functions were mainly associated with active growth and initial biofilm formation. These included biosynthetic processes, energy generation, ribosomal synthesis, iron homeostasis, and cell motility. In the mid stages, from 1.5 to 3 days, enrichment shifted toward functions related to cell projection and bacterial-type flagella, suggesting active expansion of the biofilm across the agar surface. This is consistent with earlier work

showing that flagella, motility, and chemotaxis contribute substantially to biofilm formation (Pratt & Kolter, 1998). The mid stages, therefore, appear to represent a critical developmental phase during which the biofilm actively expands and reorganizes. At mid-late and late stages, from 5 days to 1 month, enriched functions shifted toward regulation of metabolic processes, post-replication repair, and the SOS response. This transition suggests that mature biofilms increasingly experience physiological stress, nutrient limitation, and the need for repair and persistence mechanisms. Such a shift from growth-associated processes to stress-response functions is consistent with the broader biofilm life cycle, in which starvation responses and persistence-related adaptations become increasingly important as the bacterial community matures (O'Toole et al., 2000). The presence of both early growth-associated programs and late stress-associated programs indicates that the experimental design captured a broad developmental trajectory of biofilm formation.

Defining a developmental process strictly as a genetically encoded process involving irreversible commitment and checkpoint-controlled progression is too narrow for biofilm biology. Such concepts were largely derived from specialized bacterial developmental systems, such as fruiting-body formation in *Myxococcus*, which is triggered by starvation and follows a highly regulated developmental program (Monds & O'Toole, 2009; Zusman et al., 2007). Comparable programs also occur in spore-forming bacteria such as *Bacillus* (Tocheva et al., 2016; Riley et al., 2020; Futo et al., 2021), but these represent only one form of bacterial multicellular development. Biofilm formation occurs under a broader range of environmental conditions and is more flexible than these specialized developmental programs. For this reason, Bonner's definition of multicellular development, which emphasizes polarity, heritable pattern formation, and cell differentiation with division of labor, better captures the biological reality of biofilm growth (Bonner, 2001). The stage-specific transcriptional, translational, and functional changes observed in *E. coli* UTI89 biofilms are consistent with this broader developmental framework.

Therefore, the stage-organized expression patterns described here provide the developmental basis for the following evolutionary analyses. If biofilm formation is a genuine multicellular developmental process, then its temporal organization may also carry macroevolutionary signatures. By mapping evolutionary gene age onto successive stages of biofilm ontogeny, phylostratigraphic analysis can reveal developmental logic that is not

apparent from expression data alone, and may help identify principles shared across diverse biofilm-forming bacteria (Monds & O'Toole, 2009; Domazet-Lošo & Tautz, 2010b; Quint et al., 2012; Futo et al., 2021; Koska et al., 2025).

5.2 Phylogeny-ontogeny correlation in *E. coli* reveals recapitulation pattern

The evolutionary patterns observed during *E. coli* UTI89 biofilm development provide further support for the view that bacterial biofilms follow a structured developmental logic. Recapitulation was the first proposed model of ontogeny–phylogeny correlation (Abzhanov, 2013). However, most current evidence from eukaryotic developmental systems supports the molecular hourglass model, in which mid-development represents the most evolutionarily conserved phase, whereas early and late developmental stages show greater contributions of younger and more divergent genes (Domazet-Lošo & Tautz, 2010a; Kalinka et al., 2010; Irie & Kuratani, 2011; Kalinka & Tomancak, 2012; Quint et al., 2012; Cheng et al., 2015; Uesaka et al., 2022; Lotharukpong et al., 2024; Koska et al., 2025).

Both *B. subtilis* biofilms and *E. coli* UTI89 biofilms exhibit a recapitulation model, characterized by the progressive expression of evolutionarily younger and more divergent genes during later stages of biofilm development (Figure 5; Futo et al., 2021). The recovery of similar ontogeny–phylogeny correlations in these two bacterial systems is particularly notable because they are ecologically and evolutionarily distinct. *B. subtilis* is a free-living, spore-forming, soil-dwelling member of the Firmicutes phylum, whereas *E. coli* UTI89 is a uropathogenic, non-sporulating member of the Proteobacteria phylum with a saprophytic/parasitic lifestyle (Galperin et al., 2022; Yu et al., 2021). Despite these differences, both species display stage-organized biofilm development and comparable evolutionary expression patterns, suggesting that common developmental and evolutionary principles may underline biofilm formation across divergent bacterial lineages. This interpretation is consistent with the concept of a broadly conserved biofilm developmental strategy (Stoodley et al., 2002; Hall-Stoodley et al., 2004; Lappin-Scott et al., 2014). However, whether these patterns reflect an ancestral feature of bacterial multicellularity or independently evolved solutions to similar ecological challenges remains unresolved. Addressing this question will require comparative analyses across a wider range of biofilm-forming bacteria that differ in ecology, phylogeny, lifestyle, and developmental complexity.

The difference between eukaryotic hourglass patterns and bacterial recapitulation patterns may reflect fundamental differences in the architecture of multicellularity. In complex eukaryotes, development often involves irreversible cell-fate commitment, integrated three-dimensional morphogenesis, and strong constraints during mid-development. By contrast, bacterial biofilms represent a more plastic and environmentally responsive form of multicellularity. Differentiation within biofilms is often reversible, strongly influenced by local environmental gradients, and shaped by changes in nutrient availability, waste accumulation, oxygen limitation, and stress exposure. Such conditions may favor a recapitulation pattern, in which early stages rely primarily on ancient core cellular functions, whereas later stages increasingly activate lineage-specific and environmentally responsive genes. This explanation is supported by the functional transitions observed during *E. coli* UTI89 biofilm development. Early biofilm stages are dominated by core metabolic functions, exhibiting an evolutionarily older and more conserved transcriptome and proteome profile. As the biofilm matures, nutrient limitation and ecological stress induce expression shifts toward bacteria-specific and eventually *Escherichia*-specific genes, many of which remain functionally uncharacterized. This transition from general cellular processes to increasingly lineage-specific functions may explain why the evolutionary age of the expressed transcriptome decreases over biofilm ontogeny.

5.3 Evolutionary imprints in the *E. coli* biofilm regions

Previous studies of spatial gene expression in biofilms have primarily focused on single developmental time points (Wang et al., 2023; Diaz-Pascual et al., 2021; Beebout et al., 2019). In contrast, our analysis captures spatial organization across multiple stages of *Escherichia coli* UTI89 biofilm development, revealing that regional expression patterns are not fixed but change over time. Cells occupying different biofilm regions appear to shift their functional and developmental roles as the macrocolony biofilm matures. One notable example is the preferential expression of evolutionarily younger biofilm-associated genes in the inner region of mid stage biofilms, a pattern that is not observed in early or late biofilms (Figure 23). These findings indicate that the spatial distribution of biofilm-related functions is developmentally dynamic, consistent with previous reports of morphologically distinct subpopulations and functional stratification within biofilm communities (Beebout et al., 2019; Heacock-Kang et al., 2017; Yannarell et al., 2023).

While temporal progression serves as the primary driver of biofilm expression profiles,

dissecting the biofilm into distinct spatial regions (inner, middle, and outer zones) adds an additional developmental resolution. Biofilm growth in *Escherichia coli* UTI89 is characterized by polarity, discrete regional functionalities, and localized gene expression along the center-to-periphery axis. Rather than behaving as a homogenous cell mass, the macrocolony biofilm organizes into dynamic subpopulations with distinct transcriptomic and proteomic profiles across early (1-day), mid (3-day), and late (14-day) stages (Figure 17). Such spatial compartmentalization likely contributed to the coordination of stress responses, metabolic division of labor, and biofilm-associated functions across macrocolony microenvironments. Interestingly, cells situated in the outer region of mid stage (3-day-old) biofilms exhibit transcriptome profiles remarkably similar to those of young (1-day-old) biofilms (Figure 17). This suggests that bacteria at the expanding periphery retain features of an active colonization state, including processes associated with surface expansion, motility, and matrix production. In this sense, the outer edge of a more mature biofilm may preserve developmental characteristics typical of younger biofilm stages, while the inner regions undergo a different functional trajectory.

An additional spatial pattern was uncovered when our results were compared with those reported for non-pathogenic *E. coli* biofilms. Wang et al. (2023) found that genes involved in arginine biosynthesis were more highly expressed at the biofilm periphery, whereas genes associated with arginine utilization were primarily expressed in the interior. In contrast, we observed the opposite pattern during the early and mid stages of *E. coli* UTI89 biofilm development (Figure 22). By the late stage, this spatial separation was no longer evident, suggesting that arginine metabolism becomes more diffusely distributed as the biofilm matures. Since our enrichment results for other functional categories are broadly consistent with those of Wang et al. (2023), this discrepancy is likely explained by differences in experimental conditions, or strain-specific biology.

Biofilm developmental dynamics differ from the logic of many eukaryotic developmental systems. In animal development, cell fate specification often leads to terminal differentiation and the establishment of stable body plans (Tomancak et al., 2007). By contrast, spatial differentiation in bacterial biofilms appears to remain more plastic and environmentally responsive. Unlike animal embryonic stem cells, which undergo irreversible lineage commitment (Gunne-Braden et al., 2020), bacterial subpopulations can respond rapidly to local changes in nutrient availability, waste accumulation, and other microenvironmental conditions. Consequently, spatial organization in biofilms is dynamic

rather than fixed, producing a less rigid but highly adaptive form of multicellular organization.

Differences in transcriptome age indices (TAI) across spatial zones uncover distinct evolutionary signatures that emerge during biofilm development (Figure 23). In early biofilms, the inner and outer regions show little or no divergence in transcriptome age, suggesting that young macrocolonies have not yet formed evolutionarily distinct subpopulations. By the mid stage, spatial evolutionary stratification becomes more pronounced, with the outer region displaying an evolutionarily older transcriptome than the inner core. In late biofilms, this pattern changes again, indicating that the evolutionary identity of each region depends on developmental stage rather than on spatial position alone. These shifts in evolutionary signatures correspond to changes in regional functional demands during biofilm maturation (Figures 18–21). Functional enrichment analyses show that regions with relatively younger transcriptomic profiles, including the inner regions of early and mid stage biofilms and the outer region of late stage biofilms, are enriched for processes associated with biofilm development, such as motility, cell wall biogenesis, signal transduction, intracellular transport, and extracellular structure formation. Together, these results support the view that *E. coli* UTI89 biofilm development involves both temporal progression and spatial differentiation, with functional and evolutionary organization emerging through the interaction of these two dimensions.

5.4 Future directions

We examined biofilm ontogeny using a solid–air macrocolony biofilm model on agar plates (Figure 1a). While this controlled laboratory setup effectively models critical ecological stressors, such as localized nutrient limitation and starvation, it does not fully capture the complexity of natural environments. It lacks the hydrodynamic flow, host immune factors, and complex multi-species dynamics characteristic of clinical urinary tract infections. Nevertheless, the broad spectrum of observed functional categories in our analyses, spanning fundamental metabolic pathways to specialized virulence mechanisms (Figure 9, Figure 10), suggests that the identified ontogeny–phylogeny correlations remain relevant to natural biofilms. Future research incorporating multi-species bacterial cultures and dynamic host-mimicking models will be crucial to assess how temporal ecological fluctuations and interspecies interactions modulate these baseline evolutionary trajectories. Our spatial analysis focused on expression patterns variation across the horizontal two-

dimensional plane of bulk biofilm macrocolonies (Figure 17a). However, biofilms also display pronounced physiological, metabolic, and structural stratification along their vertical axis (Serra et al., 2013a, 2013b; Klauck et al., 2018; Serra & Hengge, 2019; Díaz-Pascual et al., 2021; Wang et al., 2023). Given that vertical microgradients drive spatial organization reminiscent of multicellular tissues, evolutionary system-level imprints are likely encoded along this vertical dimension as well (Serra et al., 2013a; Klauck et al., 2018). Although sampling along thin vertical profiles presents technical challenges, resolving 3D phylotranscriptomic and phyloproteomic architecture represents a next step for understanding biofilm spatial stratification.

Functional enrichment analysis of differentially expressed genes in *E. coli* showed that many evolutionarily young genes expressed during biofilm development remain functionally uncharacterized, limiting their biological interpretation (Figure 11). A primary limitation in functional enrichment analysis stems from the high proportion of evolutionarily young genes that lack functional annotation due to the absence of sequence similarity with characterized homologs (Weisman et al., 2020). In our dataset, 1,823 genes expressed during biofilm maturation currently lack assigned functions (Figure 11), highlighting critical gaps in the molecular understanding of uropathogenic *E. coli* pathogenesis (Torres-Puig et al., 2022). In this context, stratifying unannotated genes according to their evolutionary origin, as performed by Shi et al. (2020), could provide a useful framework for annotating new genes essential for *E. coli* biofilm development. Shi et al. (2020) demonstrated the value of phylostratigraphy approach in *B. subtilis* by showing that sporulation genes cluster at specific evolutionary time points and that uncharacterized genes from the same phylostrata can have related developmental functions. To bypass the limits of sequence homology, future efforts can utilize integrated non-homology frameworks, combining developmental co-expression profiling, cluster analysis, and phylotranscriptomic mapping (Shi et al., 2020). Similarly, studies on *E. coli* biofilms could target highly expressed, functionally unknown genes from relevant phylostrata, inactivate them, and examine whether these gene knockouts affect biofilm morphotype formation. Such analyses would help assign functions to previously unannotated genes involved in *E. coli* biofilm development.

6 CONCLUSION

This doctoral thesis addresses whether biofilm development in uropathogenic *Escherichia coli* represents a structured developmental program comparable to multicellular development of eukaryotes and *Bacillus subtilis* biofilms. By extending phylostratigraphic profiling beyond eukaryotic model systems to an uropathogenic bacterium, this study delivers an integrated multi-omics analysis of *Escherichia coli* UTI89 biofilm development. Through combined transcriptomic and proteomic profiling, we demonstrate that biofilm development is governed by distinct evolutionary and developmental patterns across both temporal stages and spatial regions.

Our key findings include:

- Transcriptomic and proteomic profiles spanning across 12 time points covering the full trajectory from planktonic culture to one month old biofilms, show coordinated and punctuated shifts, supporting a model in which *E. coli* biofilm development proceeds through discrete ontogenetic phases (early, mid, mid-late, and late), each characterized by a distinct molecular signature.
- Biofilm development in uropathogenic, gram-negative bacteria is not macroevolutionarily naïve. Similar to gram-positive, soil-dwelling *Bacillus subtilis*, *E. coli* shows positive phylogeny-ontogeny correlation through recapitulation pattern in which early biofilm stages are dominated by expression of evolutionarily older genes, whereas later stages show increased expression of younger genes. Evolutionary and developmental regularities between gram-negative *E. coli* and gram-positive *B. subtilis* suggests that such macroevolutionary patterns represent a broader feature of bacterial biofilm development rather than an idiosyncrasy of a single species.
- Functional enrichment analyses indicate a stage-punctuated functional architecture, with distinct stage-specific programs and an enrichment of poorly annotated genes in mid-to-late development, highlighting gaps in our understanding of mature biofilm states in this uropathogenic strain.
- While temporal progression remains the dominant driver of variation in *E. coli* UTI89 expression profiles and evolutionary patterns during biofilm ontogeny, spatially resolved sampling demonstrates an additional layer of region-specific

evolutionary and functional organization that is most pronounced during mid stage development.

Future studies across additional bacterial species, diverse environmental conditions, and spatial dimensions will be essential to determine whether these patterns reflect universal principles of bacterial multicellularity.

7 REFERENCES

- Abzhanov, A. (2013). von Baer's law for the ages: lost and found principles of developmental evolution. *Trends in genetics: TIG*, 29(12), 712–722. <https://doi.org/10.1016/j.tig.2013.09.004>
- Adeolu, M., Alnajar, S., Naushad, S., & S Gupta, R. (2016). Genome-based phylogeny and taxonomy of the 'Enterobacteriales': proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. *International journal of systematic and evolutionary microbiology*, 66(12), 5575–5599. <https://doi.org/10.1099/ijsem.0.001485>
- Ackermann, M. (2015). A functional perspective on phenotypic heterogeneity in microorganisms. *Nature reviews microbiology*, 13(8), 497–508. <https://doi.org/10.1038/nrmicro3491>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Andersson, D. I., Nicoloff, H., & Hjort, K. (2019). Mechanisms and clinical relevance of bacterial heteroresistance. *Nature reviews microbiology*, 17(8), 479–496. <https://doi.org/10.1038/s41579-019-0218-1>
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed March 6, 2024.
- Bathke, J., Konzer, A., Remes, B., McIntosh, M., & Klug, G. (2019). Comparative analyses of the variation of the transcriptome and proteome of *Rhodobacter sphaeroides* throughout growth. *BMC Genomics*, 20, 358. <https://doi.org/10.1186/s12864-019-5749-3>
- Beebout, C. J., Eberly, A. R., Werby, S. H., Reasoner, S. A., Brannon, J. R., De, S., Fitzgerald, M. J., Huggins, M. M., Clayton, D. B., Cegelski, L., & Hadjifrangiskou, M.

- (2019). Respiratory Heterogeneity Shapes Biofilm Formation and Host Colonization in Uropathogenic *Escherichia coli*. *mBio*, 10(2), e02400-18. <https://doi.org/10.1128/mBio.02400-18>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Besharova, O., Suchanek, V. M., Hartmann, R., Drescher, K., & Sourjik, V. (2016). Diversification of Gene Expression during Formation of Static Submerged Biofilms by *Escherichia coli*. *Frontiers in microbiology*, 7, 1568. <https://doi.org/10.3389/fmicb.2016.01568>
- Bonner, J. T. (2001) *First Signals: The Evolution of Multicellular Development*. Princeton University Press
- Brandis, G. (2021). Reconstructing the Evolutionary History of a Highly Conserved Operon Cluster in Gammaproteobacteria and Bacilli. *Genome biology and evolution*, 13(4), evab041. <https://doi.org/10.1093/gbe/evab041>
- Brown, J. R., & Volker, C. (2004). Phylogeny of gamma-proteobacteria: resolution of one branch of the universal tree?. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 26(5), 463–468. <https://doi.org/10.1002/bies.20030>
- Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature reviews genetics*, 21(10), 630–644. <https://doi.org/10.1038/s41576-020-0258-4>
- Bushnel B. (2014). BBMap: a fast, accurate, splice-aware aligner. Available from: <https://sourceforge.net/projects/bbmap/>. Accessed March 6, 2024.
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular biology and evolution*, 38(12), 5825–5829. <https://doi.org/10.1093/molbev/msab293>
- Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri,

- T., Meyer, R. R., Ozersky, P., Armstrong, J. R., Fulton, R. S., Latreille, J. P., Spieth, J., Hooton, T. M., Mardis, E. R., Hultgren, S. J., & Gordon, J. I. (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15), 5977–5982. <https://doi.org/10.1073/pnas.0600938103>
- Cheng, X., Hui, J. H., Lee, Y. Y., Wan Law, P. T., & Kwan, H. S. (2015). A "developmental hourglass" in fungi. *Molecular biology and evolution*, 32(6), 1556–1566. <https://doi.org/10.1093/molbev/msv047>
- Choi, Y., Ahn, S., Park, M., Lee, S., Cho, S., & Kim, H. (2023). HGTree v2.0: a comprehensive database update for horizontal gene transfer (HGT) events detected by the tree-reconciliation method. *Nucleic acids research*, 51(D1), D1010–D1018. <https://doi.org/10.1093/nar/gkac929>
- Claessen, D., Rozen, D. E., Kuipers, O. P., Søgaaard-Andersen, L., & van Wezel, G. P. (2014). Bacterial solutions to multicellularity: a tale of biofilms, filaments and fruiting bodies. *Nature reviews microbiology*, 12(2), 115–124. <https://doi.org/10.1038/nrmicro3178>
- Cole, J. A., Kohler, L., Hedhli, J., & Luthey-Schulten, Z. (2015). Spatially-resolved metabolic cooperativity within dense bacterial colonies. *BMC systems biology*, 9, 15. <https://doi.org/10.1186/s12918-015-0155-1>
- Comeron, J. M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of molecular evolution*, 41(6), 1152–1159. <https://doi.org/10.1007/BF00173196>
- Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12), 1367–1372. <https://doi.org/10.1038/nbt.1511>
- Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research*, 10(4), 1794–1805. <https://doi.org/10.1021/pr101065j>

- Čorak, N., Anniko, S., Daschkin-Steinborn, C., Krey, V., Koska, S., Futo, M., Široki, T., Woichansky, I., Opašić, L., Kifer, D., Tušar, A., Maxeiner, H. G., Domazet-Lošo, M., Nicolaus, C., & Domazet-Lošo, T. (2023). Pleomorphic Variants of *Borrelia burgdorferi* Express Evolutionary Distinct Transcriptomes. *International journal of molecular sciences*, 24(6), 5594. <https://doi.org/10.3390/ijms24065594>
- Díaz-Pascual, F., Lempp, M., Nosh, K., Jeckel, H., Jo, J. K., Neuhaus, K., Hartmann, R., Jelli, E., Hansen, M. F., Price-Whelan, A., Dietrich, L. E., Link, H., & Drescher, K. (2021). Spatial alanine metabolism determines local growth dynamics of *Escherichia coli* colonies. *eLife*, 10, e70794. <https://doi.org/10.7554/eLife.70794>
- Denamur, E., Clermont, O., Bonacorsi, S., & Gordon, D. (2021). The population genetics of pathogenic *Escherichia coli*. *Nature reviews. Microbiology*, 19(1), 37–54. <https://doi.org/10.1038/s41579-020-0416-x>
- Domazet-Lošo, T., Brajković, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in genetics: TIG*, 23(11), 533–539. <https://doi.org/10.1016/j.tig.2007.08.014>
- Domazet-Lošo, T., & Tautz, D. (2010a). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468(7325), 815–818. <https://doi.org/10.1038/nature09632>
- Domazet-Lošo, T., & Tautz, D. (2010b). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC biology*, 8, 66. <https://doi.org/10.1186/1741-7007-8-66>
- Domazet-Lošo, T., Carvunis, A. R., Albà, M. M., Šestak, M. S., Bakaric, R., Neme, R., & Tautz, D. (2017). No Evidence for Phylostratigraphic Bias Impacting Inferences on Patterns of Gene Emergence and Evolution. *Molecular biology and evolution*, 34(4), 843–856. <https://doi.org/10.1093/molbev/msw284>
- Dragoš, A., Kieseewalter, H., Martin, M., Hsu, C. Y., Hartmann, R., Wechsler, T., Eriksen, C., Brix, S., Drescher, K., Stanley-Wall, N., Kümmerli, R., & Kovács, Á. T. (2018). Division of Labor during Biofilm Matrix Production. *Current biology*, 28(12), 1903–1913.e5. <https://doi.org/10.1016/j.cub.2018.04.046>

- Duboule, D. (1994). Temporal colinearity and the phylotypic progression: A basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development*, 1994 Supplement, 135–142. <https://doi.org/10.1242/dev.1994.supplement.135>
- Dunne, K. A., Chaudhuri, R. R., Rossiter, A. E., Beriotto, I., Browning, D. F., Squire, D., Cunningham, A. F., Cole, J. A., Loman, N., & Henderson, I. R. (2017). Sequencing a piece of history: complete genome sequence of the original *Escherichia coli* strain. *Microbial genomics*, 3(3), mgen000106. <https://doi.org/10.1099/mgen.0.000106>
- Elek, A., Kuzman, M., Vlahoviček, K. (2019). coRdon: codon usage analysis and prediction of gene expressivity. Available from: <https://github.com/BioinfoHR/coRdon>. Accessed September 7, 2020.
- El-Halfawy, O. M., & Valvano, M. A. (2015). Antimicrobial heteroresistance: an emerging field in need of clarity. *Clinical microbiology reviews*, 28(1), 191–207. <https://doi.org/10.1128/CMR.00058-14>
- Flemming, H. C., & Wingender, J. (2010). The biofilm matrix. *Nature reviews. Microbiology*, 8(9), 623–633. <https://doi.org/10.1038/nrmicro2415>
- Flemming, H. C., & Wuertz, S. (2019). Bacteria and archaea on Earth and their abundance in biofilms. *Nature reviews microbiology*, 17(4), 247–260. <https://doi.org/10.1038/s41579-019-0158-9>
- Fox, D. R., Asadollahi, K., Samuels, I., Spicer, B. A., Kropp, A., Lupton, C. J., Lim, K., Wang, C., Venugopal, H., Dramicanin, M., Knott, G. J., & Grinter, R. (2025). Inhibiting heme piracy by pathogenic *Escherichia coli* using de novo-designed proteins. *Nature communications*, 16(1), 6066. <https://doi.org/10.1038/s41467-025-60612-9>
- Foxman B. (2010). The epidemiology of urinary tract infection. *Nature reviews. Urology*, 7(12), 653–660. <https://doi.org/10.1038/nrurol.2010.190>
- Foxman, B., & Brown, P. (2003). Epidemiology of urinary tract infections: transmission and risk factors, incidence, and costs. *Infectious disease clinics of North America*, 17(2), 227–241. [https://doi.org/10.1016/s0891-5520\(03\)00005-9](https://doi.org/10.1016/s0891-5520(03)00005-9)

- Futo, M., Opašić, L., Koska, S., Čorak, N., Široki, T., Ravikumar, V., Thorsell, A., Lenuzzi, M., Kifer, D., Domazet-Lošo, M., Vlahoviček, K., Mijakovic, I., & Domazet-Lošo, T. (2021). Embryo-Like Features in Developing *Bacillus subtilis* Biofilms. *Molecular biology and evolution*, 38(1), 31–47. <https://doi.org/10.1093/molbev/msaa217>
- Futo, M., Široki, T., Koska, S., Čorak, N., Tušar, A., Domazet-Lošo, M., & Domazet-Lošo, T. (2022). A novel time-lapse imaging method for studying developing bacterial biofilms. *Scientific Reports*, 12, 21120. <https://doi.org/10.1038/s41598-022-24431-y>
- Galperin, M. Y., Yutin, N., Wolf, Y. I., Vera Alvarez, R., & Koonin, E. V. (2022). Conservation and Evolution of the Sporulation Gene Set in Diverse Members of the Firmicutes. *Journal of bacteriology*, 204(6), e0007922. <https://doi.org/10.1128/jb.00079-22>
- García, V., Grønnemose, R. B., Torres-Puig, S., Kudirkiene, E., Piantelli, M., Ahmed, S., Andersen, T. E., Møller-Jensen, J., Olsen, J. E., & Herrero-Fresno, A. (2021). Genome-wide analysis of fitness-factors in uropathogenic *Escherichia coli* during growth in laboratory media and during urinary tract infections. *Microbial genomics*, 7(12), 000719. <https://doi.org/10.1099/mgen.0.000719>
- Gertler, C., Gerdt, G., Timmis, K. N., Yakimov, M. M., & Golyshin, P. N. (2009). Populations of heavy fuel oil-degrading marine microbial community in presence of oil sorbent materials. *Journal of applied microbiology*, 107(2), 590–605. <https://doi.org/10.1111/j.1365-2672.2009.04245.x>
- Gunne-Braden, A., Sullivan, A., Gharibi, B., Sherif, R. S. M., Maity, A., Wang, Y.-F., Edwards, A., Jiang, M., Howell, M., Goldstone, R., Wollman, R., East, P., & Santos, S. D. M. (2020). GATA3 mediates a fast, irreversible commitment to BMP4-driven differentiation in human embryonic stem cells. *Cell Stem Cell*, 26(5), 693–706.e9. <https://doi.org/10.1016/j.stem.2020.03.005>
- Hadjifrangiskou, M., Gu, A. P., Pinkner, J. S., Kostakioti, M., Zhang, E. W., Greene, S. E., & Hultgren, S. J. (2012). Transposon mutagenesis identifies uropathogenic *Escherichia coli* biofilm factors. *Journal of bacteriology*, 194(22), 6195–6205. <https://doi.org/10.1128/JB.01012-12>

- Haeckel, E. (1868). *Natürliche Schöpfungsgeschichte*. Berlin: Georg Reimer.
- Hall-Stoodley, L., Costerton, J. W., & Stoodley, P. (2004). Bacterial biofilms: from the natural environment to infectious diseases. *Nature reviews. Microbiology*, 2(2), 95–108. <https://doi.org/10.1038/nrmicro821>
- Hanna, A., Berg, M., Stout, V., & Razatos, A. (2003). Role of capsular colanic acid in adhesion of uropathogenic *Escherichia coli*. *Applied and environmental microbiology*, 69(8), 4474–4481. <https://doi.org/10.1128/AEM.69.8.4474-4481.2003>
- Hansen, S., Lewis, K., & Vulić, M. (2008). Role of global regulators and nucleotide metabolism in antibiotic tolerance in *Escherichia coli*. *Antimicrobial agents and chemotherapy*, 52(8), 2718–2726. <https://doi.org/10.1128/AAC.00144-08>
- Heacock-Kang, Y., Sun, Z., Zarzycki-Siek, J., McMillan, I.A., Norris, M.H., Bluhm, A.P., Cabanas, D., Fogen, D., Vo, H., Donachie, S.P., Borlee, B.R., Sibley, C.D., Lewenza, S., Schurr, M.J., Schweizer, H.P. & Hoang, T.T. (2017). Spatial transcriptomes within the *Pseudomonas aeruginosa* biofilm architecture. *Molecular Microbiology*, 106: 976–985. <https://doi.org/10.1111/mmi.13863>
- Head, I. M., Jones, D. M., & Röling, W. F. (2006). Marine microorganisms make a meal of oil. *Nature reviews. Microbiology*, 4(3), 173–182. <https://doi.org/10.1038/nrmicro1348>
- Higgins, D., & Dworkin, J. (2012). Recent progress in *Bacillus subtilis* sporulation. *FEMS Microbiology Reviews*, 36(1), 131–148. <https://doi.org/10.1111/j.1574-6976.2011.00310.x>
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, 47(D1), D309–D314. <https://doi.org/10.1093/nar/gky1085>
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new

- view of the tree of life. *Nature microbiology*, 1, 16048. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Humphries, J., Xiong, L., Liu, J., Prindle, A., Yuan, F., Arjes, H. A., Tsimring, L., & Süel, G. M. (2017). Species-Independent Attraction to Biofilms through Electrical Signaling. *Cell*, 168(1-2), 200–209.e12. <https://doi.org/10.1016/j.cell.2016.12.014>
- Hung, C., Zhou, Y., Pinkner, J. S., Dodson, K. W., Crowley, J. R., Heuser, J., Chapman, M. R., Hadjifrangiskou, M., Henderson, J. P., & Hultgren, S. J. (2013). *Escherichia coli* biofilms have an organized and complex extracellular matrix structure. *mBio*, 4(5), e00645-13. <https://doi.org/10.1128/mBio.00645-13>
- Irie, N., & Kuratani, S. (2011). Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature communications*, 2, 248. <https://doi.org/10.1038/ncomms1248>
- Janda, J. M., & Abbott, S. L. (2021). The Changing Face of the Family Enterobacteriaceae (Order: "Enterobacterales"): New Members, Taxonomic Issues, Geographic Expansion, and New Diseases and Disease Syndromes. *Clinical microbiology reviews*, 34(2), e00174-20. <https://doi.org/10.1128/CMR.00174-20>
- Jo, J., Price-Whelan, A., & Dietrich, L. E. P. (2022). Gradients and consequences of heterogeneity in biofilms. *Nature reviews microbiology*, 20(10), 593–607. <https://doi.org/10.1038/s41579-022-00692-2>
- Jubelin, G., Vianney, A., Beloin, C., Ghigo, J. M., Lazzaroni, J. C., Lejeune, P., & Dorel, C. (2005). CpxR/OmpR interplay regulates curli gene expression in response to osmolarity in *Escherichia coli*. *Journal of bacteriology*, 187(6), 2038–2049. <https://doi.org/10.1128/JB.187.6.2038-2049.2005>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>

- Justice, S. S., Hung, C., Theriot, J. A., Fletcher, D. A., Anderson, G. G., Footer, M. J., & Hultgren, S. J. (2004). Differentiation and developmental pathways of uropathogenic *Escherichia coli* in urinary tract pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(5), 1333–1338. <https://doi.org/10.1073/pnas.0308125100>
- Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., Ohler, U., Bergman, C. M., & Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature*, 468(7325), 811–814. <https://doi.org/10.1038/nature09634>
- Kalinka, A. T., & Tomancak, P. (2012). The evolution of early animal embryos: Conservation or divergence? *Trends in Ecology & Evolution*, 27(7), 385–393. <https://doi.org/10.1016/j.tree.2012.03.007>
- Kaper, J. B., Nataro, J. P., & Mobley, H. L. (2004). Pathogenic *Escherichia coli*. *Nature reviews. Microbiology*, 2(2), 123–140. <https://doi.org/10.1038/nrmicro818>
- Klauck, G., Serra, D. O., Possling, A., & Hengge, R. (2018). Spatial organization of different sigma factor activities and c-di-GMP signalling within the three-dimensional landscape of a bacterial biofilm. *Open biology*, 8(8), 180066. <https://doi.org/10.1098/rsob.180066>
- Koska, S., Leljak-Levanić, D., Malenica, N., Bigović Villi, K., Futo, M., Čorak, N., Jagić, M., Ivanić, A., Tušar, A., Kasalo, N., Domazet-Lošo, M., Vlahoviček, K., & Domazet-Lošo, T. (2025). Developmental phylotranscriptomics in grapevine suggests an ancestral role of somatic embryogenesis. *Communications biology*, 8(1), 265. <https://doi.org/10.1038/s42003-025-07712-w>
- Laganenka, L., Colin, R., & Sourjik, V. (2016). Chemotaxis towards autoinducer 2 mediates autoaggregation in *Escherichia coli*. *Nature communications*, 7, 12984. <https://doi.org/10.1038/ncomms12984>
- Lappin-Scott, H., Burton, S., & Stoodley, P. (2014). Revealing a world of biofilms—The pioneering research of Bill Costerton. *Nature reviews microbiology*, 12(11), 781–787. <https://doi.org/10.1038/nrmicro3343>

- Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8), e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>
- Leimbach, A., Hacker, J., & Dobrindt, U. (2013). E. coli as an all-rounder: the thin line between commensalism and pathogenicity. *Current topics in microbiology and immunology*, 358, 3–32. https://doi.org/10.1007/82_2012_303
- Levin, M., Hashimshony, T., Wagner, F., & Yanai, I. (2012). Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo. *Developmental cell*, 22(5), 1101–1108. <https://doi.org/10.1016/j.devcel.2012.04.004>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (Oxford, England), 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* (Oxford, England), 26(4), 493–500. <https://doi.org/10.1093/bioinformatics/btp692>
- Liu, Y., Beyer, A., & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3), 535–550. <https://doi.org/10.1016/j.cell.2016.03.014>
- Logan B. E. (2009). Exoelectrogenic bacteria that power microbial fuel cells. *Nature reviews. Microbiology*, 7(5), 375–381. <https://doi.org/10.1038/nrmicro2113>
- López, D., Vlamakis, H., Losick, R., & Kolter, R. (2009). Paracrine signaling in a bacterium. *Genes & Development*, 23(14), 1631–1638. <https://doi.org/10.1101/gad.1813709>
- López, D., Vlamakis, H., & Kolter, R. (2010). Biofilms. *Cold Spring Harbor perspectives in biology*, 2(7), a000398. <https://doi.org/10.1101/cshperspect.a000398>
- Lotharukpong, J. S., Zheng, M., Luthringer, R., Liesner, D., Drost, H. G., & Coelho, S. M.

- (2024). A transcriptomic hourglass in brown algae. *Nature*, 635(8037), 129–135. <https://doi.org/10.1038/s41586-024-08059-8>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Maharshak, N., Ryu, H. S., Fan, T. J., Onyiah, J. C., Schulz, S., Otterbein, S. L., Wong, R., Hansen, J. J., Otterbein, L. E., Carroll, I. M., & Plevy, S. E. (2015). Escherichia coli heme oxygenase modulates host innate immune responses. *Microbiology and immunology*, 59(8), 452–465. <https://doi.org/10.1111/1348-0421.12282>
- Martinez-Corral, R., Liu, J., Prindle, A., Süel, G. M., & Garcia-Ojalvo, J. (2019). Metabolic basis of brain-like electrical signalling in bacterial communities. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 374(1774), 20180382. <https://doi.org/10.1098/rstb.2018.0382>
- Mey, A. R., Gómez-Garzón, C., & Payne, S. M. (2021). Iron Transport and Metabolism in Escherichia, Shigella, and Salmonella. *EcoSal Plus*, 9(2), eESP00342020. <https://doi.org/10.1128/ecosalplus.ESP-0034-2020>
- Monds, R. D., & O'Toole, G. A. (2009). The developmental model of microbial biofilms: ten years of a paradigm up for review. *Trends in microbiology*, 17(2), 73–87. <https://doi.org/10.1016/j.tim.2008.11.001>
- Morgan, M., Hervé, P., Valerie, O., Nathaniel, H. (2017). Rsamtools: binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. Available from: <http://bioconductor.org/packages/release/bioc/html/Rsamtools.html>. Accessed September 7, 2020.
- Morikawa M. (2006). Beneficial biofilm formation by industrial bacteria *Bacillus subtilis* and related species. *Journal of bioscience and bioengineering*, 101(1), 1–8. <https://doi.org/10.1263/jbb.101.1>
- Morikawa, M., Kagihiro, S., Haruki, M., Takano, K., Branda, S., Kolter, R., & Kanaya, S. (2006). Biofilm formation by a *Bacillus subtilis* strain that produces gamma-polyglutamate. *Microbiology (Reading, England)*, 152(Pt 9), 2801–2807.

<https://doi.org/10.1099/mic.0.29060-0>

- Mukherjee, S., Seshadri, R., Varghese, N. J., Eloie-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., Coates, R. C., Hadjithomas, M., Pavlopoulos, G. A., Paez-Espino, D., Yoshikuni, Y., Visel, A., Whitman, W. B., Garrity, G. M., Eisen, J. A., Hugenholtz, P., Pati, A., Ivanova, N. N., Woyke, T., Klenk, H. P., ... Kyrpides, N. C. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature biotechnology*, 35(7), 676–683. <https://doi.org/10.1038/nbt.3886>
- Nadell, C. D., Hartmann, R., & Drescher, K. (2016). An Emerging Grip on the Growth of Grounded Bacteria. *ACS nano*, 10(10), 9109–9110. <https://doi.org/10.1021/acsnano.6b06461>
- Ogasawara, H., Yamada, K., Kori, A., Yamamoto, K., & Ishihama, A. (2010). Regulation of the *Escherichia coli* *csgD* promoter: interplay between five transcription factors. *Microbiology (Reading, England)*, 156(Pt 8), 2470–2483. <https://doi.org/10.1099/mic.0.039131-0>
- O'Toole, G., Kaplan, H. B., & Kolter, R. (2000). Biofilm formation as microbial development. *Annual review of microbiology*, 54, 49–79. <https://doi.org/10.1146/annurev.micro.54.1.49>
- Penesyan, A., Paulsen, I. T., Kjelleberg, S., & Gillings, M. R. (2021). Three faces of biofilms: a microbial lifestyle, a nascent multicellular organism, and an incubator for diversity. *NPJ biofilms and microbiomes*, 7(1), 80. <https://doi.org/10.1038/s41522-021-00251-2>
- Pesavento, C., Becker, G., Sommerfeldt, N., Possling, A., Tschowri, N., Mehli, A., & Hengge, R. (2008). Inverse regulatory coordination of motility and curli-mediated adhesion in *Escherichia coli*. *Genes & development*, 22(17), 2434–2446. <https://doi.org/10.1101/gad.475808>
- Pratt, L. A., & Kolter, R. (1998). Genetic analysis of *Escherichia coli* biofilm formation: roles of flagella, motility, chemotaxis and type I pili. *Molecular Microbiology*, 30(2), 285–293. <https://doi.org/10.1046/j.1365-2958.1998.01061.x>
- Prigent-Combaret, C., Prensier, G., Le Thi, T. T., Vidal, O., Lejeune, P., & Dorel, C.

- (2000). Developmental pathway for biofilm formation in curli-producing *Escherichia coli* strains: role of flagella, curli and colanic acid. *Environmental microbiology*, 2(4), 450-464. <https://doi.org/10.1046/j.1462-2920.2000.00128.x>
- Puri, D., Fang, X., & Allison, K. R. (2022). Evidence of a possible multicellular life cycle in *Escherichia coli*. *iScience*, 26(1), 105795. <https://doi.org/10.1016/j.isci.2022.105795>
- Puri, D., & Allison, K. R. (2024). *Escherichia coli* self-organizes developmental rosettes. *Proceedings of the National Academy of Sciences of the United States of America*, 121(23), e2315850121. <https://doi.org/10.1073/pnas.2315850121>
- Quint, M., Drost, H. G., Gabel, A., Ullrich, K. K., Bönn, M., & Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. *Nature*, 490(7418), 98–101. <https://doi.org/10.1038/nature11394>
- R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing Available from: <http://www.R-project.org>. Accessed March 6, 2024.
- Reigstad, C. S., Hultgren, S. J., & Gordon, J. I. (2007). Functional genomic studies of uropathogenic *Escherichia coli* and host urothelial cells when intracellular bacterial communities are assembled. *The Journal of biological chemistry*, 282(29), 21259–21267. <https://doi.org/10.1074/jbc.M611502200>
- Ren, D., Bedzyk, L. A., Thomas, S. M., Ye, R. W., & Wood, T. K. (2004). Gene expression in *Escherichia coli* biofilms. *Applied microbiology and biotechnology*, 64(4), 515–524. <https://doi.org/10.1007/s00253-003-1517-y>
- Rensing S. A. (2016). (Why) Does Evolution Favour Embryogenesis?. *Trends in plant science*, 21(7), 562–573. <https://doi.org/10.1016/j.tplants.2016.02.004>
- Rijavec, M., Müller-Premru, M., Zakotnik, B., & Žgur-Bertok, D. (2008). Virulence factors and biofilm production among *Escherichia coli* strains causing bacteraemia of urinary tract origin. *Journal of medical microbiology*, 57(Pt 11), 1329–1334. <https://doi.org/10.1099/jmm.0.2008/002543-0>
- Riley, E. P., Schwarz, C., Derman, A. I., & Lopez-Garrido, J. (2020). Milestones in *Bacillus*

- subtilis sporulation research. *Microbial cell* (Graz, Austria), 8(1), 1–16.
<https://doi.org/10.15698/mic2021.01.739>
- Rosen, D. A., Hooton, T. M., Stamm, W. E., Humphrey, P. A., & Hultgren, S. J. (2007). Detection of intracellular bacterial communities in human urinary tract infection. *PLoS medicine*, 4(12), e329. <https://doi.org/10.1371/journal.pmed.0040329>
- Røder, H. L., Sørensen, S. J., & Burmølle, M. (2016). Studying Bacterial Multispecies Biofilms: Where to Start?. *Trends in microbiology*, 24(6), 503–513.
<https://doi.org/10.1016/j.tim.2016.02.019>
- Sauer, K., Stoodley, P., Goeres, D. M., Hall-Stoodley, L., Burmølle, M., Stewart, P. S., & Bjarnsholt, T. (2022). The biofilm life cycle: expanding the conceptual model of biofilm formation. *Nature reviews microbiology*, 20(10), 608–620.
<https://doi.org/10.1038/s41579-022-00767-0>
- Serra, D. O., & Hengge, R. (2017). Experimental Detection and Visualization of the Extracellular Matrix in Macrocolony Biofilms. In K. Sauer (Ed.), *C-di-GMP Signaling* (Vol. 1657, pp. 133–145). Springer New York. https://doi.org/10.1007/978-1-4939-7240-1_11
- Serra, D. O., & Hengge, R. (2019). A c-di-GMP-Based Switch Controls Local Heterogeneity of Extracellular Matrix Synthesis which Is Crucial for Integrity and Morphogenesis of *Escherichia coli* Macrocolony Biofilms. *Journal of molecular biology*, 431(23), 4775–4793. <https://doi.org/10.1016/j.jmb.2019.04.001>
- Serra, D. O., Richter, A. M., & Hengge, R. (2013a). Cellulose as an architectural element in spatially structured *Escherichia coli* biofilms. *Journal of bacteriology*, 195(24), 5540–5554. <https://doi.org/10.1128/JB.00946-13>
- Serra, D. O., Richter, A. M., Klauck, G., Mika, F., & Hengge, R. (2013b). Microanatomy at cellular resolution and spatial order of physiological differentiation in a bacterial biofilm. *mBio*, 4(2), e00103–e113. <https://doi.org/10.1128/mBio.00103-13>
- Shapiro, J. A. (1998). Thinking about bacterial populations as multicellular organisms. *Annual review of microbiology*, 52, 81–104.
<https://doi.org/10.1146/annurev.micro.52.1.81>

- Sharma, V., Vashishtha, A., Jos, A. L. M., Khosla, A., Basu, N., Yadav, R., Bhatt, A., Gulani, A., Singh, P., Lakhera, S., & Verma, M. (2022). Phylogenomics of the Phylum Proteobacteria: Resolving the Complex Relationships. *Current microbiology*, 79(8), 224. <https://doi.org/10.1007/s00284-022-02910-9>
- Shi, L., Derouiche, A., Pandit, S., Rahimi, S., Kalantari, A., Futo, M., Ravikumar, V., Jers, C., Mokkaipati, V. R. S. S., Vlahoviček, K., & Mijakovic, I. (2020). Evolutionary Analysis of the *Bacillus subtilis* Genome Reveals New Genes Involved in Sporulation. *Molecular biology and evolution*, 37(6), 1667–1678. <https://doi.org/10.1093/molbev/msaa035>
- Silva, A., Sousa, A. M., Alves, D., Lourenço, A., & Pereira, M. O. (2016). Heteroresistance to colistin in *Klebsiella pneumoniae* is triggered by small colony variants subpopulations within biofilms. *Pathogens and disease*, 74(5), ftw036. <https://doi.org/10.1093/femspd/ftw036>
- Singh, R., Paul, D., & Jain, R. K. (2006). Biofilms: implications in bioremediation. *Trends in microbiology*, 14(9), 389–397. <https://doi.org/10.1016/j.tim.2006.07.001>
- Slack, J. M. W., Holland, P. W. H., & Graham, C. F. (1993). The zootype and the phylotypic stage. *Nature*, 361, 490–492. <https://doi.org/10.1038/361490a0>
- Stewart, P. S., & Franklin, M. J. (2008). Physiological heterogeneity in biofilms. *Nature reviews. Microbiology*, 6(3), 199–210. <https://doi.org/10.1038/nrmicro1838>
- Stoodley, P., Sauer, K., Davies, D. G., & Costerton, J. W. (2002). Biofilms as complex differentiated communities. *Annual review of microbiology*, 56, 187–209. <https://doi.org/10.1146/annurev.micro.56.012302.160705>
- Supek, F., & Vlahovick, K. (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC bioinformatics*, 6, 182. <https://doi.org/10.1186/1471-2105-6-182>
- Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, 34(Web Server issue), W609–W612. <https://doi.org/10.1093/nar/gkl315>

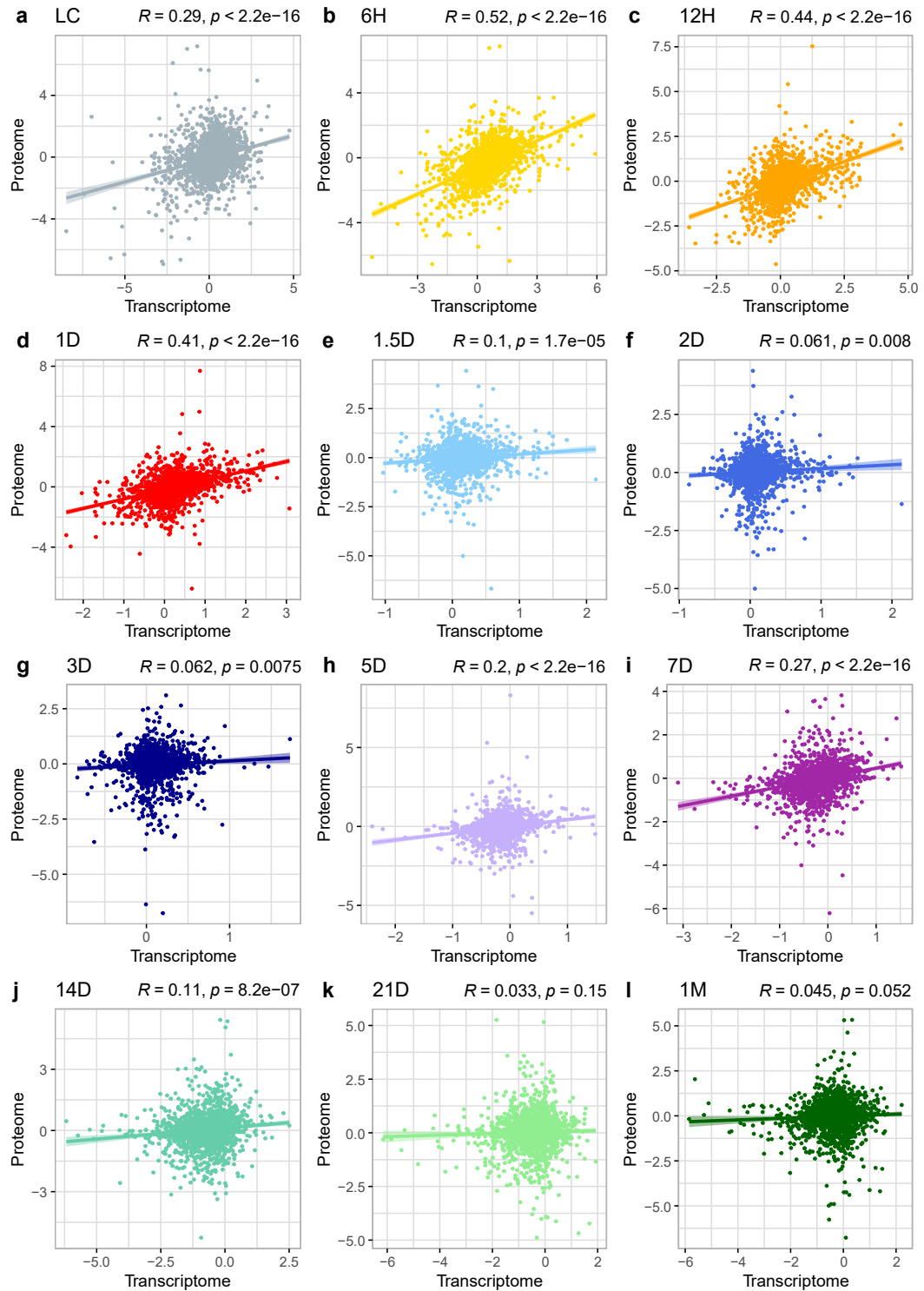
- Tajbakhsh, E., Ahmadi, P., Abedpour-Dehkordi, E., Arbab-Soleimani, N., & Khamesipour, F. (2016). Biofilm formation, antimicrobial susceptibility, serogroups and virulence genes of uropathogenic *E. coli* isolated from clinical samples in Iran. *Antimicrob Resist Infect Control*. 5:11. <https://doi.org/10.1186/s13756-016-0109-4>
- Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature reviews Genetics*, 12(10), 692–702. <https://doi.org/10.1038/nrg3053>
- Tocheva, E. I., Ortega, D. R., & Jensen, G. J. (2016). Sporulation, bacterial cell envelopes and the origin of life. *Nature reviews microbiology*, 14(8), 535–542. <https://doi.org/10.1038/nrmicro.2016.85>
- Tomancak, P., Berman, B. P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S. E., & Rubin, G. M. (2007). Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome biology*, 8(7), R145. <https://doi.org/10.1186/gb-2007-8-7-r145>
- Torres-Puig, S., García, V., Stærk, K., Andersen, T. E., Møller-Jensen, J., Olsen, J. E., & Herrero-Fresno, A. (2022). “Omics” technologies—What have they told us about uropathogenic *Escherichia coli* fitness and virulence during urinary tract infection? *Frontiers in Cellular and Infection Microbiology*, 12, 824039. <https://doi.org/10.3389/fcimb.2022.824039>
- Tušar, A., Koska, S., Futo, M., Kasalo, N., Svetličić, E., Čorak, N., Domazet-Lošo, M., Franjević, D., Mijaković, I., & Domazet-Lošo, T. (in press - 2026). Evolutionary expression patterns in developing biofilms of uropathogenic *Escherichia coli* reveal embryo-like features. *Evolution Letters*.
- Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, 11(12), 2301–2319. <https://doi.org/10.1038/nprot.2016.136>
- Uesaka, M., Kuratani, S., & Irie, N. (2022). The developmental hourglass model and recapitulation: An attempt to integrate the two models. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 338(1-2), 76–86. <https://doi.org/10.1002/jez.b.23027>

- van Gestel, J., Vlamakis, H., & Kolter, R. (2015). Division of Labor in Biofilms: the Ecology of Cell Differentiation. *Microbiology spectrum*, 3(2), microbiolspec.mb-0002-2014. <https://doi.org/10.1128/microbiolspec.MB-0002-2014>
- Vann, W. F., Daines, D. A., Murkin, A. S., Tanner, M. E., Chaffin, D. O., Rubens, C. E., Vionnet, J., & Silver, R. P. (2004). The NeuC protein of *Escherichia coli* K1 is a UDP N-acetylglucosamine 2-epimerase. *Journal of bacteriology*, 186(3), 706–712. <https://doi.org/10.1128/JB.186.3.706-712.2004>
- Vlamakis, H., Chai, Y., Beauregard, P., Losick, R., & Kolter, R. (2013). Sticking together: building a biofilm the *Bacillus subtilis* way. *Nature reviews. Microbiology*, 11(3), 157–168. <https://doi.org/10.1038/nrmicro2960>
- Wang, T., Shen, P., He, Y., Zhang, Y., & Liu, J. (2023). Spatial transcriptome uncovers rich coordination of metabolism in *E. coli* K12 biofilm. *Nature chemical biology*, 19(8), 940–950. <https://doi.org/10.1038/s41589-023-01282-w>
- Weel-Sneve, R., Kristiansen, K. I., Odsbu, I., Dalhus, B., Booth, J., Rognes, T., Skarstad, K., & Bjørås, M. (2013). Single transmembrane peptide DinQ modulates membrane-dependent activities. *PLoS genetics*, 9(2), e1003260. <https://doi.org/10.1371/journal.pgen.1003260>
- Weisman, C. M., Murray, A. W., & Eddy, S. R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLOS Biology*, 18(11), e3000862. <https://doi.org/10.1371/journal.pbio.3000862>
- Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag. Available from: <http://ggplot2.org>. Accessed March 6, 2024.
- Williams, K. P., Gillespie, J. J., Sobral, B. W., Nordberg, E. K., Snyder, E. E., Shallom, J. M., & Dickerman, A. W. (2010). Phylogeny of gammaproteobacteria. *Journal of bacteriology*, 192(9), 2305–2314. <https://doi.org/10.1128/JB.01480-09>
- Wilson, B. A., Foy, S. G., Neme, R., & Masel, J. (2017). Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nature ecology & evolution*, 1(6), 0146–146. <https://doi.org/10.1038/s41559-017-0146>

- Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B. J., Hooper, S. D., Pati, A., Lykidis, A., Spring, S., Anderson, I. J., D'haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., ... Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276), 1056–1060. <https://doi.org/10.1038/nature08656>
- Xia, S., Chen, J., Arsala, D., Emerson, J. J., & Long, M. (2025). Functional innovation through new genes as a general evolutionary process. *Nature genetics*, 57(2), 295–309. <https://doi.org/10.1038/s41588-024-02059-0>
- Yanai, I. (2018). Development and Evolution through the Lens of Global Gene Regulation. *Trends in genetics : TIG*, 34(1), 11–20. <https://doi.org/10.1016/j.tig.2017.09.011>
- Yannarell, S. M., Beaudoin, E. S., Talley, H. S., Schoenborn, A. A., Orr, G., Anderton, C. R., Chrisler, W. B., & Shank, E. A. (2023). Extensive cellular multi-tasking within *Bacillus subtilis* biofilms. *mSystems*, 8(4), e0089122. <https://doi.org/10.1128/msystems.00891-22>
- Yu, D., Banting, G., and Neumann, F. N. (2021). A review of the taxonomy, genetics, and biology of the genus *Escherichia* and the type species *Escherichia coli*. *Canadian Journal of Microbiology*. 67(8): 553-571. <https://doi.org/10.1139/cjm-2020-0508>
- Zagaglia, C., Ammendolia, M. G., Maurizi, L., Nicoletti, M., & Longhi, C. (2022). Urinary Tract Infections Caused by Uropathogenic *Escherichia coli* Strains-New Strategies for an Old Pathogen. *Microorganisms*, 10(7), 1425. <https://doi.org/10.3390/microorganisms10071425>
- Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541(7637), 353–358. <https://doi.org/10.1038/nature21031>
- Zusman, D. R., Scott, A. E., Yang, Z., & Kirby, J. R. (2007). Chemosensory pathways, motility and development in *Myxococcus xanthus*. *Nature Reviews Microbiology*, 5, 862–872. <https://doi.org/10.1038/nrmicro1770>

8 APPENDICES

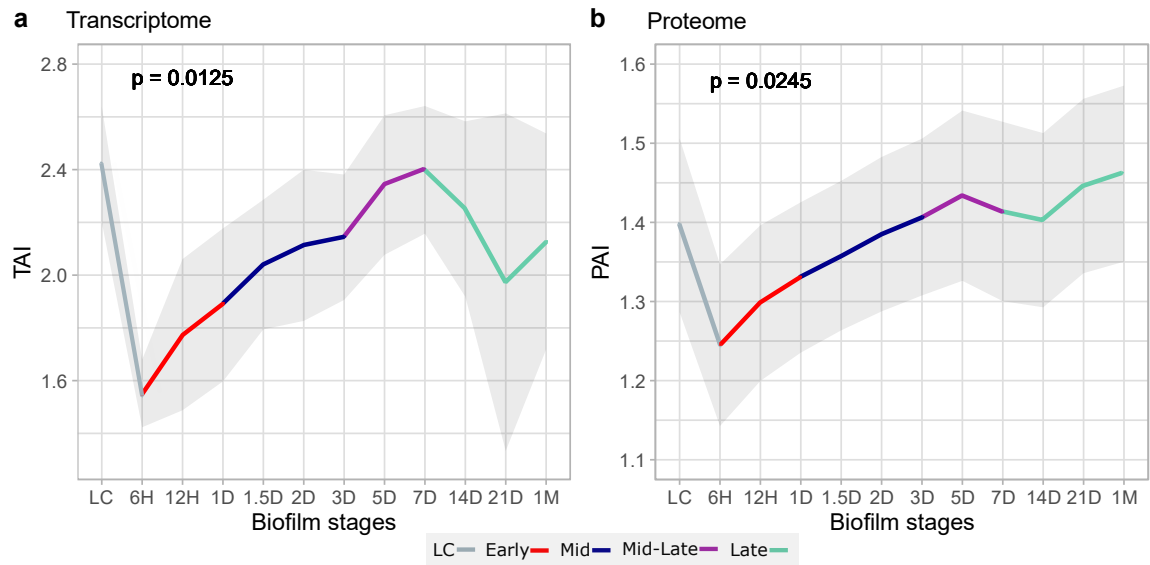
8.1 Figures



Appendix 1. Transcriptome–proteome correlations in *E. coli* UTI89 biofilm samples. Correlation strength varies across temporal samples, ranging from moderate (6H, 12H, 1D; panels

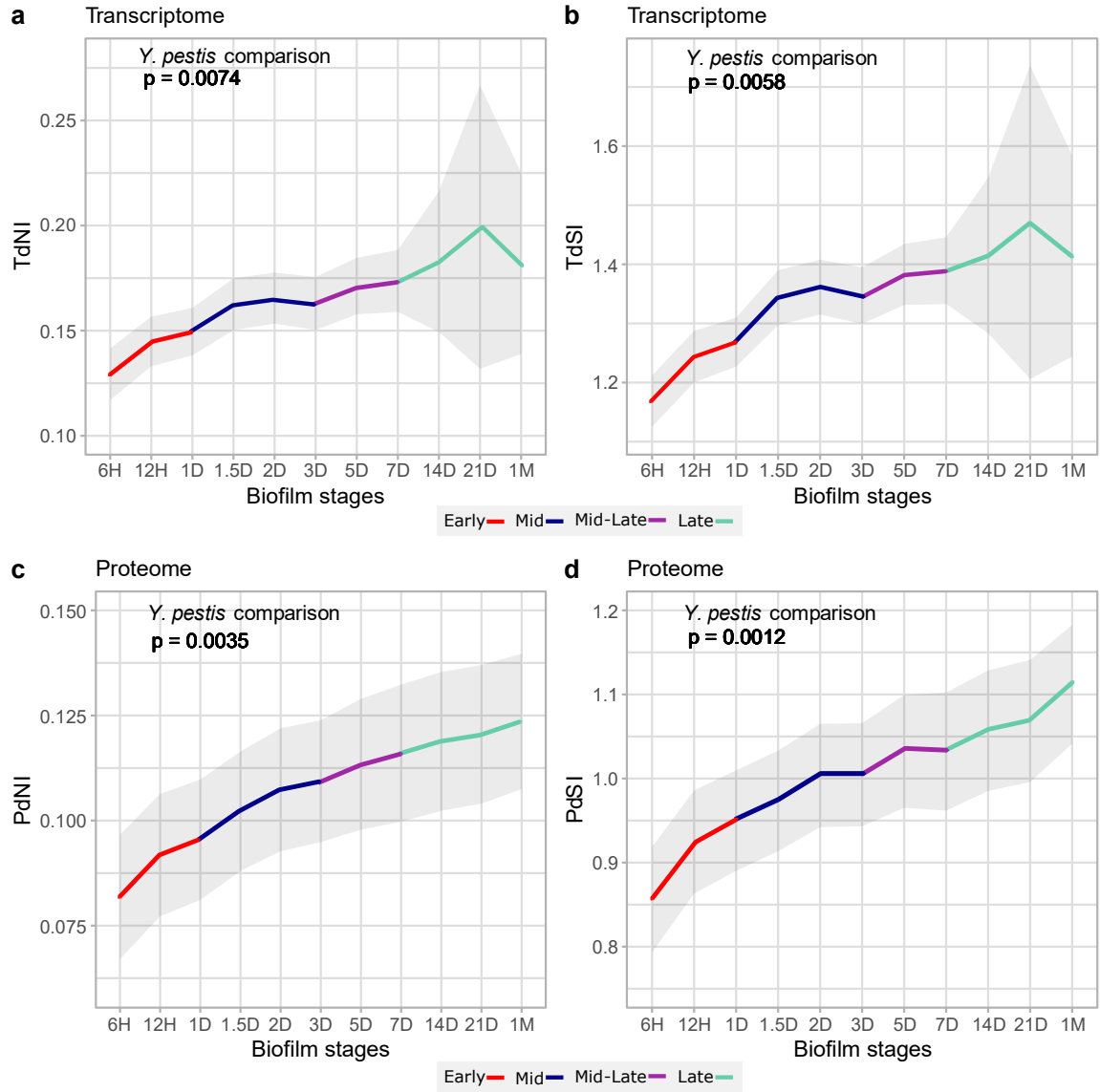
b–d), to low (LC, 5D, 7D; panels **a**, **h**, **i**), to very low (1.5D, 2D, 3D, 14D, 21D, 1M; panels **e–g**, **j–l**). Pearson’s correlation coefficients (R) were calculated from normalized transcriptome and proteome expression values on a $-\log_2$ scale, and the corresponding p-values are reported. Genes or proteins with zero expression in either the transcriptome or proteome dataset were excluded from the analysis. Supplementary dataset is provided in Tušar et al. (2026).

Appendix 2. Correlation analysis of expression values across biofilm time points and spatial regions using unresolved biological replicates. For these analyses, biological replicates were treated as independent samples. Pearson's correlation coefficients between biofilm ontogeny time points are shown for the (a) transcriptome and (b) proteome datasets. Pearson's correlation coefficients across spatial regions (inner, middle, outer) and developmental stages (1D, 3D, 14D) are shown for the (c) transcriptome and (d) proteome datasets. Replicates are indicated by numerical suffixes (1, 2, 3) in the sample labels. Supplementary dataset is provided in Tušar et al. (2026).



Appendix 3. Evolutionary indices calculated including the starting liquid culture sample (LC).

(a) Transcriptome age index (TAI) and (b) proteome age index (PAI) profiles of *E. coli* UTI89 biofilm ontogeny reveal a recapitulation pattern, in which early biofilm stages predominantly express evolutionarily older genes, whereas later stages express younger genes (Appendix_). The significance of these evolutionary patterns was evaluated using the early conservation test (see Methods). Grey shaded areas indicate $\pm$ one standard deviation based on permutation analysis. Biofilm developmental stages are color-coded as follows: early (red), mid (blue), mid-late (purple), and late (green). Supplementary dataset is provided in Tušar et al. (2026).



Appendix 4. Synonymous and non-synonymous rates estimated from pairwise comparisons between *Escherichia coli* UTI89 and *Yersinia pestis*. (a) Transcriptome non-synonymous divergence index (TdNI) and (b) transcriptome synonymous divergence index (TdSI) exhibit an early conservation pattern. (c) Proteome non-synonymous divergence index (PdNI) shows a similar early conservation pattern, consistent with (d) proteome synonymous divergence index (PdSI). Reported p-values were obtained using early conservation test, and grey shading indicates $\pm$ one standard deviation estimated by permutation analysis. Biofilm stages are color-coded as early (red), mid (blue), mid to late (purple), and late (green). Supplementary dataset is provided in Tušar et al. (2026).

8.2 Tables

Appendix A. Distribution of HGT-derived *E. coli* UTI89 genes across phylostrata using strict and standard methods for horizontal gene transfer (HGT) genes detection. The table shows the distribution of candidate HGT-derived genes across phylostrata. Candidate HGT-derived genes were identified using the HGTree database, applying both the standard and strict classification thresholds, which differ according to reference database size (Choi et al., 2023). This analysis was used to evaluate whether horizontal gene transfer could influence the inferred evolutionary age distribution of genes and the resulting TAI and PAI profiles. The table was adapted from Tušar et al. (2026).

Target	Method type	Query	Locus_tag	Protein_ID	PS
WP_000104291	standard,strict	pid 0000000000364106931 tx 364106	UTI89_C0956	ABE06441.1	5
WP_000123468	standard,strict	pid 0000000000364106924 tx 364106	UTI89_C0949	ABE06434.1	3
WP_000127029	standard	pid 00000000003641062261 tx 364106	UTI89_C2303	ABE07771.1	1
WP_000153208	standard	pid 0000000000364106280 tx 364106	UTI89_C0288	ABE05790.1	1
WP_000158968	standard	pid 0000000000364106936 tx 364106	UTI89_C0961	ABE06446.1	6
WP_000206716	standard	pid 0000000000364106917 tx 364106	UTI89_C0942	ABE06427.1	7
WP_000211273	standard	pid 0000000000364106922 tx 364106	UTI89_C0947	ABE06432.1	6
WP_000219338	standard	pid 00000000003641062635 tx 364106	UTI89_C2679	ABE08145.1	6
WP_000386637	standard	pid 00000000003641062624 tx 364106	UTI89_C2668	ABE08134.1	5
WP_000389073	standard	pid 00000000003641065031 tx 364106	UTI89_C5141	ABE10541.1	3
WP_000431541	standard	pid 00000000003641062597 tx 364106	UTI89_C2641	ABE08107.1	5
WP_000623548	standard	pid 00000000003641062263 tx 364106	UTI89_C2305	ABE07773.1	1

Appendix A. - continued.

Target	Method type	Query	Locus_tag	Protein_ID	PS
WP_000632354	standard	pid 000000000364106911 tx 364106	UTI89_C0936	ABE06421.1	3
WP_000697342	standard	pid 0000000003641062262 tx 364106	UTI89_C2304	ABE07772.1	1
WP_000741755	standard	pid 0000000003641065025 tx 364106	UTI89_C5135	ABE10535.1	2
WP_000767115	standard	pid 0000000003641064994 tx 364106	UTI89_C5104	ABE10504.1	2
WP_000776265	standard	pid 000000000364106937 tx 364106	UTI89_C0962	ABE06447.1	5
WP_000853458	standard,strict	pid 000000000364106893 tx 364106	UTI89_C0918	ABE06403.1	2
WP_000954672	standard	pid 0000000003641065030 tx 364106	UTI89_C5140	ABE10540.1	9
WP_000972139	standard	pid 000000000364106898 tx 364106	UTI89_C0923	ABE06408.1	5
WP_001036814	standard	pid 000000000364106929 tx 364106	UTI89_C0954	ABE06439.1	6
WP_001163786	standard	pid 000000000364106921 tx 364106	UTI89_C0946	ABE06431.1	2
WP_001555908	standard	pid 0000000003641061704 tx 364106	UTI89_C1736	ABE07214.1	9
WP_011478365	standard	pid 0000000003641065022 tx 364106	UTI89_C5132	ABE10532.1	2

Appendix B. Functional enrichment of functionally uncharacterized genes within phylostrata (PS) expressed during biofilm development was assessed using the hypergeometric test: quant = the number of unannotation genes expressed during biofilm development per phylostratum, sample = the number of all unannotated genes expressed during biofilm development, hit = the number of all expressed genes per phylostrata, total = the number of all expressed genes in all phylostrata. P values were adjusted for multiple comparisons using the Benjamini and Hochberg procedure (FDR_p_value).

PS	PS_name	quant	sample	hit	total	log_odds	raw_p_value
1	Cellular_organism	552	1823	2819	4863	-1,909944827	4,16E-205
2	Bacteria	431	1823	942	4863	0,426820906	4,47E-09
3	Proteobacteria	97	1823	174	4863	0,771301233	4,87E-07
4	Gammaproteobacteria	78	1823	151	4863	0,597046171	0,000220844
5	Enterobacterales	199	1823	277	4863	1,537568472	5,01E-33
6	Enterobacteriaceae	147	1823	175	4863	2,244422384	4,69E-38
7	Escherichia_Shigella_group	181	1823	186	4863	4,203355532	4,33E-72
8	Escherichia_coli_group_B2_F	57	1823	58	4863	4,58586268	1,08E-23
9	Escherichia_coli_UTI89	81	1823	81	4863	Inf	9,82E-36

Appendix C. Calculated averages for GO-annotated and unannotated genes across phylostrata for secondary-structure measures and AlphaFold confidence scores. The table summarizes average structural properties of *E. coli* UTI89 proteins assigned to each phylostratum (PS), comparing proteins encoded by GO-annotated genes with those encoded by genes without GO annotations. Structural support was evaluated using AlphaFold predictions (Jumper et al., 2021). Mean local prediction confidence scores were extracted from the B-factor fields of AlphaFold models, while the number of helices and sheets was determined using the DSSP algorithm. These measures were used to assess whether unannotated genes, particularly evolutionarily younger genes, show comparable structural support to annotated genes. Supplementary data supporting this analysis are provided in Tušar et al. (2026).

Proteins with GO-annotation				
PS	helix	sheet	mean_confidence	median_confidence
1	16,09	11,50	90,77	94,40
2	11,49	8,90	86,13	90,72
3	7,95	7,91	82,31	86,24
4	5,78	6,83	81,13	86,16
5	4,56	3,84	76,78	80,09
6	3,85	3,23	76,72	80,70
7	3,55	2,36	76,51	79,99
8	2,00	0,00	76,84	80,88
9	NA	NA	NA	NA

Appendix C. - continued.

Proteins without GO-annotation

PS	helix	sheet	mean_confidence	median_confidence
1	15,15	10,19	88,89	92,81
2	10,76	9,73	83,80	88,38
3	7,05	7,17	79,61	84,53
4	6,32	5,95	82,13	87,41
5	4,93	5,00	79,33	83,12
6	3,76	2,07	67,10	68,95
7	2,49	1,07	61,20	62,09
8	3,26	2,28	62,88	63,81
9	2,32	1,10	53,50	53,52

Appendix D. Differences between proteins with known functional annotation and proteins without GO annotation tested using Mann–Whitney tests.

The table summarizes statistical comparisons between GO-annotated and unannotated *E. coli* UTI89 proteins for AlphaFold prediction confidence and secondary-structure measures. Differences between the two protein groups were tested using Mann–Whitney U tests, with the alternative hypothesis that values were lower in the unannotated group. Tested measures included AlphaFold confidence scores and the number or density of secondary structures, including helices and sheets. P-values were corrected for multiple testing using the Benjamini–Hochberg false discovery rate procedure. This analysis was used to assess whether unannotated proteins show reduced structural confidence or lower secondary-structure content compared with proteins with known functional annotation. Supplementary data supporting this analysis are provided in Tušar et al. (2026). knownF = known functional annotation; unknownF = unknown functional annotation

measure	n_unknownF	n_knownF	median_unknownF	median_knownF	mean_unknownF	mean_knownF	W	p_adj
mean confidence	1823	3383	82,81	90,56	76,61202874	88,67435066	1405601	4,16E–152
median confidence	1823	3383	88,75	95,12	79,77084146	92,51026308	1407415	5,72E–152
helix	1823	3383	5	13	8,137751787	14,12160187	1453238	2,26E–141
sheet	1823	3383	4	9	6,249512671	10,37065186	1768556	6,52E–78

9 CURRICULUM VITAE

Anja Tušar earned her BSc degree in Microbiology (2016) and subsequently completed a MSc in Biotechnology (2019) at the Biotechnical Faculty, University of Ljubljana, Slovenia. As a bachelor student, she expanded her interest in microbiology with a two-month Erasmus internship at the Max-Planck Institute in Marburg, Germany (2015). In 2018, she undertook another Erasmus internship at the Technical University of Denmark, working with *Bacillus subtilis*, as a key model organism for studying biofilm formation and sporulation. For Master's thesis project at the National Institute of Chemistry in Slovenia (2019) and a 2.5-month ASEF fellowship with the Kobe research team at the University of Queensland (2019), she focused on various molecular cloning techniques, and delve into protein production, isolation, and characterization. From 2019 to 2021, she was researcher for a collaborative project between the Biotechnical Faculty in Slovenia and the pharmaceutical company Lek d.d. (Sandoz Group AG), where she helped develop molecular biology tools for mammalian cell lines. In 2021, Anja began her doctoral studies in Biology at the Faculty of Science and became a Marie-Curie doctoral student within the PEST-BIN project (Pioneering Strategies Against Bacterial Infections, No. 955626) at the Ruđer Bošković Institute in Croatia. Under the mentorship of the Domazet-Lošo group, her research has expanded to include bioinformatics, phylogenetics, and the evolutionary and developmental biology of bacterial biofilms.

Publications:

1. Futo, M., Široki, T., Koska, S., Čorak, N., Tušar, A., Domazet-Lošo, M., & Domazet-Lošo, T. (2022). A novel time-lapse imaging method for studying developing bacterial biofilms. *Scientific reports*, 12(1), 21120. <https://doi.org/10.1038/s41598-022-24431-y>
2. Čorak, N., Anniko, S., Daschkin-Steinborn, C., Krey, V., Koska, S., Futo, M., Široki, T., Woichansky, I., Opašić, L., Kifer, D., Tušar, A., Maxeiner, H. G., Domazet-Lošo, M., Nicolaus, C., & Domazet-Lošo, T. (2023). Pleomorphic Variants of *Borrelia* (syn. *Borrelia*) burgdorferi Express Evolutionary Distinct Transcriptomes. *International journal of molecular sciences*, 24(6), 5594. <https://doi.org/10.3390/ijms24065594>
3. Koska, S., Leljak-Levanić, D., Malenica, N., Bigović Villi, K., Futo, M., Čorak, N., Jagić, M., Ivanić, A., Tušar, A., Kasalo, N., Domazet-Lošo, M., Vlahoviček, K., & Domazet-Lošo, T. (2025). Developmental phylotranscriptomics in grapevine suggests an ancestral role of somatic embryogenesis. *Communications biology*, 8(1), 265. <https://doi.org/10.1038/s42003-025-07712-w>
4. Tušar, A., Koska, S., Futo, M., Kasalo, N., Svetličić, E., Čorak, N., Domazet-Lošo, M., Franjević, D., Mijaković, I., & Domazet-Lošo, T. (in press - 2026). Evolutionary expression patterns in developing biofilms of uropathogenic *Escherichia coli* reveal embryo-like features. *Evolution Letters*.