PROBING MRNA-PROTEIN RELATIONSHIPS ACROSS PROKARYOTES: FROM *PSEUDOMONAS* TO *SULFOLOBUS*

A Dissertation
Presented to
The Academic Faculty

by

Mengshi Zhang

In Partial Fulfillment
of the Requirements for the Degree
Doctorate of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology

August 2024

PROBING MRNA-PROTEIN RELATIONSHIPS ACROSS PROKARYOTES: FROM *PSEUDOMONAS* TO *SULFOLOBUS*

Approved by:

Dr. Marvin Whiteley, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Peter Yunker
School of Physics
Georgia Institute of Technology

Dr. Sam Brown
School of Biological Sciences
Georgia Institute of Technology

Dr. Steve Diggle
School of Biological Sciences
Georgia Institute of Technology

Dr. David Weiss
School of Medicine
Emory University

Date Approved: July 10th, 2024

ACKNOWLEDGEMENTS

First of all, I would express my greatest appreciation to my mentor, Dr. Marvin Whiteley, for keen and invaluable guidance and help throughout every step of my Ph.D. journey. I am grateful to all the members of the Whiteley lab, past and present, for their assistance, advice, good times, and occasional distractions. I extend my thanks to all my committee members for providing valuable advice on my research and to Dr. William Ratcliff for being supportive as a program director.

I extend my gratitude to my wonderful collaborators, including Dr. Changyi Zhang and Dr. Rachel Whitaker's lab at UIUC, Dr. Kelly Michie at the Mayo Clinic, Dr. Stephen Dolan's lab at Clemson University, and Dr. Daniel Cornforth at CDC. Special thanks to Duc Duong at Emory Integrated Proteomics Core, and Molecular Evolution Core at Georgia Tech. This work would not be possible without their help and funding from the National Institutes of Health, the Cystic Fibrosis Foundation, and the Burroughs Wellcome fund. I also want to thank my mentors, Dr. Guanyu Wang, and Dr. Micheal Deem for getting me started on my scientific journey and helping my Ph.D. application process. Additionally, I am grateful to my gym friends, career mentors (Dr. Avery Bell, Dr. Elizabeth Leslie and Dr. Jimena Andersen) from the women+ computational biology mentorship program, and all those wonderful people who have supported and heard me throughout this journey.

Lastly, a special thank you to embroidery, and other handicrafts for providing moments of relaxation and creativity amidst the distress of academia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

A. actinomycetemcomitans	*Aggregatibacter actinomycetemcomitans*
ATP	Adenosine triphosphate
BPG	1,3-bisphosphoglycerate
C. albicans	*Candida albicans*
CDC	Centers for Disease Control and Prevention
CF	Cystic fibrosis
CLASI-FISH	Combinatorial labeling and spectral imaging-FISH
E.coli	*Escherichia coli*
E4P	Eerythrose-4-phosphate
F. nucleatum	*Fusobacterium nucleatum*
F6P	Fructose-6-phosphate
$FADH_2$	Reduced flavin adenine dinucleotide
FISH	Fluorescence in situ hybridization
G3P	Glyceraldehyde 3-phosphate
LASSO	Least Absolute Shrinkage and Selection Operator
LC-MS/MS	Liquid chromatography coupled to tandem mass spectrometry
M. pneumoniae	*Mycoplasma pneumoniae*
M. tuberculosis	*Mycobacterium tuberculosis*
Mbp	Megabase pair
MIC	Minimum inhibitory concentration
MS	Mass spectrometry
NADH	Nicotinamide Adenine Dinucleotide and Hydrogen

NADPH	Nicotinamide adenine dinucleotide phosphate
nanoLC-MS/MS	Nanoscale liquid chromatography coupled to tandem mass spectrometry
NGS	Next-generation Sequencing
P. aeruginosa	*Pseudomonas aeruginosa*
P. catoniae	*Porphyromonas catoniae*
P. gingivalis	*Porphyromonas gingivalis*
P. pasteri	*Porphyromonas pasteri*
PEP	Phosphoenolpyruvate
PTMs	Post-translational modifications
PTR	Protein-to-mRNA ratio
QS	Quorum Sensing
R5P	Ribose-5-phosphate
Ri5P	Ribulose-5-phosphate
RNA-seq	RNA sequencing
RTP	RNA-to-protein
S. aureus	*Staphylococcus aureus*
S. gordonii	*Streptococcus gordonii*
S. islandicus	*Sulfolobus islandicus*
S7P	Sedoheptulose-7-phosphate
SD	Standard deviation
SR	Standardized residuals
UIUC	University of Illinois Urbana-Champaign
X5P	Xylulose-5-phosphate
2PG	2-phosphoglycerate
3PG	3-phosphoglycerate

SUMMARY

Understanding the biology of native microbial communities is hindered by the lack of robust functional data for the microbes within these communities. One way to tackle this problem is to quantify gene expression in native communities and use this data to infer microbial function. Although RNA-seq has been widely used to study bacterial physiology in situ, a critical concern arises regarding whether mRNA levels accurately predict protein levels, which are the primary functional units of a cell. Here, we addressed this challenge systematically by using comprehensive transcriptome and proteome datasets from Gram-negative bacteria, Gram-positive bacteria, and an archaea. This thesis explores three questions: (i) How does growth rate impact mRNA-protein correlations in the human pathogen *Pseudomonas aeruginosa*?; (ii) How do mRNA-protein correlations change across six prokaryotes?; (iii) Can protein level prediction from mRNA levels be improved? Here, we discovered that the overall correlation of mRNA and protein is similar across different growth rates in *P. aeruginosa* and across diverse prokaryotes, with mRNA and protein positively correlated. However, genes essential for viability have higher mRNA-protein correlations, and both mRNAs and proteins from these essential genes are produced at higher levels compared to non-essential genes. We used statistical methods to identify 'outlier' genes in which mRNA and protein were poorly correlated in six prokaryotes and showed that RTP conversion factors can be used to improve the predictivity of protein levels across strains and growth conditions. Indeed, RTP conversion factors calculated from bacteria were shown to improve protein predictivity in a hyperthermophilic archaea, providing proof-of-principle that this approach is robust across

domains of life. Collectively, our results provide new insights into mRNA-protein relationships and provide valuable tools for inferring *in situ* bacterial function from transcriptome data.

CHAPTER 1. INTRODUCTION

1.1 Overview

This thesis aims to explore the mRNA-protein relationship in two aspects, 1) How growth rate impacts the mRNA-protein relationship in the model system *P. aeruginosa*; 2) How the mRNA-protein relationship changes across bacteria and archaea.

Chapter 1, the introduction, focuses on two distinct model systems: *P. aeruginosa* and the polymicrobial system oral microbiota. The chapter initiates by providing a foundational overview of *P. aeruginosa* microbiology, emphasizing its key characteristics during infections and growth physiology. Subsequently, it delves into the methodologies employed to study bacterial physiology during infection, such as the development of sophisticated laboratory models that mimic infection sites, transcriptional profiling, and proteomic analysis. Chapter 2 presents research on the impact of the growth rate on mRNA and protein contents in *P. aeruginosa*. Chapter 3 advances the research by tackling two research goals: (i) to explore mRNA-protein relationship across diverse prokaryotes; and (ii) determine if mRNA-protein ratios from a specific bacterium or a group of bacteria can be used to improve protein predictivity from mRNA levels in other prokaryotes. Last but not least, Chapter 4 concludes the thesis by summarizing the main findings and discussing potential future directions.

1.2 Model system *P. aeruginosa*

1.2.1 P. aeruginosa microbiology

P. aeruginosa is a Gram-negative, rod-shaped, motile, aerobic, and non-spore forming bacterium that can cause various infections in both healthy and immunocompromised hosts. It is commonly found in a variety of natural environments and host systems, such as soil (1, 2) and water (3, 4), as well as plants, insects and humans (5-8). The ability to adapt and thrive in highly diverse ecological niches stems from its extremely versatile metabolic adaptability, which is associated with a large and complex genome (9). As a highly resourceful pathogen, *P. aeruginosa* encodes multiple mechanisms for antibiotic resistance and virulence factors, including efflux systems, antibiotic-inactivating enzymes, and biofilm formation. Unlike other bacteria, the large genome of *P. aeruginosa* is not merely a result of genome expansion induced by gene duplication (10), but is attributed to the presence of distinct gene families, reflecting a large functional diversity and genetic capacities (11, 12).

The genome size of *P. aeruginosa* ranges from approximately 5.5-7 Mbp, containing a highly conserved core genome and a diverse accessory genome that varies among different strains (13). The core genome of *P. aeruginosa* contains genes enriched with pathways related to the metabolism of amino acids and carbohydrates, cell maintenance and cellular processes. In addition, most *P. aeruginosa* strains exhibit a large number of genes involved in the processing of the genetic and environmental information, particularly ABC transporters and two-component systems (14). The accessory genomes of *P. aeruginosa* reflect the functional diversity of this species and the acquisition of new metabolic functions to adapt to nutrient-limited environments. The plasticity of accessory genomes among *P. aeruginosa* strains at a specific infection site is mostly results from frequent genetic exchange between strains through horizontal gene transfer predominantly mediated by bacteriophages, which give rise to a non-clonal population of *P. aeruginosa* within a specific habitat (12, 15).

1.2.2 P. aeruginosa physiology during infection

Opportunistic bacterial pathogens like *P. aeruginosa* are a leading cause of life-threatening infections in healthcare settings. *P. aeruginosa* causes both acute and chronic infections (16), which are often difficult to eliminate using conventional antimicrobial therapies (17, 18).

Acute infection

To establish an acute infection, *P. aeruginosa* must quickly survive and adapt in the host environment and avoid elimination by the host immune system. This short-term adaptation requires a finely tuned interplay of a group of regulators that ensure survival and promote pathogenicity. For example, the expression of toxins and proteases can facilitate bacterial migration within damaged tissues (19).

Chronic infection

If *P. aeruginosa* successfully colonizes an infection site and evades elimination by host factors or antibiotics, it can establish long-term chronic infections and accumulate patho-adaptive mutations. Strains isolated from chronic infections, particularly in CF lungs, display a large phenotypic variability, and enhanced biofilm-forming capabilities, such as mucoid strains with overproduction of alginates, or small colony variants (20, 21). The ability to enter biofilm state enables *P. aeruginosa* to thrive in hostile environments and provides fitness advantages to bacteria compared to individual, free-living planktonic cells. The biofilm lifestyle is associated with increased resistance to antibiotic treatments, environmental stressors, and host immune responses (22). Biofilm-associated chronic infection typically involve high bacterial loads and worsening lung function (23).

1.2.3 Bacterial growth physiology and growth rates

Bacterial growth physiology involves regulation of macromolecule synthesis, cell division, and adaption to changes of external environments, such as nutrient depletion. In natural environments, bacteria frequently encounter limited nutrients, leading them to shift into a state of very slow growth or no growth. Bacterial generation times can range from days to months due to extremely low levels of nutrients (24, 25).

Adaptive responses to nutrient limitations

In response to essential nutrient depletion, bacteria adapt by two strategies to keep growing: 1) induce another uptake system with higher efficiency to utilize the low concentration of nutrients; 2) synthesize enzymes to use alternative nutrients [4, 7]. For example, *P. aeruginosa* activates a more efficient high-affinity zinc uptake system (ZnuABC) and the zinc uptake regulator (Zur) when suffering from zinc limitations in the sputum of CF patients (26). If bacteria fail to bring in essential nutrients for growth and cell division, then bacteria enter stationary phase or from spores or cyst (27).

Diauxic growth

Bacterial growth on complex media with multiple carbon sources may undergo diauxic growth, characterized by a biphasic growth curve. Diauxic growth occurs when bacteria initially grow on the preferential carbon source and then switch to a secondary carbon source. The most common preferential carbon source is glucose. Many bacteria, including *E.coli*, first grow exponentially on glucose, and sufficiently depleted the glucose in the medium. Then bacteria enter to a lag phase for synthesizing enzymes necessary for the growth on the lactose. After the lag phase, bacteria resume exponential growth on the second carbon source (28, 29).

1.3 Measurement of bacterial physiology during infection

1.3.1 Chemostats and steady state growth

Steady state growth rates characterize the bacterial population of a culture that double at each cell division event and maintain a constant doubling time. Steady state happens when a cell culture undergoes exponential growth by subculturing before entering stationary phase or during continuous growth in a device called a chemostat.

In a chemostat, bacteria maintain growth in a chamber that continuously supplies fresh medium at a defined growth rate maintained by a controller. The concentration of the limiting nutrient is kept low to ensure no excess growth. While maintaining a constant chemostat volume, the growth in the bacteria in the growth chamber is limited by the nutrient availability. When a drop of fresh medium flow into the growth chamber, the bacteria immediately utilized the grow-limiting nutrient and pause growth until the next drop of fresh nutrient inflow. Therefore, the rates of nutrient inflow provides a way to control the growth rates. The carrying capacity of bacteria is controlled by concentration of limiting nutrient in the chemostat chamber (30, 31).

<u>*Growth rates of P. aeruginosa*</u>

Our understanding of the basic aspects of bacterial physiology is highly biased toward extremely fast-growing bacteria. Not only are high growth rates not generally the norm in the natural environment, but growth rates are also highly dynamic and can vary widely within a population (32-35). Indeed, previous studies revealed that the growth rates of bacteria in the human body during chronic infection can vary widely depending on the nutritional environment of the infection site and the immune response. For example, the doubling time of *P. aeruginosa* chronically infecting the lungs of people with cystic fibrosis

(CF) has been shown to vary from <1 h to >24 h (33), with population averages of between 2.5 and 4.6 h (35).

1.3.2 Development of laboratory models that mimic infection sites

The niche of human host environment significantly impacts bacterial physiology during infection. To closely resemble the host habitat, extensive efforts have been dedicated to development of *in vitro* experimental model to resemble the host habitat. For example, synthetic cystic fibrosis sputum has been developed to reflect the nutrient composition and physical properties of CF sputum (36). Various experimental models have been utilized to integrate host factors into the medium, such as burn wound exudate (37), mucin glycans (38), native human mucins (38), and expectorated mucin from human respiratory tract (39).

The *P. aeruginosa* physiology is further shaped by co-infecting pathogens and the presence of commensals, particularly through their secreted metabolites. Mix-species models has been utilized to study the co-infection dynamics, unveiling the intricate reciprocal interactions and crosstalk. For example, the peptidoglycan secreted by *S. aureus* can induce the production of extracellular virulence factors in *P. aeruginosa* (40). The phenazine biosynthesis genes were upregulated in *P. aeruginosa* responding to secreted ethanol when co-cultured with *C. albicans (41)*.

Host factors also play a pivotal role in inducing physiological changes in *P. aeruginosa*. Those observations are supported by *in vitro* experiments and further advanced through animal model systems. For example, acute mouse pneumonia model and acute murine burn model have been developed to resemble acute infection scenarios

(42). For chronic infections, models like the mouse tumour model (43) and murine surgical wound model have been employed to resemble the host factors present during prolonged infection (44). Besides animal models, lung models have been unutilized to closely mimic the physiology of the infected human body sites, such as *ex vivo* porcine lung model and in vivo burn wound models in pigs (45).

1.3.3 Transcriptional profiling of P. aeruginosa infections: from in vitro to ex vivo

Transcriptional profiling is a key tool for studying bacterial physiology, enabling the identification of environmentally driven transcriptional responses in bacterial pathogens. Early studies examined compared *in vitro* transcriptional profiles of a laboratory reference strain under two experimental conditions. A classic application of this approach involves identifying gene expression signatures in biofilms by comparing the gene expression changes as bacteria transition from a motile to a sessile lifestyle (46). However, biofilm-specific gene expression profiles can vary significantly and depend on the experimental design of biofilm assay, such as media composition, static or dynamic cultivation conditions. Meta-analysis of early microarray data across studies highlights the substantial impact of the growth media and other environmental factors on the measured transcriptome (47). In addition to growth conditions, the variability of gene expression is also assessed across diverse strain backgrounds (48). Specifically, the variability of gene expression profiles among different *P. aeruginosa* strain backgrounds under the same growth condition appears to be smaller than that observed for PA14 under diverse environmental conditions. Those early transcriptomic studies concluded that *P. aeruginosa* transcriptome response is a function of the environmental conditions (49).

Recent advancements leverage the accessibility and high-throughput capabilities of RNA-seq to dissect bacterial function during human infection, providing new insights into bacterial physiology within the human host. The optimization of sample preparation techniques and advancements of NGS technologies now enable transcriptome analysis on bacterial mRNA directly from infection sites like human tissues and organs. Transcriptional profiling of *ex vivo* samples is considered the gold standard for studying bacterial phenotypic adaptations directly from infection sites. Transcriptome work on *ex vivo* samples were initially focused on some model systems, such as *S. aureus* from prosthetic joints and *E. coli* from urinary tract infection (50). Optimization of the sample preparation protocol has made it possible to extract intact bacterial mRNA from specimens of tissues and organs (51, 52). Recent research in *P. aeruginosa* infection has delved into *ex vivo* transcriptional profiles from CF sputum, chronic wounds, burn wounds, and directly from explanted chronically infected lungs from CF patients. All of these studies contained substantial reads (~95%) representing human transcripts, posing a challenge to *ex vivo* RNA-sequencing of bacterial data. Despite this, these three studies successfully acquired sufficient bacterial reads to produce a robust *ex vivo* transcriptome of P. aeruginosa. Notably, *Ex vivo P. aeruginosa* transcriptome in the respiratory tract revealed bacterial adaptations to oxidative stress and microaerophilic conditions. Furthermore, ex vivo transcriptional signatures also reflect the up-regulation of genes involved in iron acquisition and uptake system, and antimicrobial resistance. Moreover, *P. aeruginosa* population has shown heterogeneity in bacterial motility genes, with a subpopulation appeared non-motile exhibiting low expression of *ex vivo* motility-related genes (51-53).

With the recent developments of NGS method and increasing availability of RNA-seq data (54), there is now a holistic understanding of gene expression profiles in bacterial model systems under various environmental conditions. With the rapid increase of new bioinformatics tools, the focus of future transcriptomics analysis can be shifted from descriptive studies on gene-level to comprehensive studies that involves regulatory networks of co-regulated genes or functional pathways on a global scale.

1.3.4 Development of proteomics to understand bacterial physiology

The first genome sequence of *P. aeruginosa* strain PAO1 was completed in 2000 (11), and provided a global and comprehensive view of this pathogen and its functional capacities. Over the past two decades, we have learned a lot about this species by combining comparative genomics, transcriptomics, and bioinformatics. While genomics and transcriptome present blueprints of life, the proteins provide the real life of bacterial physiology. Interest in *P. aeruginosa* presents an attractive and valuable system for studying regulation of gene expression due to its versatile metabolic capabilities and our extensive knowledge of the genetics, biochemistry and physiology (55, 56).

Proteomic research is well-developed for certain bacteria, including well-studied organisms like *E. coli* (57-59), *B. subtilis* (60) and some pathogens like *M. pneumoniae* (57, 61, 62)*, S. aureus* (63), and *M. tuberculosis* (64). Early proteomic studies primarily focused on simpler model systems like *M. pneumoniae*, achieving a considerably high coverage and measuring 81% of the genomically predicted ORFs in a 2004 study (57). *M. pneumoniae* offers a simplified model for detection of the entire proteome, however, it is not as attractive as a model for physiological proteomics because it has limited regulation of gene expression. Specifically, *M. pneumoniae* has limited numbers of sigma factors and two-component regulatory factors. While *P. aeruginosa* provides a valuable system due to its versatile metabolic capacities, earlier work was primarily focused on the outer membrane and extracellular proteins (65, 66).

Subproteome vs total theoretical proteome

An ideal proteomic analysis aims to quantify all proteins synthesized by a cell and gain a complete and comprehensive understanding of bacterial cellular events and metabolic potentials. Unlike DNA arrays that contain all the blueprints of the entire genome, proteomics can only access parts of synthesized proteins. Such limitation can be mitigated, and majority of the proteome can be measured by combining multiple subproteomic fractions, such as membrane proteins, extracellular proteins, weakly neutral/weakly acidic/alkaline cytosolic proteins (55, 67).

From a physiological perspective, the identification of the entire proteome is also limited by physiology constraints because only part of genome is active at any time and leads to the expression of a subpopulation of proteins. Two classes of proteomes can be defined based on their physiological states, proteomes of growing cells (vegetative proteome) and those of non-growing bacteria (60, 68, 69). The proteome of the growing bacteria primarily consists of large proportions of housing-keeping proteins required for growth and reproduction. Another fraction of vegetative proteomes consists of proteins used for degradation of energy sources and synthesis of nucleotides and amino acids that dependent on the composition of the growth medium. In addition to cellular events required for bacterial growth, there are also fractions of proteome that depends on the growth rates, such as the DNA replication and RNA synthesis machinery, translational apparatus, and many anabolic reactions (70). In contrast, a large number of proteomic signatures of non-growing bacteria were found to respond to various stress or starvation stimuli. It indicates that non-growing cells experience oxidative stress, glucose starvation, heat stress or phosphate deprivation (69).

Together, the combination of various subproteomic fractions serves as physiological signatures of various cellular events, and remains a fundamental approach to capturing the full array of proteins expressed in bacteria.

1.3.5 Historical proteomic work to understand P. aeruginosa physiology

The colonization and persistence of *P. aeruginosa* during human infection is facilitated by the production of a variety of factors, including proteins present on the cell surface or secreted into the external environments. Membrane proteins and extracellular proteins are important for bacterial survival, proliferation and virulence (55).

Extracellular proteins of P. aeruginosa

P. aeruginosa, like many bacterial pathogens, utilizes proteins and other non-proteinaceous substances to promote pathogenicity, including extracellular pigments such as pyocyanin that are known to release reactive oxygen species and induce oxidative stress in host cells (71). Moreover, *P. aeruginosa* secretes a large spectrum of proteins to enhance pathogenicity. These proteins includes protease (lasA) that degrade and damage proteins on the surface of host cells and tissues and toxins (exoA) that induce necrosis of specific host cells (72). In addition to releasing extracellular proteins to the external environments, *P. aeruginosa* can directly inject extracellular proteins into host cells by type III secretion system, such as exoenzyme (exoS). Protein secretion in *P. aeruginosa* is primarily observed during or near the stationary phase, with numerous virulence factors controlled by quorum-sensing systems (73). Extracellular proteins are relatively easy to isolate and solubilize due to their hydrophilic nature.

Historical work has been undertaken to study quorum sensing (QS) mutants using extracellular proteome (65, 66). The QS system (lasI/lasR and rhlI/rhlR) in *P. aeruginosa* regulates almost 5% of the genome, including various virulence factors (74, 75). Virulence factors can be regulated by either system in the absence of the other or can be solely regulated one one regulon. Those virulence factors regulated by QS include LasA protease (lasA), elastase (lasB), alkaline phosphatase (phoA), hemolysis and other non-protein extracellular molecules, such as pyocyanin (75, 76). After identification of the QS-regulated genes, proteomics analysis of multiple mutants consisting of one or multiple gene depletion can be utilized to understand their effects on the expression on other QS-regulated genes.

Membrane proteins of P. aeruginosa

Membrane proteins and surface proteins are vital for bacterial survival, proliferation, and virulence and have been widely studied use diverse approaches in bacteria. Studying these proteins using proteomics to study those proteins offer insights into bacterial physiology. However, isolating membrane proteins from *P. aeruginosa* can be challenging due to their nature of amphipathic, having hydrophobic transmembrane spanning regions and hydrophilic regions. It is difficult to solubilize membrane proteins because the ionization efficiencies and peptide recoveries vary broadly for hydrophobic transmembrane-spanning helices and soluble tail/loop peptides (77). Notably, these *P. aeruginosa* proteins play crucial roles in processes such as adhering to surfaces (pili), removing antibiotics using efflux pumps (MexAB-OprM), and serving as receptors and channels for essential nutrients (eg. Sugar-transporting porins and vitamin B12 receptor) (78, 79).

Early proteomic studies of *P. aeruginosa* proteome focused on membrane proteins. In those analysis of outer membrane proteins, they have identified 104 unique ORF of 189 characterized proteins, consisting of three major categories, receptors, porins, and protein of unknown function (80, 81). Those expression of porins proteins indicates the presence of specific nutrients in the growth environment (82, 83). For example, the expression of OprD shows that the presence of the amino acids like arginine and histidine as the sole carbon or nitrogen sources influences its expression. As a non-specific porin, oprF is the major pseudomonal outer membrane proteins and have diverse function, including maintenance of cell shape as a structural protein and association with the peptidoglycan. Several receptors identified in early proteomic studies turns out to be pyochelin, pyoverdine, and aerobactin. The expression of these proteins can serve as indicators of iron-starvation in the media (84, 85).

Post-translational modifications

Post-translational modifications (PTMs) play a critical role in protein functioning because they impact protein stability, protect from proteases, and conduct signal transduction. Bacterial proteins undergo various potential modification like phosphorylation, methylation, glycosylation, deamidation, glycosylation and biotinylation. PTMs are generally regulated by specific enzymes but it can also occur non-enzymatically. First described in two-component regulatory systems, phosphorylation can happen on a response regulator vis histidine kinase and lead to activation of a serious of genes and proteins (86). In *P. aeruginosa*, the first proteome analysis to determine the overall level of protein phosphorylation was performed in PAO1 using phosphopeptides enrichment methods and Nanoscale liquid chromatography coupled to tandem mass spectrometry (nanoLC-MS/MS) analysis. This study identified 55 phosphorylation sites on serine, threonine, and tyrosine residues and 23 phosphorylated proteins (86). A separate study has investigated the phosphorylation of all proteins (phosphoproteome) of PA14 using both intracellular and extracellular subproteomic fractions. In PA14, 131 phosphorylated sites and 81 phosphoproteins were identified (86, 87). Surprisingly, a very low amount of those phosphorylated proteins is shared among two *P. aeruginosa* strains PAO1 and PA14 (6 proteins, phosphomannomutase *AlgC*, succinyl-CoA synthetase, isocitrate dehydrogenase, phosphoenolpyruvate synthase, hypothetical proteins PA14_20770 and PA14_02130).

Protein glycosylation is another important PTM in bacteria, but it is less studied due to these challenges in analysis: 1) glycosylated proteins are lowly abundant compared to the unmodified proteins; 2) high heterogeneous in moiety (88). For example, glycosylation happens in bacteria adhesions such as flagellins with different moiety. Type-A flagellin were modified at a rhamnose residue with an O-linked heterogenous glycan consisting of up to 11 monosaccharide units (89), reflecting the heterogeneity of the glycosylation modification. However, type-B flagellin were modified with a glycan moiety linked through deoxyhexose residual (90).

1.4 Polymicrobial model system: Oral Microbiota[1]

Oral diseases like caries and periodontitis affect the majority of adults in the United States and are two of the most prevalent infections worldwide (91, 92). The oral cavity, characterized by diverse microenvironments influenced by dramatic nutrient and temperature changes, regular disturbances (such as hygiene practices), and host interactions, provides a valuable model system for studying biofilm development, ecological succession, and the eco-evolutionary dynamics (93-96). Consequently, research on oral diseases has spurred more than a century of innovative studies at the interface of pathogenesis and microbial ecology. The oral microbiota presents a crucial model for advancing our knowledge of microbial communities and their interactions.

1.4.1 Discovery of polymicrobial nature

For as long as we have known that microbes exist, we have recognized the polymicrobial nature of the oral microbiota. Dating back to the discovery of microorganisms in 1683, Antonie van Leeuwenhoek observed a tartar specimen from his tooth using a primitive microscope (97), describing diverse bacterial morphologies now recognized as cocci, spirochetes, and fusobacteria (98). As microbiology research advanced in the late 1800s and early 1900s, scientists aimed to understand the roles of individual microbes and the collective community as the causative agents of oral diseases (99-101). Initially focusing on easily cultivable aerobes and facultative anaerobes like

[1] Parts of this subsection were adapted from the following reference: Zhang M, Whiteley M, Lewin GR, 2022. Polymicrobial Interactions of Oral Microbiota: a Historical Review and Current Perspective. mBio 13:e00235-22. DOI: 10.1128/mbio.00235-22. Copyright © 2022 Zhang et al. Reused with permission. I was the primary author of this work.

Streptococcus (101) and Actinomyces (102, 103), the cultivation challenges posed by a variety of oral bacteria were eventually overcome through the development of anaerobic cultivation techniques for obligate anaerobes and the refinement complex media to enhance the growth of fastidious bacteria. The successful growth of diverse oral bacteria in laboratory settings enabled detailed genotypic and phenotypic characterization of individual species and laid the foundations for polymicrobial studies.

1.4.2 Foundations of oral bacterial colonization, a spatiotemporal manner

Research on microbial interactions within the oral microbiota traces back to 1970 when Gibbons and Nygaard observed clumping sedimentation within seconds after mixing two pure cultures of different species, initially known as interbacterial aggregation (104). Now termed coaggregation, this phenomenon of cell-cell interactions was first described in the oral microbiome but is prevalent across microbial communities, including the human gut, urogenital tract, and freshwater environments (105). Coaggregation of oral microbes with each other and adherence to the tooth surface are crucial for the formation of oral biofilms (plaque) and were extensively characterized even before the introduction of the term "biofilm" in 1978 (106).

To further understand how these interactions influence bacterial colonization of the oral cavity, researchers physically removed mature human oral biofilms and monitored bacterial reappearance on the tooth surface over time. The temporal order of bacterial species appearance, combined with known coaggregation partnerships, enabled Kolenbrander to propose a spatiotemporal model of oral bacterial colonization in 1993 (**Figure 1**) (107-109). Following a 4-hour timeframe, the microbes repopulating cleaned teeth and adhering to host surfaces were recognized as "early colonizers" (109). Subsequently, species colonizing at later time points, spanning from hours or days to months or years, were termed as "late colonizers." While coaggregation is key to linking these species together, Kolenbrander noted that early colonizers do not bind to late colonizers and thus proposed the existence of a bridge organism that coaggregates with both early and late colonizers (110, 111). The genus Fusobacterium rarely isolated within the first 12 hours after professional teeth cleaning but is frequently found in both healthy and diseased dental plaques and can bind most genera of early and late colonizers. Thus, Fusobacterium is depicted as a coaggregation bridge in the oral microbiome (**Figure 1**) (109).

The ability of oral bacteria to bind to the tooth surface and coaggregate are critical for oral bacterial survival in the oral cavity, where they face challenges such as the continuous flow of saliva, regular oral hygiene practices, fluctuating nutrient availability, and drastic temperature changes. Additionally, adhesion, coaggregation, and environmental differences across mouth surfaces create multiple biogeographical niches in the oral cavity (107, 112). While the development of biofilms is important in various ecosystems and was first chronicled in aquatic systems in the 1930s, the oral biofilm stands out as an accessible and clinically relevant biofilm system (113, 114). Research on the oral biofilm has played a pivotal role in advancing our understanding of biofilm formation, microbe-microbe interactions, and the interplay between biofilms and pathogenesis (107).

1.4.3 Polymicrobial species associations in vivo

Although the physical interactions in the oral biofilm were widely studied by coaggregation, microbial interaction work specific to plaque from below the gumline (subgingival) was limited until Socransky and his colleagues performed a comprehensive and impactful study in 1998. These authors studied over 13,000 human subgingival biofilm samples from periodontally healthy and diseased patients using checkerboard DNA-DNA hybridization techniques and then grouped 40 culturable bacterial species into six "colored" complexes using cluster analysis and community ordination techniques (**Figure 1**; **Figure 2**) (115). For example, the bacteria in the red complex, *Porphyromonas gingivalis*, *Treponema denticola*, and *Tannerella forsythia* (formerly Bacteroides forsythus), were strongly associated with each other and strongly correlated with periodontal disease. Among the colored complexes, this red complex has attracted significant interest because all three members are putative periodontal pathogens, while other complexes contain both commensals and pathogens (116). The proposal of a red complex contradicted the prevalent idea that disease resulted from the presence of a single pathogen and instead supported the alternative hypothesis that periodontal diseases are associated with communities of bacteria. In addition to identifying the association among species within each complex, Socransky et al.'s study identified the relationship among different complexes. For instance, this work showed that the species in the red complex were rarely found without the presence of orange complex bacteria. This study on subgingival species associations *in vivo* was pioneering and comprehensive, and a similar study was performed using plaque from above the gumline (supragingival) in 2008 (117). While these studies only identified known, culturable bacteria and were constrained by the limitations of correlative studies, they provide important context for future research to study complexes beyond the species level (i.e., strain level or genus level) and to incorporate the role of phenotypic plasticity, including virulence factor production and other responses to host factors (118).

1.4.4 *Probing community composition and function*

The development of next-generation sequencing techniques and bioinformatic analytic tools allows researchers to study the complexity of diverse oral communities without cultivation, although these studies are challenging due to limited bacterial genetic material in clinical samples and disagreement on the best analysis approaches. The application of 16S rRNA gene amplicon sequencing and metagenomics allows differentiation of bacteria at the species or even strain level and determination of their functional potential.

Work using these techniques has revealed enormous genetic diversity in the human oral cavity: over 700 bacterial species are found in the mouth across individuals, with approximately 200 species per person. Extensive work has resulted in the ability to culture 74% of these microbes (119-121). For instance, the candidate phylum TM7, now Saccharibacteria, was discovered 20 years ago using 16S rRNA gene sequence analyses (122), eventually leading to the recent successful cultivation of this microbe with its obligate host bacteria in lab conditions (123, 124). Thus, the initial sequencing analyses led to experimental characterizations and the finding that Saccharibacteria downregulates the pathogenicity of its host bacteria and reduces inflammatory bone loss (95).

While the metagenome shows the possible microbial function of the community, it does not reflect the community's actual activity. A full understanding of oral microbiome dynamics requires a combined picture of microbial composition (the metagenome) and global gene expression (the metatranscriptome). Importantly, the accessibility of the human oral cavity allows for the instantaneous preservation of microbial RNA with temporal and spatial control. Studies using metatranscriptomics on oral specimens have comprehensively shaped our understanding of the oral microbiome. These studies found that although the community composition during periodontitis can vary drastically, the community's functional activities are conserved (125). This finding suggests that rather than focusing on specific pathogens, we should consider the community as a pathogen (126). Additional support for "the community as pathogen" comes from community-wide metatranscriptomic studies that found organisms traditionally considered commensals transcribe the majority of virulence factors during periodontal diseases (7, 127). Metatranscriptomics can also provide insight into the in situ functionality of individual species (128). One striking finding is that a species can have similar relative abundance but different functional activities across different conditions. For example, F. nucleatum does not significantly change in relative abundance between healthy and diseased samples, but metatranscriptomics shows that its metabolism is altered (129).

The oral cavity has become one of the best-described microbial sites in the human body due to its accessibility and clinical relevance; sequencing-driven studies of oral health and disease have broadened our understanding of the community composition and functional activities of the oral microbiota. However, these studies are only starting points, and the driving factors that lead to the progression from healthy to diseased states remain unknown (**Figure 3**, Table 1). Longitudinal studies could provide clues about initial stages of oral diseases and markers of disease progression (127, 130). Finally, there is a further need to simultaneously measure microbial and host gene expression, which could partially explain the development of disease or the shift back to health after therapeutic interventions. These future directions are important for developing therapies to prevent and treat oral infections and for broadly understanding host-associated microbial communities.

1.4.5 Microbiogeography

To understand microbe-microbe and microbe-host interactions, it is additionally important to determine the spatial organization of bacteria with each other and with host factors at scales of microns to hundreds of microns (**Figure 2; Figure 3**). The advent of fluorescence in situ hybridization (FISH) and immunofluorescence allowed researchers to distinguish the identities of microbes in complex oral communities (131). By use of combinatorial labelling and spectral imaging-FISH (CLASI-FISH) (132), the corncob structure was redefined by visualizing nine genera simultaneously in supragingival plaque. The structure was more complex than previously identified: multiple corncobs together formed "hedgehog structures" with aerotolerant taxa on the outside and anaerobic taxa on the inside toward the tooth surface (**Figure 2**) (133). This pioneering study suggested differential oxygen and nutrient usage across the plaque biofilm and provided a spatial framework to incorporate metabolic and ecological factors. Further work quantitatively assessed the proximity of microbes relative to host factors through computing microbial pairwise correlation functions using CLASI-FISH imaging of the tongue dorsum (134). These studies provide important in situ benchmarks to experimentally test hypotheses about the spatiotemporal development of the oral biofilm (112, 135).

There are further opportunities to use imaging approaches such as FISH-based techniques to study disease progression and community dynamics in the oral cavity. In vitro applications of CLASI-FISH can distinguish up to 120 different species in a single image (50), and high-phylogenetic-resolution microbiome mapping by FISH (HiPR-FISH) can distinguish over 1,000 closely related bacterial strains in vitro and has been applied to study oral microbial spatial patterning across longitudinal samples (136). Finally, as metatranscriptomics showed the importance of measuring bacterial function in addition to community composition, there are increasing opportunities to measure bacterial physiology at the micron level (135, 137, 138).

Table 1. Open questions and approaches in oral microbial ecology.

Questions	Bottom-up approaches	Top-down approaches
1. How to build a simplified model system for oral microbiota? What are appropriate metrics for model evaluation?	Experimental model systems can be constructed at the species level (96, 107), genus level (96, 133), or functional level.	Quantitative approaches must be used to benchmark models using the human oral community (128, 135, 139).
2. What are the similarities and differences in community dynamics between supragingival and subgingival plaque?	Models can illuminate differences in microbial interactions between the two environments. For instance, the Zurich model suggested that the subgingival plaque model could be derived from the supragingival plaque model (140).	Comparative studies using sequencing and microscopy can show intra-patient and inter-patient differences between environments.
3. What are the factors that drive the progression from healthy to diseased states or from diseased to healthy states? How do oral microbes colonize and invade different oral habitats?	Model systems can test the factors that lead to pathogen abundance and the production of virulence factors.	Detailed analyses of longitudinal studies in human patients can further show how communities change over time and which healthy communities become diseased.
4. How many microenvironments exist in the oral cavity? How do microenvironments impact biodiversity?	Perturbation of laboratory models can test the importance of different environmental factors.	*In situ* measurement of chemical gradients and oxygen levels can indicate different niches. Also, differential abundance of microbes can indicate site-specialists with distinct niches (112).

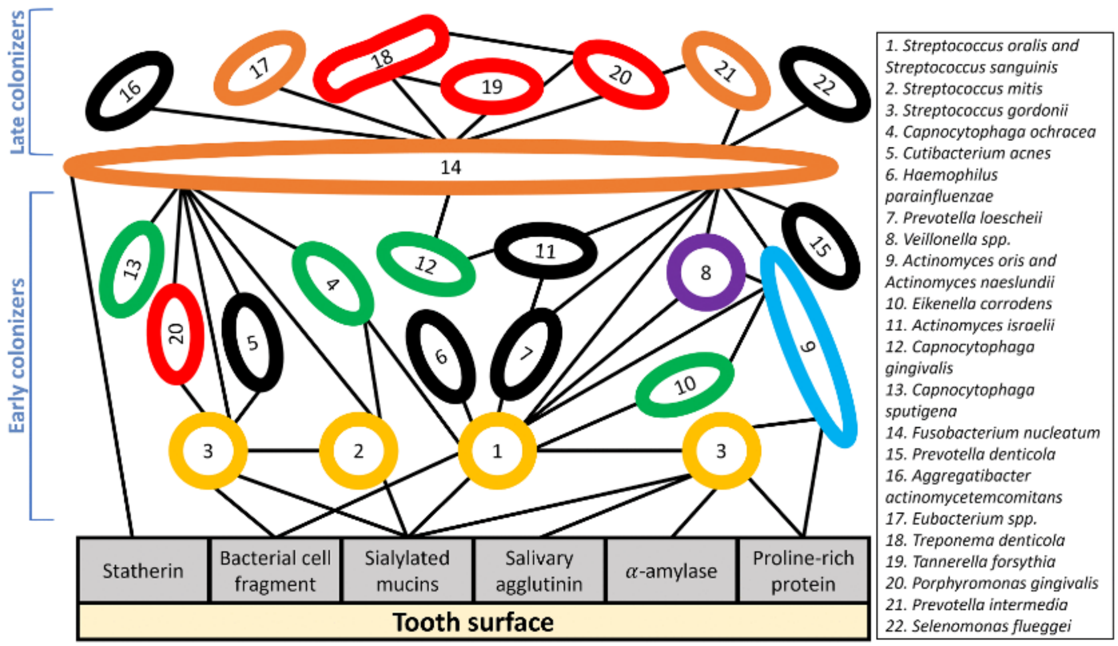


Figure 1. Spatiotemporal model of oral bacterial colonization and pathogenicity.

A schematic of the Kolenbrander model of long-term microbial succession from the early stage of biofilm colonization on the tooth surface to the establishment of mature supragingival and subgingival biofilms and, ultimately, to the formation of diseased bacterial communities (107-109). Rods and circles indicate microbial taxa, and lines indicate physical interactions, including binding to the tooth surface or known coaggregation. This schematic integrates over 1,000 coaggregation connections found in the oral cavity, involving microbes that are primarily found both in supragingival and subgingival plaque. The colors of the microbial taxa indicate their corresponding Socransky complex from subgingival plaque, which consists of six categories, yellow, green, blue, purple, orange, and red (115). Specifically, orange and red complexes are more often associated with clinical parameters of gum disease. Species not covered in this Socransky model are colored in black. This schematic highlights that early colonizers and late colonizers are classified into different-colored complexes (141-143) and the proposed importance of F. nucleatum as a bridge species in linking early and late colonizers. Although this model has been highly influential on our understanding of biofilm formation, current work employing advanced microscopy and sequencing techniques continues to refine our understanding of oral biofilm biogeography and development.

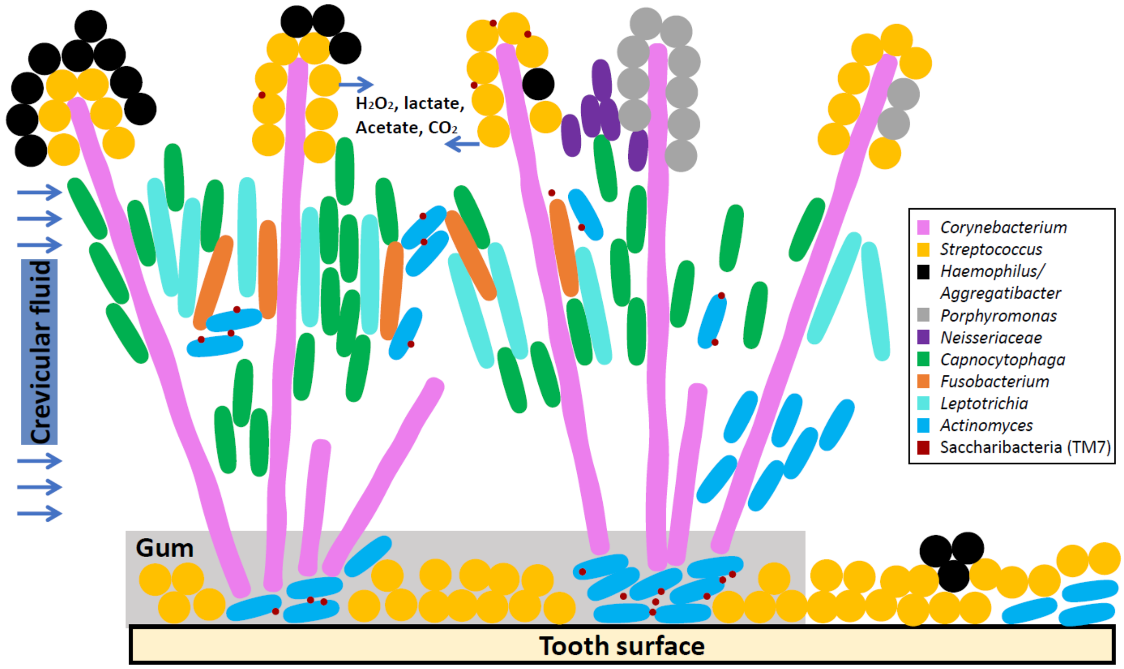


Figure 2. The supragingival oral biofilm architecture observed using CLASI-FISH, incorporating proposed biochemical gradients and episymbiotic Saccharibacteria

In this diagram, the rods and circles indicate microbial taxa, and their locations are based on microscale imaging of supragingival plaque using CLASI-FISH (95, 133). The 'hedgehogs' are structured by clusters of *Corynebacterium* filaments that bind to *Streptococcus* and *Actinomyces* near the base and then expand to the 'corncob'-structured perimeter. This spatial patterning also divides the environment into different chemical environments, as shown. The bacterial genera are colored with their corresponding colored complexes from mature supragingival biofilms: *Streptococcus spp.* (yellow), *Neisseriaceae spp.* (purple), *Capnocytophaga spp.* (green), *Fusobacterium spp.* (orange), and *Actinomyces spp.* (blue) (117). Other taxa, including Saccharibacteria, not included in the Socransky complexes are shown in other distinct colors. Note that *Porphyromonas* is not included in a colored complex as *Porphyromonas* here is likely aerotolerant *Porphyromonas catoniae* and/or *Porphyromonas pasteri* (133).

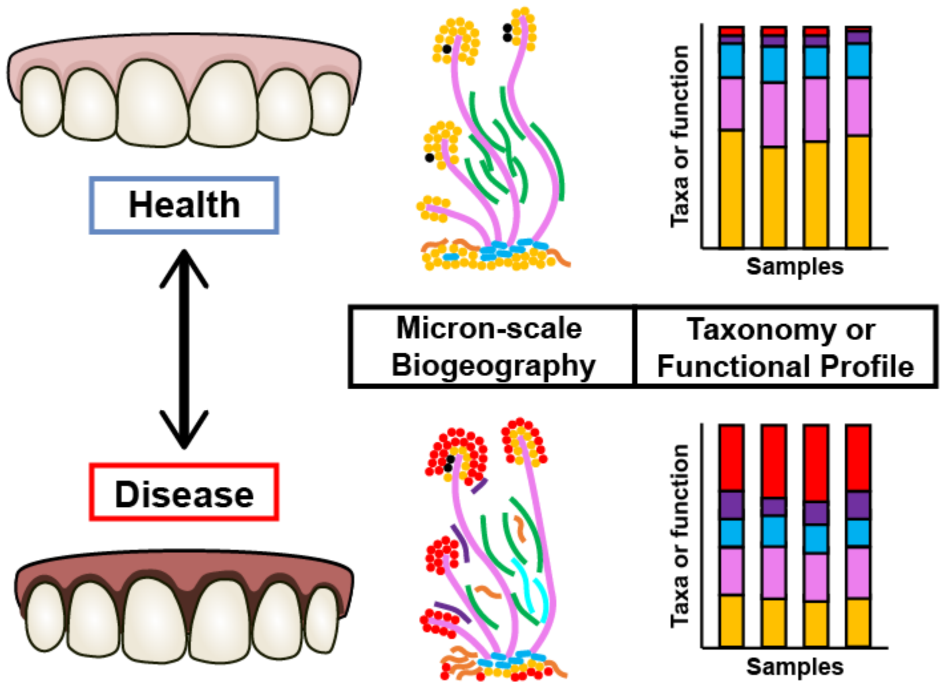


Figure 3. Approaches to study the dynamics of the oral microbiota during health and disease.

Microscopy and sequencing-based community profiling are powerful approaches that can be leveraged for spatiotemporal studies of oral microbial ecology to further understand the relationship of the oral microbiota with health and disease. Confocal scanning microscopy imaging of labeled strains reveals changes in micron-scale biogeography and the corresponding changes in microbial interactions. Community profiling, for instance using metagenomics and metatranscriptomics, shows changes in the composition and functional activities of samples across space or time. These methods are important in both top-down and bottom-up approaches. For example, top-down approaches could sample the oral biofilm over time after a professional cleaning (144). For bottom-up approaches, emergent spatiotemporal dynamics can be observed using a small number of cells directly removed from oral specimens with micromanipulators or using communities constructed from pure cultures of strains (145).

CHAPTER 2. MRNA-PROTEIN RELATIONSHIPS IN *P. AERUGINOSA*[2]

2.1 Background and significance

A major challenge in understanding the mechanisms controlling colonization and persistence of bacterial pathogens such as *P. aeruginosa* is a lack of knowledge regarding their physiology during human infection. In recent studies, we have leveraged the accessibility and high-throughput nature of RNA sequencing (RNA-seq) to profile bacterial function during human infection, providing new insights into bacterial physiology in the human host (7, 125, 128-130, 139, 146-148). However, as protein, not mRNA, is the primary functional component of a bacterial cell, these transcriptome studies are reliant on the limited assumption that mRNA levels are predictive of protein abundance.

The advent and widespread availability of high-throughput mRNA and protein quantification has allowed for investigation of the correlation between mRNA and protein levels across different phylogenetic taxa, including bacteria, plants, and humans (149-163). These studies have revealed a positive correlation between mRNA and protein

[2] Parts of this subsection were adapted from the following reference: Zhang M, Michie KL, Cornforth DM, Dolan SK, Wang Y, Whiteley M, 2023. Impact of Growth Rate on the Protein-mRNA Ratio in Pseudomonas aeruginosa. mBio 14:e03067-22. DOI: 10.1128/mbio.03067-22. Copyright © 2022 Zhang et al. © 2022 Zhang et al. Reused with permission. I was the primary co-first author of this work.

across kingdoms, with genome-wide Spearman rank correlation coefficients ranging from 0.45-0.65. However, in eukaryotes, particular subsets of genes have been shown to be more or less predictive of protein levels (154, 164, 165). For example, cell cycle-regulated genes have a weaker correlation in human cells (166), while in yeast, cell cycle and nucleolar-localized genes display higher correlation (167). Compared to eukaryotes, investigation of mRNA-protein correlations in bacteria is somewhat limited (150-153, 156, 158, 159, 168). Two previous studies quantified protein and mRNA abundance for *P. aeruginosa* grown in test tubes, revealing correlations of 0.64-0.65 (168, 169). However, this and other bacterial studies have been limited to a single growth environment and have not explored mRNA-protein correlations of subsets of genes.

Bacterial physiology is primarily studied in vitro conditions, and the knowledge gained is biased towards fast-growing bacteria. However, this approach oversimplied bacterial behaviors in their natural environments, where growth rates are dynamic and can vary significantly (32-35). Research on chronic infections, like *P. aeruginosa* in cystic fibrosis patients, demonstrates that growth rates can range from less than an hour to over 24 hours (33),

Our understanding of the basic aspects of bacterial physiology, including mRNA-protein correlations, is largely based on *in vitro* growth and is thus highly biased towards extremely fast-growing bacteria. Not only are fast growth rates not generally the norm in the natural environment, but growth rate is also highly dynamic and can vary widely within a population (32-35). Indeed, previous studies revealed that the growth rate of bacteria in the human body during chronic infection can vary widely, depending on the nutritional environment of the infection site and the immune response. For example, the doubling time of *P. aeruginosa* chronically infecting the lungs of people with cystic fibrosis (CF) has been shown to vary from less than 1 hr to greater than 24 hrs (33), with population averages between 2.5 and 4.6 hrs (35). Thus, there is a gap in knowledge as to what extent changes in growth rate affect basic bacterial physiology, including the correlation of mRNA and protein. Here, we address this knowledge gap by assessing the impact of growth rate on *P. aeruginosa* mRNA-protein correlation by performing high-throughput mRNA and protein quantification of bacterial grown at four growth rates in chemostats. We discovered that while growth rate did not impact the overall correlation of mRNA and protein, there were subsets of genes that exhibited higher correlation at all growth rates. In addition, we developed and employed an RNA-to-protein correction factor that allows for more accurate prediction of protein levels of lowly correlated genes, thus increasing the biological utility of human infection transcriptomes from *P. aeruginosa* and other bacterial pathogens.

Significance statement

Our understanding of bacterial physiology during human infection is limited by the difficulty in assessing bacterial function in the infection site. Recent studies have begun to address this question by quantifying bacterial mRNA levels in human-derived samples using transcriptomics. One challenge for these studies is the poor predictivity of mRNA for protein levels for some genes. Here, we addressed this challenge by measuring transcriptomes and proteomes of P. aeruginosa grown at four growth rates. Our results revealed that growth rate doesn't impact the genome-wide correlation of mRNA and protein. We used statistical methods to identify genes in which mRNA and protein were poorly correlated and developed an RTP conversion factor that improved predictivity of protein levels across strains and growth conditions. Our results provide new insights into mRNA-protein correlations and tools to enhance understanding of bacterial physiology from transcriptome data.

2.2 Impact of growth rate on *P. aeruginosa* transcriptomes and proteomes

2.2.1 Growth of P. aeruginosa at four growth rates

To evaluate how growth rate impacts mRNA and protein levels, we grew *P. aeruginosa* strain PA14 in carbon-limited chemostats. Chemostats were used as they allow continuous-culture growth of a planktonic bacterial population at steady state. The generation (doubling) time and bacterial density are precisely controlled, thus providing the opportunity to study the impact of growth rate on *P. aeruginosa* mRNA and protein levels in the absence of confounding variables often used to control growth rate (e.g. different growth media and different temperatures). A defined, minimal medium containing succinate as the sole carbon and energy source was used (170), and bacteria were cultured in chemostats with generation times of 3, 6, 14, and 25 hrs at a density of 7×10^8 cells/ml. For each generation time, bacterial cells were harvested for transcriptomic and proteomic analyses at steady-state.

2.2.2 Comparison of the transcriptomes of fast and slow growing P. aeruginosa

To begin to assess the impact of growth rate on the *P. aeruginosa* transcriptome, we first compared gene expression in the fastest (3 hr) and slowest (25 hr) growing chemostat cultures. Transcripts from 5894 genes were detected from cells grown at both growth rates. 2362 genes were differentially expressed ($p_{adj} < 0.05$) when comparing the slowest and fastest growth rate, indicating that growth rate has a tremendous impact on *P. aeruginosa* gene expression. Enrichment analysis of differentially regulated genes using the TIGRFAM functional gene annotation database categories (171) revealed no enrichment of any category, indicating a global change in gene expression (Fisher's exact test). Using a more stringent cutoff, (fold change ≥ 4, $p_{adj} < 0.05$), 343 genes were differentially expressed, with 27 genes upregulated and 216 downregulated at the faster growth rate. Differentially regulated genes included a large number of operons that have been functionally characterized, including three gene clusters encoding components of type VI secretion systems, two of which (H1-T6SS and H2-T6SS) were up-regulated and one (H3-T6SS) that was down-regulated at the faster growth rate. Another operon up-regulated at the faster growth rate was *pqsABCDE,* which encodes proteins required for the biosynthesis of quinolines that have antimicrobial and quorum signaling activities. Operons downregulated at the faster growth rate include those involved in glycogen biosynthesis, type III secretion, nitrate metabolism, sulfur metabolism, and branched chain amino acid metabolism. Operons encoding two low affinity terminal oxidases were also differentially expressed, including upregulation of the bo_3 quinol oxidase (*cyoABCDE*) and downregulation of the aa_3 cytochrome c oxidase (*coxA, coxB, PA14_01310, coxC/coIII*) at the faster growth rate.

2.2.3 Comparison of the proteomes of fast and slow growing P. aeruginosa

In addition to *P. aeruginosa* transcriptome, we then compared protein levels in the fastest (3 hr) and slowest (25 hr) growing chemostat cultures to assess the impact of growth rate on the *P. aeruginosa* proteome. 3903 proteins were identified in cells from both the fastest and slowest growing *P. aeruginosa* chemostat cultures, with 2023 of these proteins differentially produced ($p_{adj} < 0.05$). Enrichment analysis of differentially regulated proteins using TIGRFAM categories revealed an enrichment of several 'main' categories (171), including protein synthesis and several metabolic categories, as well as the 'subrole' categories of ribosomal proteins and tRNA aminoacylation. Overall, most proteins showed a small magnitude of change, with only 391 proteins having a fold change greater than 2. Of these proteins, 147 were upregulated and 244 downregulated at the fastest growth rate. Five of the ten most differentially increased and six of the ten most differentially decreased proteins are hypothetical. As observed in the mRNA comparison, many of the proteins were encoded by operons that have been functionally characterized. Proteins with increased levels at the faster growth rate included 55 components of the 30S or 50S ribosomal subunits. Proteins with decreased levels at the faster growth rate included the aa_3 cytochrome c oxidase, branched chain amino acid metabolism, and proteins involved in phenazine biosynthesis. These results reveal that similar to mRNA, changes in growth rate have a tremendous effect on the *P. aeruginosa* protein composition.

2.2.4 Overlap of differentially expressed genes and proteins

We next asked, of the genes and proteins that are differentially expressed between the fastest and slowest growth rates, how many are shared? We focused on the 3903 genes in which both mRNA and protein were detected for each growth rate. Genes were considered to overlap only if the protein and mRNA levels changed in the same direction (up- or down-regulated). Of the 2697 genes that showed differential mRNA or protein levels ($p_{adj} < 0.05$), 1030 showed overlapping changes, including 558 that were up-regulated and 472 that were down-regulated (**Figure 4**; **Figure 5**). Enrichment analysis of these shared genes revealed enrichment of three TIGRFAM main categories: protein synthesis, amino acid synthesis, and purine/pyrimidine/nucleotide/nucleoside synthesis. Using a more stringent cutoff for differential regulation (RNA fold change ≥ 4, protein fold change ≥ 2, $p_{adj} < 0.05$) revealed 98 shared genes between the transcriptomes and proteomes with 68 genes downregulated, and 30 genes upregulated (**Figure 4**B). Many of the shared genes were expressed in operons and include the cytochrome c oxidase aa3 (*coxA, coxB, PA14_01310*), glycogen biosynthesis (*glgB, glgP, PA14_36730, PA14_36740*) and sulfur metabolism (*PA14_19540, ssuB, ssuD, PA14_19590*).

2.2.5 Assessment of mRNA and protein levels across four growth rates

The fact that we quantified gene expression at four growth rates provides the unique opportunity to assess gene expression across a range of growth rates. Thus, we next assessed expression of the 2362 genes differentially expressed at the fastest and slowest growth rates across all growth rates, specifically focusing on whether these genes showed monotonic expression as growth rate increased. Using statistical methods, we identified 424 genes whose expression increased and 317 genes whose expression decreased monotonically with increasing growth rate (**Figure 6**A and C). The majority of genes that showed increased mRNA levels with increasing growth rate encode functions important for cell growth, including energy metabolism (F_0F_1 ATP synthase: *atpA, atpD, atpE, atpF, atpG, atpH*), oxidoreductase reactions (*nqrB, nqrC, nqrF*), amino acid metabolism (arginine biosynthesis: *argB, argF, argG, argH*), metabolism of cofactors and vitamins (porphyrin metabolism: *cobB, cobC, cobI, cobJ, cobM, cobN, cobQ, cobT, cobU*), and fatty acid biosynthesis (*fabA, fabB, fabF1, fabG, fabZ*). Expression of a large number of genes involved in translation also increased expression with increasing growth rate, including genes encoding 50S ribosomal proteins (*rplI, rplJ, rplK, rplL, rplW, rpmE, rpmF*), 30S ribosomal proteins (*rpsE, rpsF, rpsI, rpsL, rpsN, rpsQ, rpsU*), and genes responsible for aminoacyl-tRNA biosynthesis (*alaS, glyQ, glyS, gatA, gatC, glnS, fmt, ileS, lysS, serS, thrS, trmD, valS*). 118 genes (37.2%) that showed a monotonic decrease in mRNA levels with increasing growth rate encode putative proteins of unknown function (**Figure 6**). The remaining genes encode proteins involved in chemotaxis (*cheA, cheB, cheW*), nitrate metabolism (*napA, napB, napC, napD*), the aa_3 cytochrome c oxidase (*coxA, coxB, PA14_01310*), and type III secretion (*pcrG, pcrV*, *pscB, pscC, pscD, pscJ, pscK*).

We next asked if genes that showed monotonic mRNA expression across all growth rates also showed monotonic changes in protein levels. Only 54 of the 424 genes whose mRNA increased monotonically with increasing growth rate also displayed monotonic increases in protein levels (**Figure 6**B). These genes encoded functions including small organic acid transport (*dctA*), amino acid metabolism (*trpB, lysC, lysA, hisC2, purB*), lipid metabolism (*fabA*), nucleotide metabolism (*purB*) and ribosomal proteins (*rplI, rplJ, rplK, rpmE, rpsI, rpsL, rpsN, rpsQ*). Only 72 of the 317 genes whose mRNA decreased with increasing growth rate also displayed monotonic decreases in protein abundances (**Figure 6**B), and 31 of these genes encode proteins of unknown function. Genes with known or putative functions included some involved in secretion (*pscC, pscJ, hcp3, hsiJ3*), carbohydrate metabolism (*pslG, pslI, atoB, bkdA1, glgB, prpD*), lipid metabolism (*glpK, pmtA, fadE),* nitrate metabolism (*napC*), amino acid metabolism (*arcC, aspC*), translation (*fusA2*) and other functions (*algR, exsB, osmC*, *osmE*). Our results reveal that while there are large numbers of genes whose mRNA levels change monotonically across four growth rates, only 13% of these genes also showed monotonic changes in protein levels.

We also developed statistical methods to discover genes whose expression followed more complex shapes. 85 genes had high expression levels during both the fastest and slowest growth rates but have low gene expression during the moderate growth rates, a pattern which we termed “U-shape” (**Figure 7**A). Many of these genes encode proteins involved in siderophore biosynthesis and pyochelin biosynthesis (*pchDCBA, pchEFG*) and were found to be organized in operons. 93 proteins were shown to follow a U-shape pattern (**Figure 7**C), but only 4 of these genes overlapped with those that showed U-shape mRNA levels. We also identified 242 genes that have low expression levels during both the fastest and slowest growth rates but have high gene expression during moderate growth rates, termed “upside down U-shape” (**Figure 7**B). No proteins showed an upside-down U-shape pattern (**Figure 7**D). These data indicate that while there are over 300 genes that show more complex U-shape expression patterns, U-shape protein patterns are very rare.

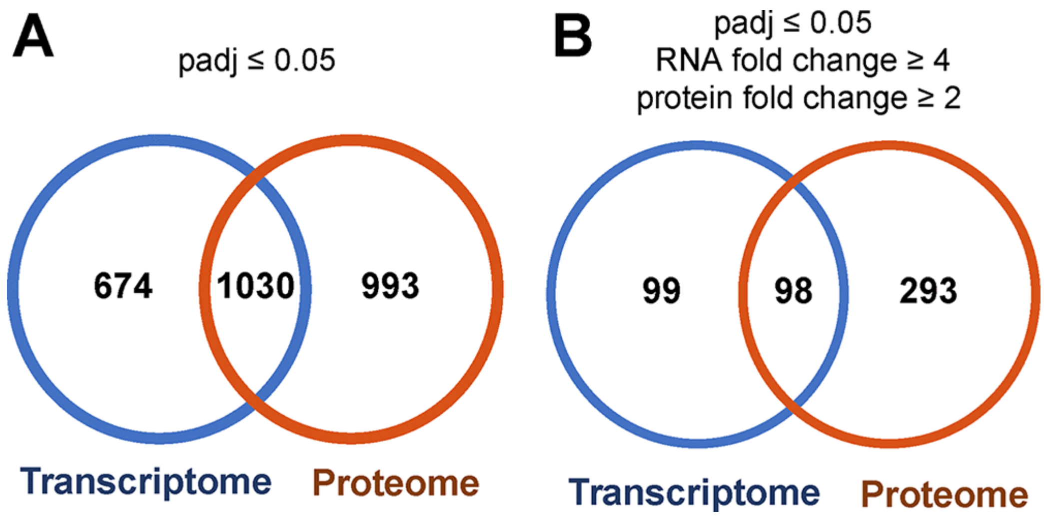


Figure 4. Differentially expressed mRNAs and proteins show modest overlap.

Venn diagrams show the overlap between differentially expressed mRNAs and proteins among fast-growing (3-h doubling time) and slow-growing (25-h doubling time) *P. aeruginosa* cultures. (A) Overlap of differentially produced mRNAs and proteins using a P_{adj} value of <0.05. (B) Overlap of differentially produced mRNAs and proteins using a more stringent cutoff (RNA fold change of ≥4, protein fold change of ≥2, and P_{adj} of <0.05).

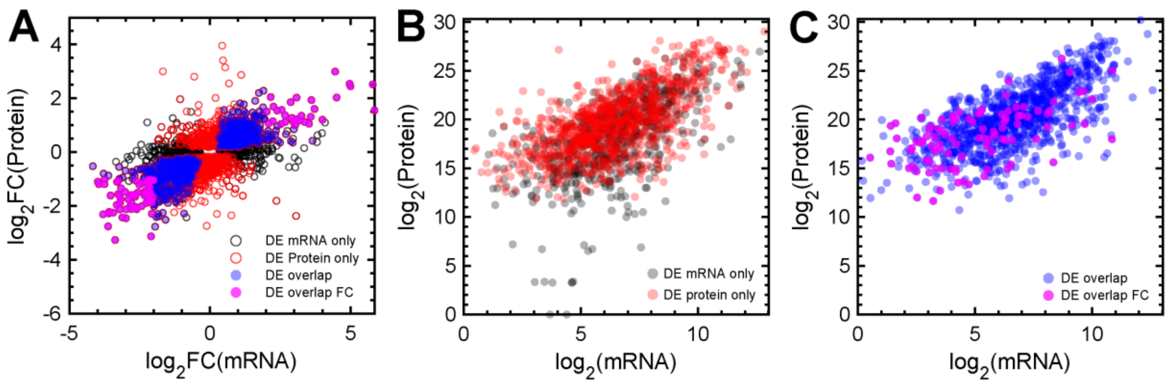


Figure 5. Assessing the RNA-protein relationships for differentially expressed mRNAs and proteins among fast and slow growing *P. aeruginosa*.

(A) Scatterplot of mRNA and protein fold changes among fast growing (3hr doubling time) and slow growing (25hr doubling time) *P. aeruginosa*. Genes with differentially expressed mRNA only (black open circle), protein only (red open circle), both mRNA and protein ("DE overlap", filled blue and pink circle) using the standard pval cutoff ($p_{adj} < 0.05$) or a more stringent cutoff ("DE overlap FC", mRNA fold change ≥ 4, protein fold change ≥ 2, $p_{adj} < 0.05$, filled pink circle) are shown. (B-C) Scatterplots of mRNA and protein abundances for genes with differentially expressed (B) mRNA only (gray), protein only (red), or (C) overlap using either the standard pval cutoff (blue) or the more stringent cutoff (pink) in *P. aeruginosa*. The Spearman rank correlation coefficients between mRNA and proteins among those gene subsets are: $\rho_{DE\_mRNA\_only}$ = 0.50-0.64, $\rho_{DE\_protein\_only}$ = 0.55-0.62, $\rho_{DE\_overlap}$ = 0.58-0.68, $\rho_{DE\_overlap\_FC}$ = 0.29-0.53, $p<10^{-7}$.

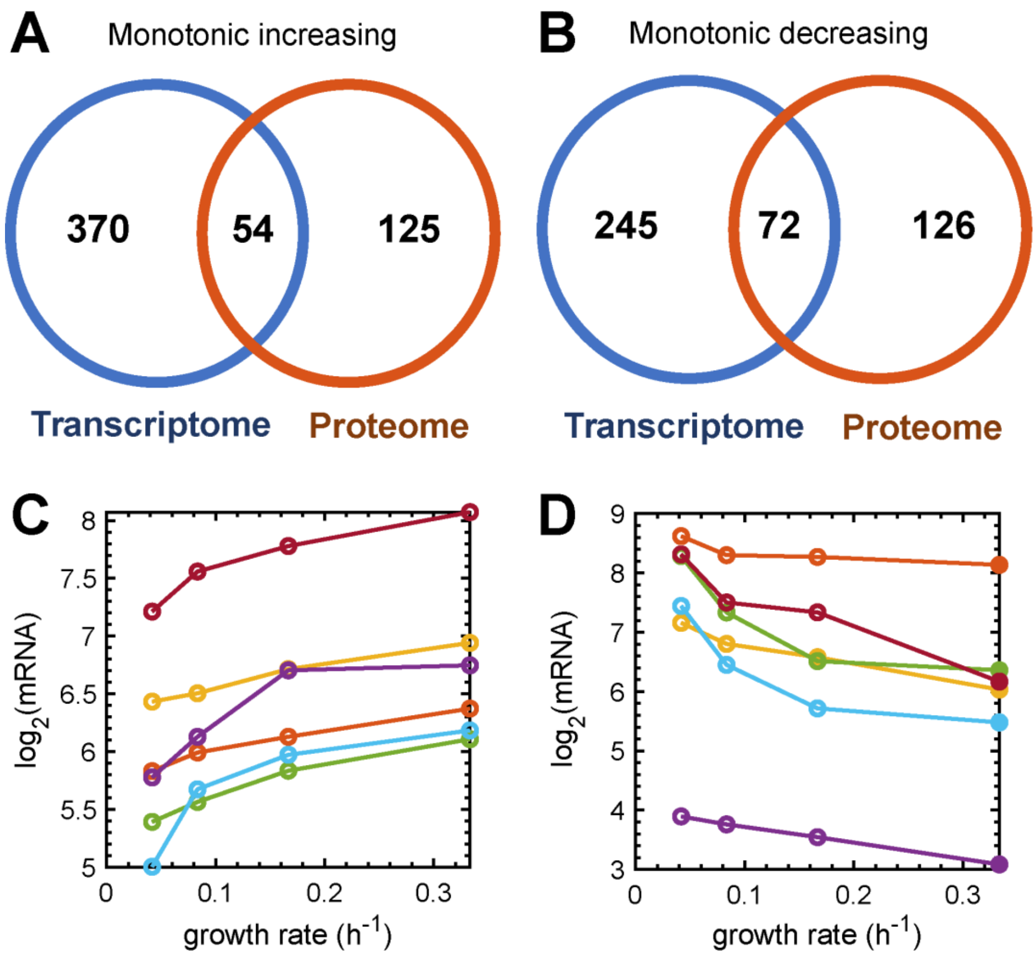


Figure 6. Identification of *P. aeruginosa* mRNAs and proteins that change in abundance monotonically with the growth rate.

(A and B) Venn diagrams showing the overlap of monotonic increasing (A) and decreasing (B) levels as a function of the growth rate in both the transcriptome and proteome. (C) Representative genes that showed monotonic increasing gene expression as the growth rate increased. Genes listed from top to bottom are PA14_61720 (red), uraA (yellow), murI (purple), upp (orange), moeB (green), and prmC (blue). (D) Representative genes that showed monotonic decreasing gene expression as the growth rate decreased. Genes listed from top to bottom are PA14_66320 (orange), gcvP1 (red), arcD (green), arcC (blue), phaC2 (yellow), and hutH (purple).

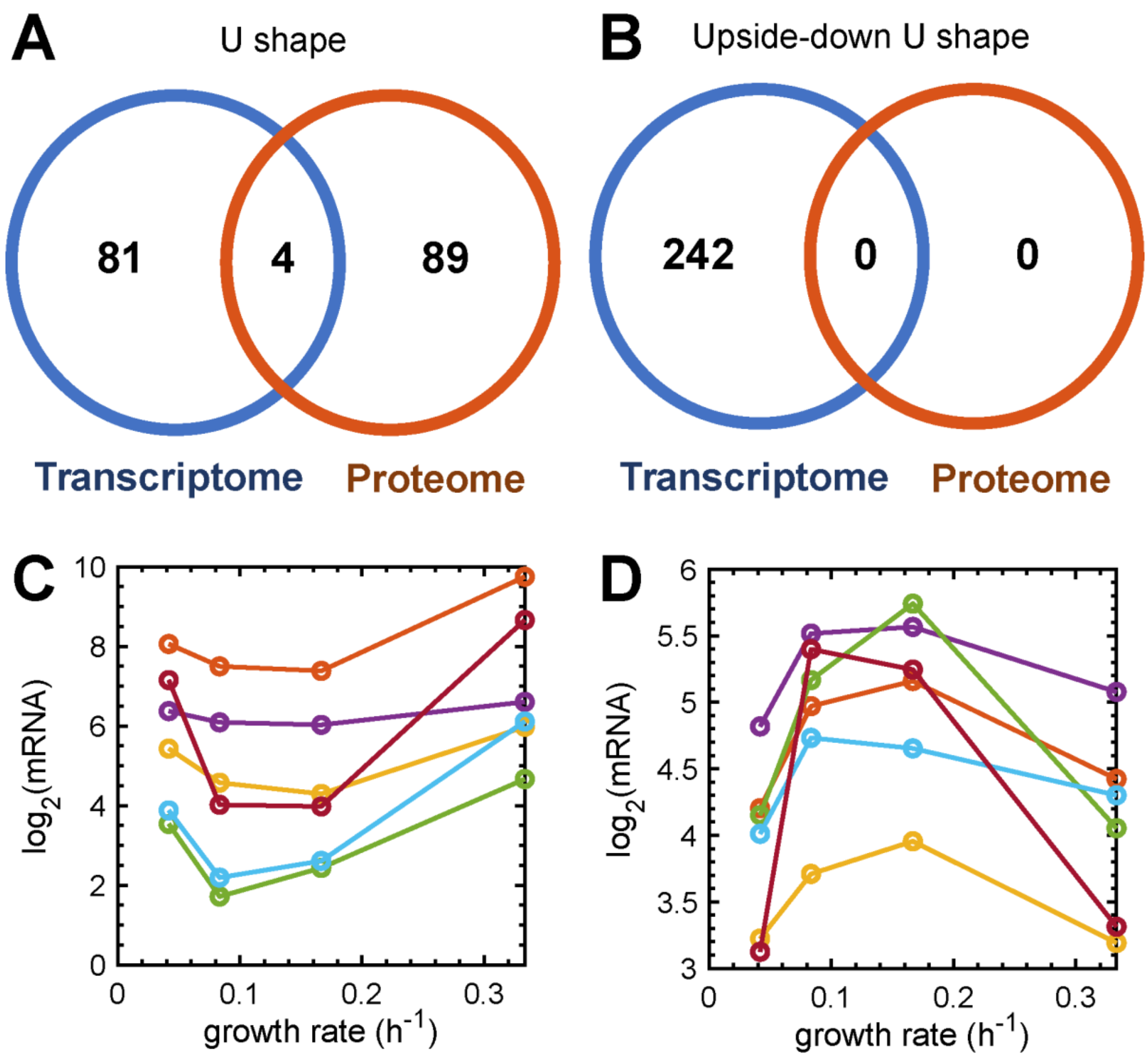


Figure 7. mRNAs and proteins with U-shape or upside-down U-shape expression patterns have little or no overlap.

Venn diagrams showing the overlap of (A) U-shape and (B) upside-down U-shape patterns of expression across growth rates in P. aeruginosa. Representative genes that showed (C) U-shape and (D) upside-down U-shape pattern of gene expression.

2.3 mRNA-protein relationships across genome-wide and gene subsets

2.3.1 mRNA and protein levels are positively correlated across genome-wide level

While the relationship between mRNA and protein abundances has been studied in bacteria under single growth conditions (150-153, 156, 158, 159, 168), our data provide the opportunity to systematically evaluate if mRNA levels are a good proxy to predict protein abundance in *P. aeruginosa* across a range of steady-state growth conditions that cover *in situ* growth rates during chronic human infection. For these analyses, we restricted our analysis to 3903 genes that were identified in both the mRNA and protein datasets in all growth rates. The paired datasets were restricted because the MS-based proteomics method is a “non-amplification” process and is limited by the sensitivity and dynamic range of mass spectrometry, thus it leads to preferential detection of abundant proteins and peptides with robust ionization. For example, we found that membrane proteins (172) are less detected in our proteomics datasets than predicted ($p < 10^{-41}$, Fisher’s exact test). Examination of the normalized distributions of detected mRNAs and proteins revealed a narrower distribution of mRNA than protein, with measured protein abundance varying by ~10^6 while mRNA levels varied by ~10^3 (**Figure 8**A-D, right panels). Thus, when comparing the least and most abundant mRNAs, the magnitude of this difference is smaller for mRNA than for protein. Whether this is a result of biological or methodological factors is not known, although this pattern is conserved across all growth rates.

We next quantified the correlation between protein and mRNA abundances by calculating the Spearman rank correlation coefficient (ρ) of the log2-transformed data, testing the hypothesis that gene expression levels correlate with protein levels (**Figure**

8A-D, middle and right panels). The correlation of mRNA and protein abundance were similar across growth rates (ρ = 0.55-0.63), indicating that growth rate has little impact on this correlation. The coefficient of determination (R^2) between mRNA and protein abundances revealed that on average 33% of the variability (28%-39%) in protein levels can be explained by the variability in mRNA levels.

2.3.2 Essential genes show higher correlation of mRNA and protein abundances

While we identified a genome-wide positive correlation between mRNA and protein abundances, we hypothesized that certain subsets of genes would exhibit stronger correlations. To test this hypothesis, we first assessed the RNA-protein relationships among genes and proteins differentially expressed at the fastest and slowest growth rates and found a similar correlation (**Figure 5**). We then examined the correlation of mRNA and protein abundances for genes within TIGRFAM functional classifications. Protein synthesis showed a higher mRNA-protein correlation (ρ_{ps} = 0.79, p<0.00001, N_{ps}=150) while regulatory functions displayed a lower correlation (ρ_{rf} = 0.38, p<0.0009, N_{rf}=76). However, these analyses were largely limited by the number of genes present in each category, which spanned from tens to hundreds of genes. We next examined genes that are known to be essential for *P. aeruginosa* viability, defined using a *P. aeruginosa* PA14 transposon insertion mutant library (170, 173). Of the 434 essential genes, we used 404 that were present in both our mRNA and protein datasets. Both the mRNA (Figure 9A) and protein (Figure 9B) levels of essential genes were significantly higher than those of the nonessential genes. In addition, the correlation between mRNA and protein levels of essential genes was significantly higher (0.64-0.73) than nonessential genes (0.49-0.58) across all growth rates (Figure 9C and D, p=0.029, Wilcoxon rank sum test). We also demonstrated that the higher correlation of essential genes compared to nonessential genes is not due to differences in gene numbers (**Figure 11**A and B) or expression levels (**Figure 11**C and D). In addition, despite increased levels of both mRNA and protein, the essential genes displayed significantly lower variance compared to non-essential genes (Figure 9E and F). These results demonstrate that genes essential for *P. aeruginosa* PA14 viability are highly expressed, are more stable in their expression levels, and have stronger mRNA-protein correlations across a range of growth rates.

2.3.3 Identification of outlier genes that have extreme protein-to-mRNA ratios

Which genes are most responsible for reducing the genome-wide correlation between mRNA and protein? To answer this question, we developed a statistical method to identify genes whose mRNA abundances were least predictive of protein levels. We identified genes (**Figure 10**A, pink and blue dots) that most deviated from the best fit linear regression lines of the genome-wide mRNA and protein levels at each growth rate (**Figure 8**) using a statistical measure called standardized residual. Standardized residuals are a quantity that describes the difference among the observed protein levels minus the predicted protein levels where the prediction is based on the genome-wide linear regression among mRNA and protein (174). These low-predictive-genes are categorized into two groups based on whether the differences in the protein to mRNA ratio (PTR) is higher or lower than expected (**Figure 10**B). The distribution of PTR was positively skewed, with a long left-tail containing 137 of the 160 genes (86%) with low PTRs at the 3hr growth rate (**Figure 10**B), indicating that it is more common to have genes with high mRNA levels but low protein levels rather than low mRNA and high protein levels. There were 94 genes with extreme PTRs that were shared in all four growth rates and 40 genes shared in three of four growth rates (**Figure 10**C), including genes encoding proteins involved in genetic information processing (*rodA, dcd2, rpoH, dnaE2, bolA*), metabolism (*aroQ1, pabB, ubiA, aprI, phzE1, trxB2*), and environmental information processing (*fha1, secG, tssF1, nppC*)(169). There were 70 genes that were specific to a single growth rate and another 35 genes that were only shared in two growth rates, including genes involved in siderophore biosynthesis (*pvdL, pvdH, PA14_33550, PA14_33560, PA14_33610, fpvA, pvdE, pvdO, pvdN, pvdA*). Enrichment analysis didn't reveal any significant enrichment of TIGRFAM functional categories among genes that are shared or not shared between growth rates (Fisher's exact test, $p<0.05$), likely due to the fact that most genes with extreme PTRs encode proteins of unknown function.

We next explored potential mechanisms that lead to genes with extreme PTRs. We first tested if genes with low PTRs are enriched for secreted proteins, with the rationale being that our proteomic experiments did not quantify secreted proteins. We did not find an enrichment of extracellular proteins ($p = 0.71$-1.0, Fisher's exact test) in genes with low PTRs across all growth rates. We next tested for enrichment of proteins that are not well detected by mass spectrometry, but again did not observe enrichment of small <10kDa ($p = 1.0$, Fisher's exact test) or membrane proteins ($p = 1.0$, Fisher's exact test) in genes with low PTRs. We then assessed whether genes with extreme PTRs could be explained by the sequence of the ribosome binding site (also referred to as the Shine-Dalgarno sequence), which has been shown to be important for the efficiency of translation initiation in *E. coli* (175). For this analysis, we calculated the predicted Gibbs Free energy of binding between the Shine-Dalgarno of each gene and the 16S rRNA component of the ribosome, with the hypothesis that genes with high PTR would have strong binding (low Gibbs free energy) and genes with low PTR would have weak binding (high Gibbs free energy) (176). However, the Gibbs free energy showed weak correlation with protein levels ($\rho = -0.12$ to

-0.09) and PTR (ρ = -0.18 to -0.17) (**Figure 10**D), explaining only ~1.0% (0.80%-1.4%) of the variability in PTR levels and ~2.1% (2.0%-2.3%) of the variability in protein levels. These data indicate that the secreted proteins, proteins poorly detected by mass spectrometry, and the Shine-Dalgarno sequence are not a primary driver of the PTR of these extreme genes.

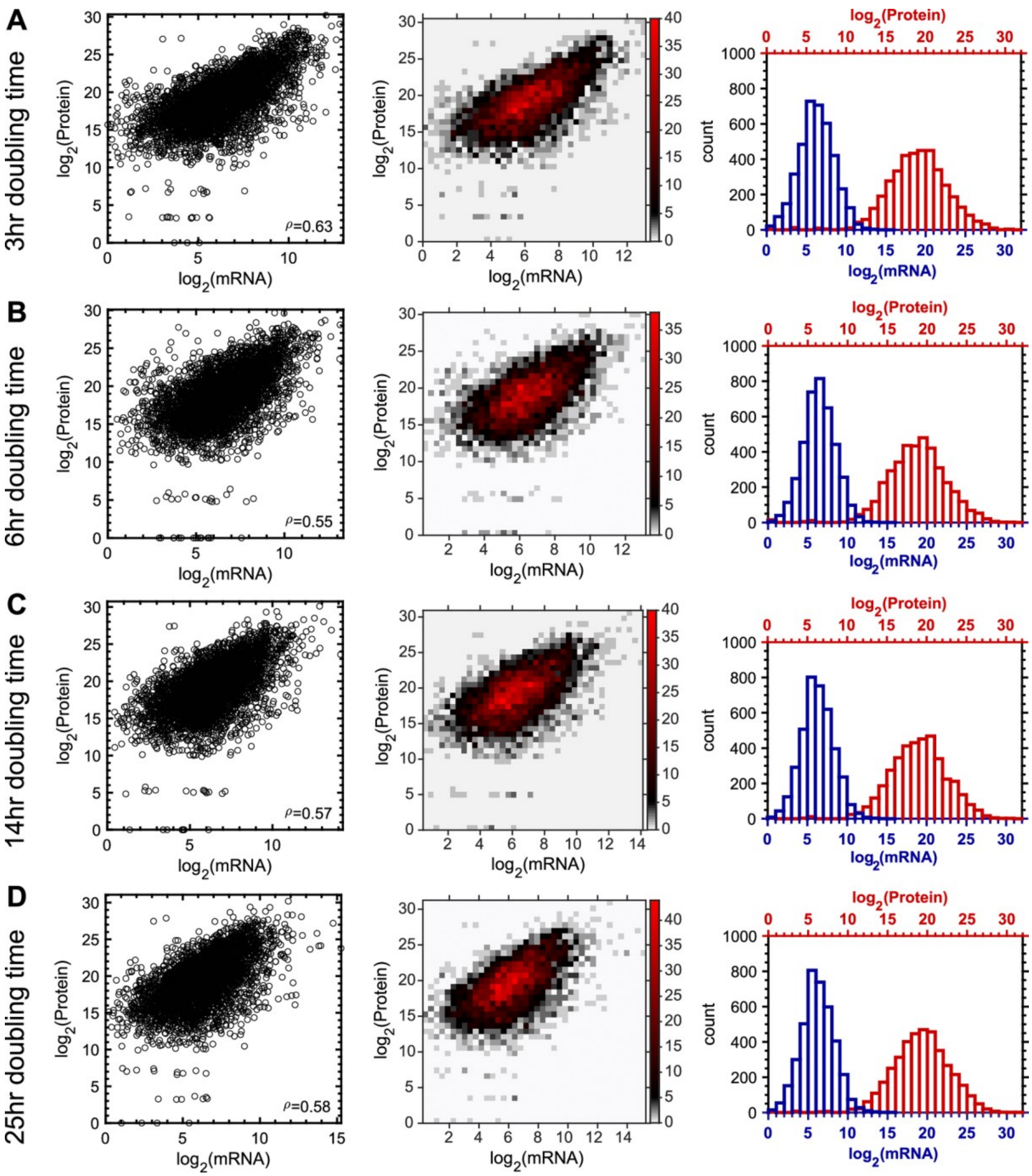
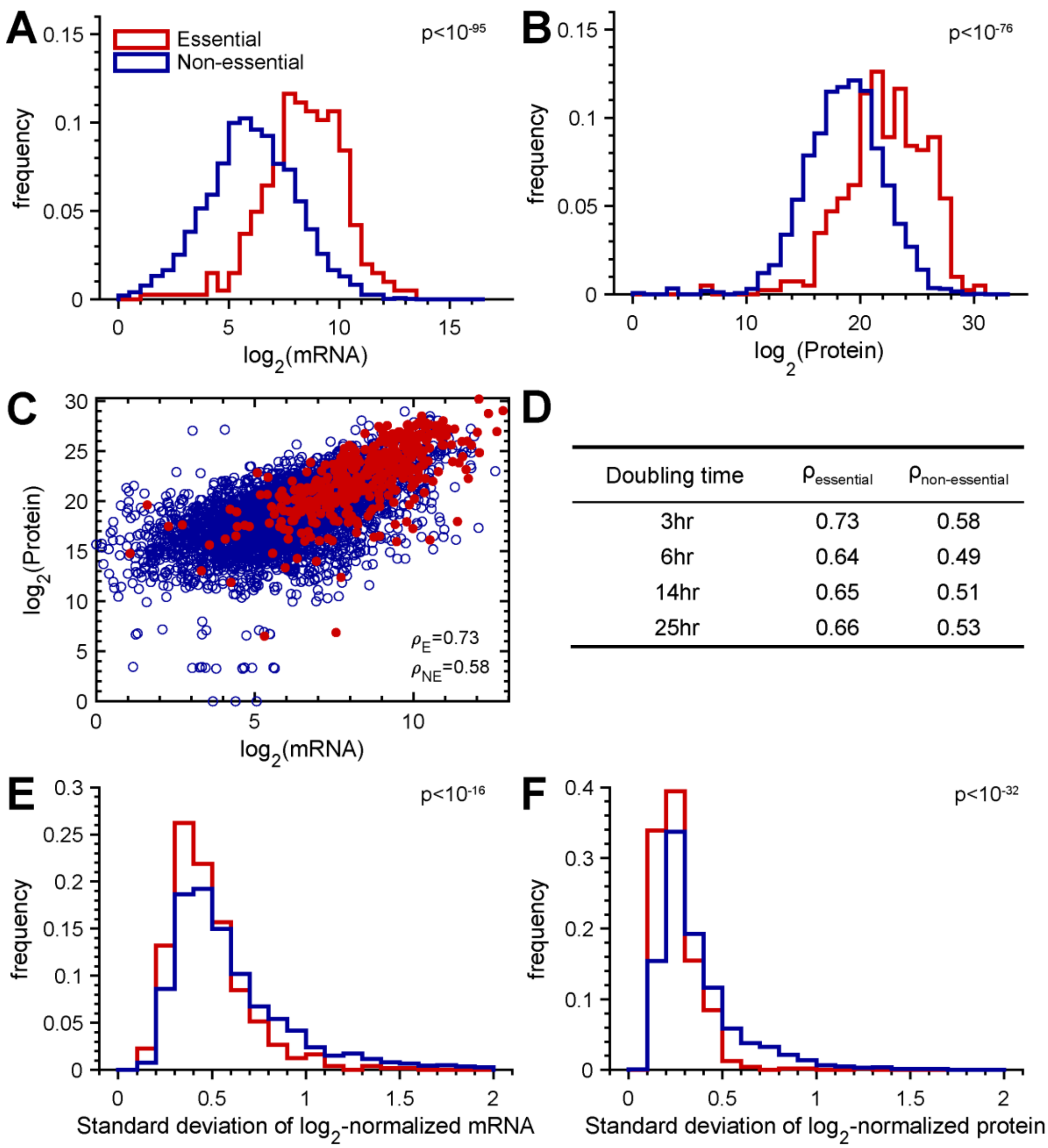


Figure 8. mRNA and protein are positively correlated at all growth rates

mRNA-protein relationships across a diverse range of growth rates spanning doubling times from 3 h to 25 h (A-D) are shown. Genes used for these analyses were detected at all growth rates in both the mRNA and protein expression data sets (n = 3,903). (Left) Scatterplots of mRNAs and proteins with associated Spearman rank correlation coefficients (ρ) demonstrating that mRNA and protein levels are positively correlated. (Middle) Binned scatterplots showing the relationship between measured protein abundances and mRNA abundances. The numbers of genes within each correlation are represented by the color scale in the right ordinate, where the number indicates the number of genes at a specific x-y coordinate. (Right) Distributions of mRNA and protein levels show that protein abundances vary to a greater extent than mRNA abundances.



Doubling time	$\rho_{essential}$	$\rho_{non\text{-}essential}$
3hr	0.73	0.58
6hr	0.64	0.49
14hr	0.65	0.51
25hr	0.66	0.53

Figure 9. ***P. aeruginosa* essential genes are highly expressed and have increased mRNA-protein correlations.**

(A and B) Distributions of mRNA (means ± standard deviations [SD], 8.4 ± 1.8 for essential genes and 6.0 ± 2.0 for nonessential genes) (A) and protein (means ± SD, 22.3 ± 3.5 for essential genes and 18.6 ± 3.4 for nonessential genes) (B) abundances for *P. aeruginosa* essential and nonessential genes at the 3-h doubling time reveal higher levels of both mRNA and protein for essential genes. (C) Scatterplot with associated Spearman rank correlation coefficient (ρ) for essential genes (red circles) (n_E = 484) and nonessential genes (blue circles) (n_{NE} = 3,419). (D) Spearman rank correlation coefficients for *P. aeruginosa* essential and nonessential genes at all growth rates. (E and F) Histograms of the standard deviations in mRNA (means ± SD, 0.48 ± 0.22 for essential genes and 0.62 ± 0.38 for nonessential genes) (E) and protein (means ± SD, 0.28 ± 0.32 for essential genes and 0.42 ± 0.49 for nonessential genes) (F) abundances for essential and nonessential genes demonstrate lower variance for essential genes than for nonessential genes. Histograms were constructed with a bin size of 0.05 using the variation in $\log_2$-normalized mRNA and protein levels from all growth rates. *P* values were calculated using a Wilcoxon rank sum test.

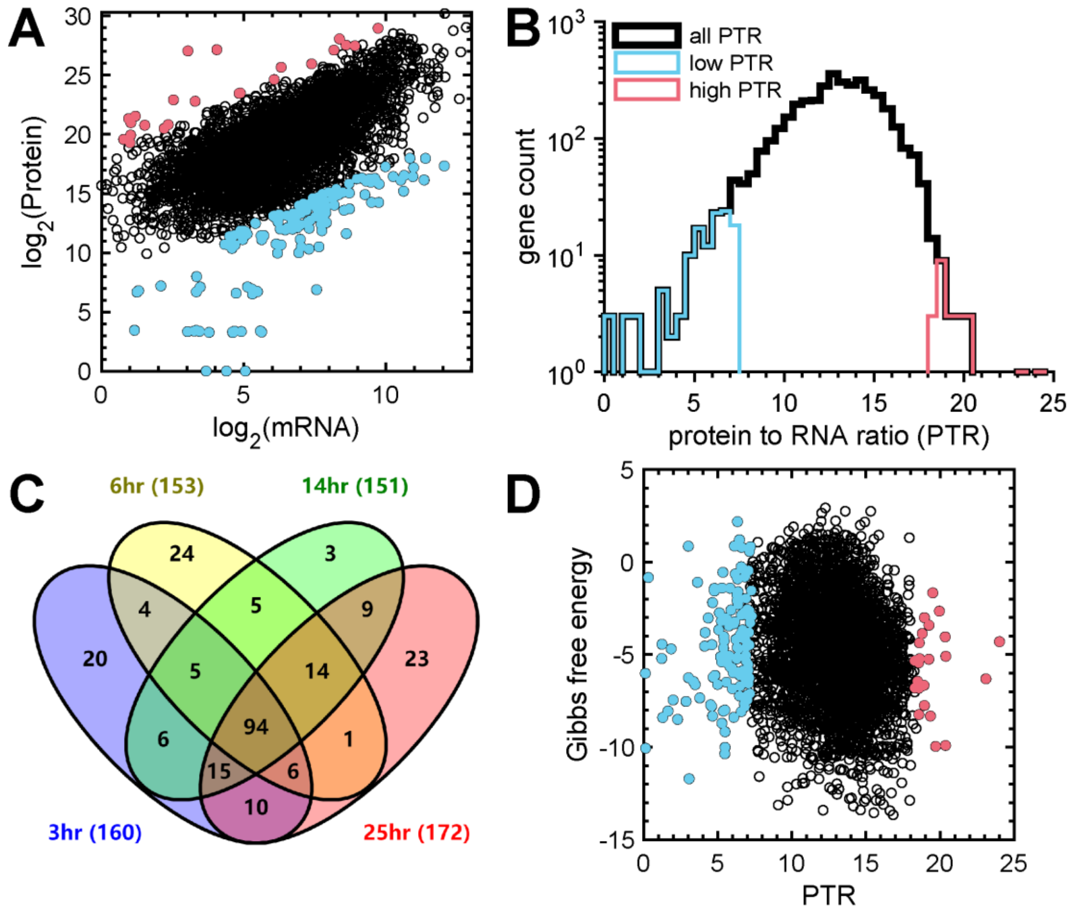


Figure 10. Identification of genes with poor mRNA-protein correlations.

(A) Scatterplot of mRNA and protein abundances noting 160 genes with high (pink) or low (blue) PTRs at the 3-h doubling time. Genes with high and low PTRs were defined as being 2 standard residuals from the predicted correlation. (B) Distribution of genes with high (pink) and low (blue) PTRs at the 3-h doubling time. (C) Venn diagram of genes with poor mRNA-protein correlations across all growth rates. Ninety-four genes are shared at all growth rates. (D) Graph showing the predicted Gibbs free energy of the Shine-Dalgarno sequence binding to the 16S rRNA of the ribosome as a function of the PTR. The predicted binding energy of the Shine-Dalgarno sequence was weakly correlated with the PTR for all genes, including the 160 genes with high (pink) and low (blue) PTRs.

2.4 Prediction of proteins from mRNA levels in vitro and during infection

2.4.1 Using a gene-specific RTP conversion factor to increase predictability of protein from mRNA levels

In addition to assessing the correlation of protein and mRNA for each gene, a goal of this work is to determine if we can improve the predictability of some poorly correlated genes using an RTP conversion factor. The basic idea is to identify poorly correlated genes that have consistent PTRs across growth conditions by calculating an inconsistency factor, then devising a conversion factor that can be applied to mRNA measurements to improve protein level prediction (154). The RTP conversion factor is a gene-specific conserved quantitative metric and is calculated by assessing the protein-to-mRNA ratios across all growth conditions for those genes identified using consistency factors. In addition to the chemostat transcriptomic and proteomic data generated in this study, we also used paired, publicly available transcriptomic and proteomic data for *P. aeruginosa* strains PA14 and PAO1 during batch culture growth in a medium (SCFM, Synthetic CF sputum Medium) (168) that mimics the nutritional environment of the CF lung (36, 170). The rationale for including these data is that they were acquired in a different media condition and at a different growth rate (40-minute generation time) from those used in the chemostat experiments. They also included data from a different *P. aeruginosa* strain (PAO1). These additional data provide the opportunity to assess the robustness of RTP conversion factors.

We identified 309 poorly correlated genes that had consistent RNA levels, protein levels, and PTRs (**Figure 14**) and calculated an RTP conversion factor for each of these genes. Examination of the function of these genes revealed that many encode proteins important for central carbon metabolism (*rpiA, tktA, pgk, eno, tal, aceA, idh, sucB, acnA,* *aceF, mqoA, mqoB*), purine biosynthesis (*ndk, guaA, purT, adk, purF, purB, purC, purM, purD, purA, purK*), branched chain amino acid biosynthesis (*ilvD, leuA, leuB, ilvH, ilvI, ilvE*) and aminoacyl-tRNA biosynthesis (*glyS, fmt, leuS, hisS, lysS, metG, gltx, thrS, serS, cysS, aspS, alaS, argS*). 114 out of 309 genes (~40%) are essential in *P. aeruginosa*, indicating that many poorly correlated essential genes have consistent PTRs that allow for calculation of a robust RTP conversion factor.

We next assessed the utility of the RTP conversion factor. As expected, application of the RTP to both *P. aeruginosa* PAO1 and PA14 mRNA datasets improved the mRNA predictivity of protein levels (**Figure 12**). Applying the RTP conversion factors to mRNA abundances of batch-grown *P. aeruginosa* PAO1 in SCFM and strain PA14 grown in chemostats improved the mRNA-protein Spearman rank correlation coefficient in PAO1 from 0.69 to 0.91 and in PA14 from 0.77 to 0.98 (**Figure 12**A and B). To test the robustness of this improvement, we performed leave-one-out model validation of our RTP conversion factors. In this analysis, we left one of our six test datasets out of the model, then applied these estimated conversion factors to the left-out transcriptome to determine if the corrected transcriptome was more predictive of measured protein levels (Table 2). We discovered that the correlations among predicted RTP-corrected mRNA and protein levels are significantly higher than the observed mRNA-to-protein correlations ($p<0.003$, Wilcoxon rank sum test) for all conditions. Prediction of left-out chemostat data was slightly better than the batch culture SCFM data (Table 2), likely because our dataset is dominated by chemostat data. However, the mRNA predictability of protein levels of the SCFM batch data of both strains PA14 and PAO1 were also highly improved (Table 2).

2.4.1 Application of RTP conversion factors to P. aeruginosa human infection transcriptomes

Since methods for assessing bacterial physiology during human infection are limited, the ability to perform RNA-seq on *P. aeruginosa* during human infection is highly informative. A prime motivator of this work was to develop RTP conversion factors to enhance the physiological relevance of these transcriptomes, with the goal of enhancing our understanding of the functions of *P. aeruginosa* during human infection. Thus, we applied the RTP conversion factors to transcriptomes gathered from *P. aeruginosa* infecting the human CF lung (139, 146). The results revealed that central metabolic pathways were significantly affected by application of the RTP conversion factor, including an increase in enzymes of the pentose phosphate pathway (**Figure 12**C and D). However, other pathways, including the tricarboxylic acid enzymes, were minorly impacted by application of the RTP conversion factor (**Figure 12**C and D; **Figure 13**).

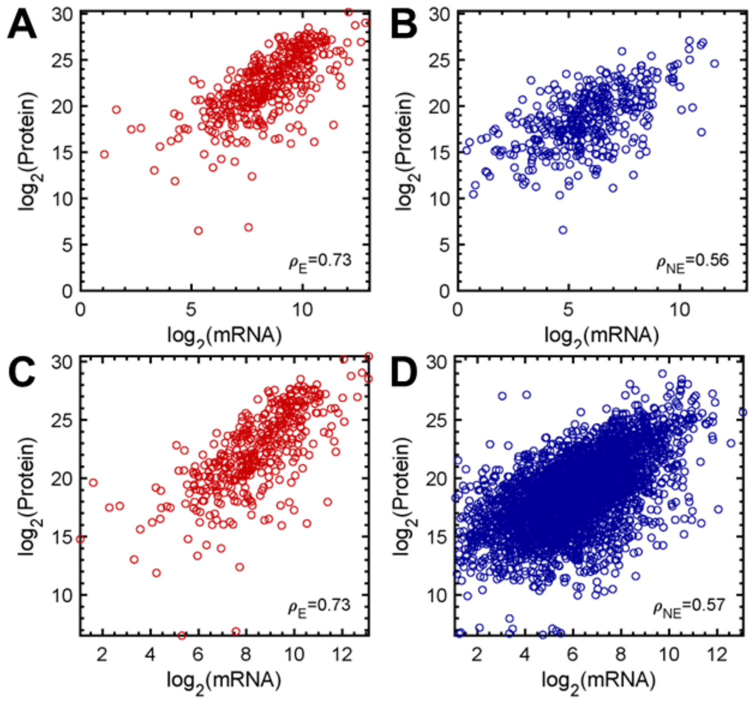


Figure 11. Higher correlation between mRNA and protein abundances for essential genes compared to nonessential genes are not due to differences in the number of genes and comparable expression levels.

(A-B) Scatterplot of mRNA and protein abundances in *P. aeruginosa* grown at the 3hr doubling time for (A) essential genes and (B) an identical number of nonessential genes randomly subsampled ($N_E = N_{NE} = 404$). The Spearman rank correlation coefficient for essential genes (0.73) is higher than the average correlation of nonessential genes subsampled 100 times (mean±SD, 0.57±0.039). (C-D) Higher correlations between mRNA and protein abundances for essential genes compared to nonessential genes are not due to comparable expression levels. Scatterplot of mRNA and protein abundances in P. aeruginosa grown at the 3hr doubling time with their Spearman rank correlation coefficients for (C) essential genes and (D) nonessential genes with comparable mRNA and protein expression levels ($p < 10^{-61}$, t test).

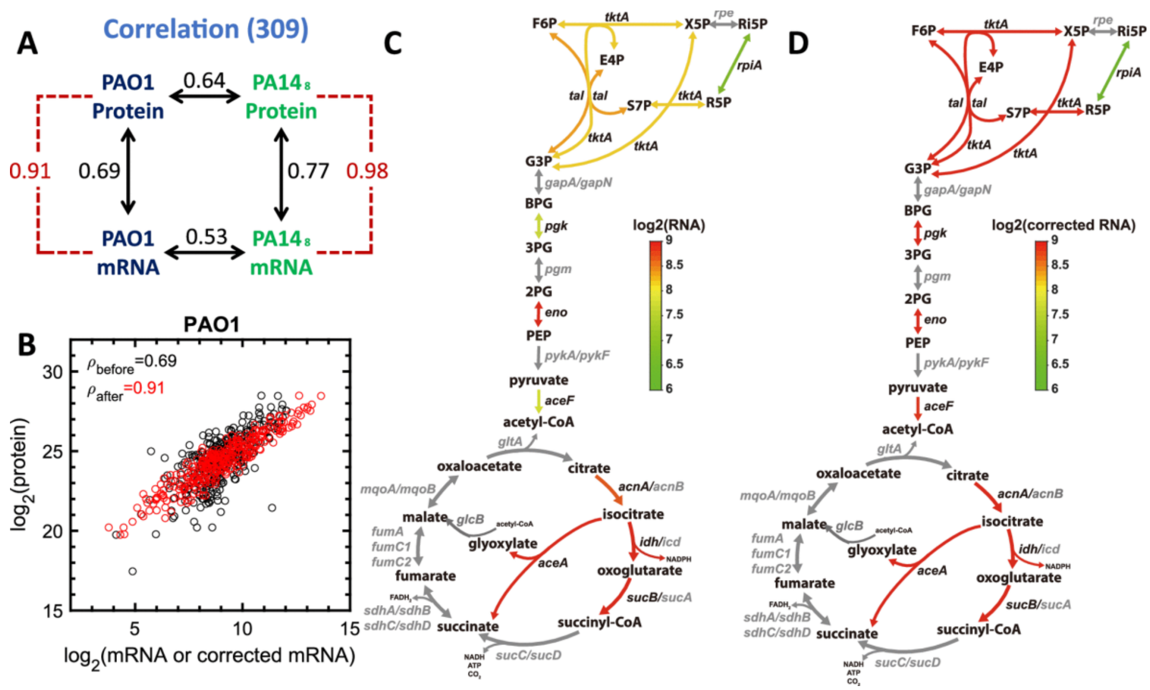


Figure 12. Application of gene-specific RTP conversion factors to *P. aeruginosa* transcriptomes from other strains, growth conditions, and human CF infections. (A) Prediction of protein levels by the application of the RTP conversion factors to mRNA levels of *P. aeruginosa* strain PAO1 grown in a synthetic CF sputum medium (SCFM) and strain PA14 grown in chemostats. The application of the RTP conversion factors improved the mRNA-protein Spearman rank correlation coefficients for PAO1 from 0.69 to 0.91 and for PA14 from 0.77 to 0.98. (B) Scatterplot of RNA-protein relationships among measured mRNA and protein levels (black circles) and mRNA corrected by the application of the RTP conversion factors and measured protein levels (red circles). Data are from *P. aeruginosa* strain PAO1 grown in a synthetic CF sputum medium. The associated Spearman rank correlation coefficients (ρ) among RNA and protein abundances increased from 0.69 to 0.91 after the application of the RTP conversion factors to mRNA levels. (C and D) Central metabolic pathway, with colored arrows showing predicted metabolic fluxes using mRNA levels from *P. aeruginosa* human CF infection transcriptomes (C) and mRNA

levels from *P. aeruginosa* human CF infection transcriptomes after applying the RTP conversion factor (D). Gray indicates the absence of genes or proteins detected in the data sets.

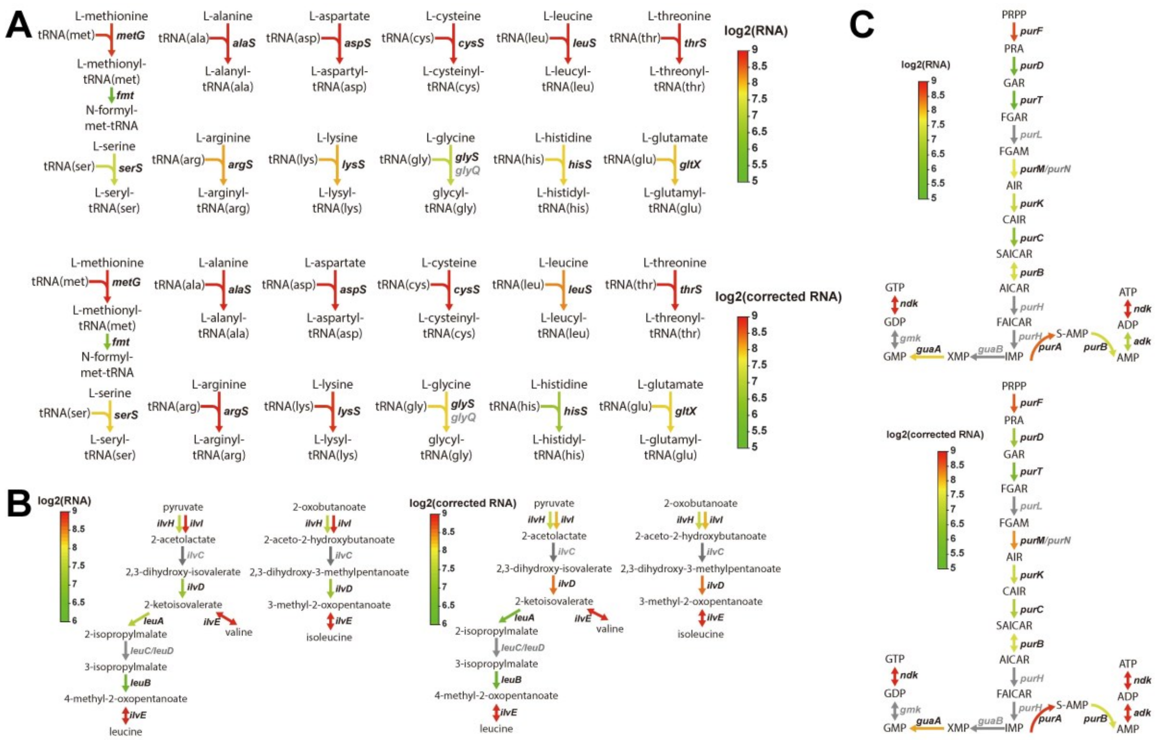


Figure 13. RTP conversion factor application to predicted metabolic fluxes from human infection transcriptome data

Important metabolic pathways with colors showing predicted metabolic fluxes using mRNA levels from P. aeruginosa human CF infection transcriptomes (top panel or left panel) and corrected mRNA levels (bottom panel or right panel) after applying the RTP conversion factor. The gray color indicates the absence of genes or proteins detected in the datasets. These pathways encode important genes for bacterial physiology, such as (A) aminoacyl-tRNA biosynthesis (B) branched chain amino acid biosynthesis (C) purine metabolism.

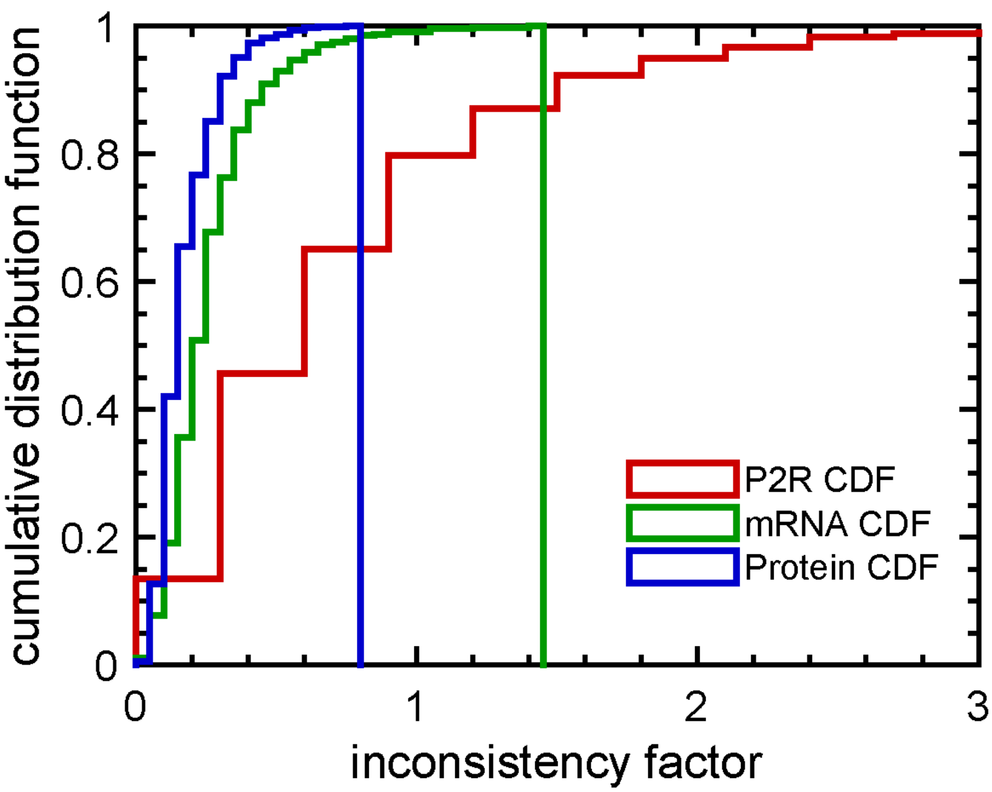


Figure 14. Application of inconsistency factors to threshold genes appropriate for RTP conversion factors

Cumulative distribution function curves for inconsistency factors applied to chemostat transcriptomes (green), chemostat proteomes (blue) and PTR ratios for all conditions (red).

Table 2. Leave-one-out model validation of RTP conversion factors.

Strain[a]	Growth condition[b]	N_{gene}[c]	$\rho_{measured}$[d]	$\rho_{corrected}$[d]
PA14	Chemostat, MOPS-succinate, 25hr	289	0.76	0.95
PA14	Chemostat, MOPS-succinate, 14hr	280	0.73	0.97
PA14	Chemostat, MOPS-succinate, 6hr	298	0.74	0.97
PA14	Chemostat, MOPS-succinate, 3hr	270	0.77	0.97
PAO1	Shaking flask, SCFM	381	0.72	0.89
PA14	Shaking flask, SCFM	388	0.72	0.90

[a] Strain of *P. aeruginosa.*

[b] Growth condition that was left out of RTP conversion factor calculation, then corrected using the RTP conversion factor.

[c] Number of genes chosen for individual leave-one-out model

[d] Spearman rank correlation coefficient of measured ($\rho_{measured}$) and RTP-corrected ($\rho_{corrected}$) mRNA-protein levels.

2.5 Discussions

The goal of this work was to explore the impact of the growth rate on mRNA and protein contents in *P. aeruginosa*. Our highly controlled chemostat experiments allowed us to comprehensively quantify the mRNA and protein compositions of cells at steady state across a range of growth rates observed during human infection. It is clear that growth rate significantly alters the protein and mRNA composition of *P. aeruginosa*, with hundreds of mRNAs and proteins differentially present in cells growing at fast and slow growth rates (**Figure 4**). Both mRNAs and proteins from genes encoding ribosomal proteins were elevated in fast growing cells. This is an expected result because ribosome numbers generally increase in faster growing bacteria (177-179). Some of these ribosomal proteins were members of a group of 54 genes in which both mRNA and protein levels increased monotonically with growth rate. These genes, along with 72 genes in which mRNA and protein levels decreased monotonically with growth rate (**Figure 6**), have the potential to be used to predict the average growth rate of *P. aeruginosa* populations in complex bacterial communities in which growth rate can't easily be measured. One aspect of our future work is to leverage these genes to predict growth rate in *P. aeruginosa* human infections using linear regression and deep learning approaches.

Like many other studies in organisms from all domains of life (149-163), we found that mRNA and protein levels are generally positively correlated. Correlations were similar across growth rates, with the highest Spearman rank correlation coefficient (ρ=0.63) observed at the fastest growth rate (**Figure 8**). However, there are a number of genes that showed poor correlation (**Figure 8**). While the majority of the most poorly correlated genes have low PTRs (**Figure 10**B), these genes were not enriched for secreted proteins, indicating that the low PTR genes cannot be explained by the fact that our proteomics experiments did not quantify secreted proteins. We also assessed protein properties that might reduce their detection with mass spectrometry (180-183), and found that low PTR genes are not enriched for membrane or small (<10 kDa) proteins. Nor could the extreme low or high PTR genes be fully explained by the sequence of the Shine-Dalgarno (**Figure 10**D), which is consistent with previous studies in which the Shine-Dalgarno had little explanatory power for genes with extreme PTRs (168, 169). Thus, while the mechanism is unknown, it is likely that many of these genes are the target of post-transcriptional regulation, either via mRNA/protein degradation or by small regulatory RNAs.

The observation that the abundance of mRNAs from *P. aeruginosa* essential genes were more predictive of protein levels than non-essential genes (Figure 9) provides new insights into these conserved genes. Consistent with studies in yeast and human cell lines (184, 185), we also noted that essential genes are more highly expressed compared to non-essential genes. mRNA from only 14 of the 404 essential genes were present at levels below 2^5, and only 10 of 404 essential proteins were detected at levels below 2^{15} (Figure 9C). One potential explanation for the high protein predictability of essential gene mRNA is that the high levels of mRNA and protein reduce detection bias. However, non-essential genes with similar protein and mRNA levels as essential genes have lower correlation (**Figure 11**C&D). Instead, we propose that essential genes have stable PTRs as they serve core functions in the cell and act as hubs in protein-protein interaction networks, similar to that observed in human cells (184). As such, we predict that most essential genes are not subject to significant post-transcriptional regulation, but instead rely on mRNA levels to modulate protein numbers.

Previous work has shown that gene-specific RTP conversion factors can be used to better predict protein abundances from mRNA levels across human tissues and cell lines (154). Here, we identified 309 *P. aeruginosa* genes that have consistent RTP conversion factors across strains, media and growth rates (Figure 12B). Notably, application of these correction factors to RNA-seq data acquired from a different strain of *P. aeruginosa* grown in non-chemostat conditions improved the protein predictivity of mRNA levels (Table 2). We anticipate that this RTP conversion factor approach can be applied to many more than the 309 genes identified in this study. Here, our identification of these genes was limited by the poor depth of coverage of the proteomics data from the previous study (168), where only 703 proteins were detected. Overall, this approach appears to be robust, and further work generating high-coverage, simultaneous transcriptome and proteomes from additional *P. aeruginosa* strains in a range of growth conditions will strengthen the identification of appropriate genes and robustness of our RTP conversion factor approach.

This study furthers our recent work using RNA-seq to assess the physiology of bacteria during human infection (128-130, 139, 146, 186). While there are challenges to using mRNA levels to predict protein levels in microbial communities due to poor mRNA-protein correlation of some genes, this work provides a corrective framework to increase the predictability of mRNA by assessing protein/RNA relationships across strains and growth conditions. Our discovery that essential genes have stronger mRNA-protein correlations is encouraging given that many of these genes are involved in fundamental processes critical for pathogenesis. This allows protein levels of these genes to be inferred from transcriptomic data with high accuracy in relevant environments (i.e. human infection), which we suggest will be valuable for efforts focused on development of antimicrobials targeting these essential functions.

In sum, we propose that there are significant gaps in knowledge regarding the physiology and behavior of microbes during human infection that hampers progress in diagnosing and treating infectious diseases. For example, a better understanding of the biology of infection could be used to improve the physiological relevance of the pre-clinical model systems needed to develop new therapeutics. Transcriptomic studies, which are one of the few methods that can be performed on complex microbial communities *in situ*, are leading the way in generating this knowledge. It is therefore critical to understand and address the limitations of these datasets, as we have done here. While it is likely that technological advancements will allow robust, high-throughput assessment of the microbial proteome during human infection, these approaches are likely years away from generating sufficient data for robust analyses. Thus, our approach fills an important gap in knowledge for understanding bacterial physiology in native environments.

2.6 Materials and methods

Strains, media, and growth conditions

Pseudomonas aeruginosa strain UCBPP-PA14 (PA14) (187, 188) was routinely grown on Lysogeny Broth-Miller (LB) agar plates and LB broth (Sigma-Aldrich) at 37°C. The chemostat growth media was a morpholinepropanesulfonic acid (MOPS) buffered minimal medium (189) with 10 mM succinate as the sole carbon source (MOPS-succinate). Chemostats were performed in a 300 ml Berzelius beaker (Kimble) containing 100 ml of culture volume, stirred, and aerated with filtered air as previously described (189). Chemostat cultures were inoculated, monitored, and grown to a steady state in a 37ºC warm room. The flow rate was set to 3.9 mL/hr, 7.3 mL/hr, 16.6 mL/hr, 30.3 mL/hr using a peristaltic pump, which corresponds to 25.3 hr (referred to as 25 hr), 13.8 hr (referred to as 14 hr), 6 hr, and 3.3 hr (referred to as 3 hr) generation times. The chemostat was maintained at steady-state for 48 hrs before sampling.

Sample collection, library preparation for RNA-seq

A 5 ml sample of culture was taken from at least two chemostat experiments with 4 replicates in total and mixed with RNAlater, which was incubated overnight at 4°C and then stored at –80°C. RNA-seq libraries were prepared as previously described for *in vitro* samples (139). Ribosomal RNA was depleted using the MICROBExpress bacterial mRNA enrichment kit (Sigma). The depleted RNA was fragmented for 2 min with the NEBNext Magnesium RNA fragmentation module kit, and cDNA libraries were prepared using the NEBNext multiplex small RNA library prep kit (New England Biolabs). The final RNA concentration for each sample was assessed with a Qubit dsDNA HS assay kit (Thermo) and the High Sensitivity Bioanalyzer kit (Agilent). Libraries were sequenced at the Molecular Evolution Core at the Georgia Institute of Technology on an IlluminaNextSeq500 using 75-bp single-end runs.

Bioinformatic analysis for RNA-seq

RNA-seq reads were mapped and analyzed as previously described by the Whiteley group (139). Gene counts were tallied to the coding sequences (CDS) using the R package featureCounts (available under subread package) using the strand-specific option. For correlational analyses, raw data were normalized using the reads per million mapped reads (RPM) method. Differential expression between the two extreme growth rates was determined using DESeq2 (Bioconductor). Transcripts were considered as differentially increased or decreased in abundance only when their padj was <0.05 and log2 fold change was $>|2|$. All figures were created in MATLAB using function scatter, histogram, imagesc, and PlotPub.

Sample collection and preparation for proteomics mass spectrometry

For each sample, 5 ml of chemostat culture was collected on ice, followed immediately by centrifugation at 4ºC at 14,000g for 5 min. The supernatant was removed, and cells were washed once in 1 ml of ice-cold sterile PBS. Cells were pelleted again, supernatants removed, and cell pellets were stored at -80ºC. To allow relative comparison of proteins between samples, proteins in all samples were labelled, pooled, and quantified in a single run by the Georgia Institute of Technology's Systems Mass Spectrometry Core Facility. Proteins were extracted from frozen cell pellets using a urea lysis buffer and digested with trypsin overnight. Peptides were labeled with the TMT isobaric labels (ThermoFisher Scientific), pooled, and separated by HPLC (High-Pressure Liquid Chromatography, Beckman). Specifically, a total of 120 fractions were collected and pooled. Both the unfractionated and the pooled fractions were analyzed by nano-LC/MS-MS using Dionex nanoLC system coupled to a Q-Exactive Plus mass spectrometer (Thermo Scientific). Reverse-phase chromatography was performed using an in-house packed column (40 cm long X 75 µm ID X 360 OD, Dr. Maisch GmbH ReproSil-Pur 120 C18-AQ 1.9 µm beads) and a 120 min gradient. Mass spectrometry was performed in a data-dependent mode, including a full MS1 scan and 12 MS2 scans. The MS1 scan was performed using parameters with 70,000 resolution, 3 × 106 AGC, and 20 ms maximal ion time. The MS2 scans were performed with 35,000 resolution, 5 × 104 AGC, 120 ms maximal ion time, HCD, 30% normalized collision energy, 1.6 m/z isolation window with 0.3 m/z offsets, and 20s dynamic exclusion.

Protein search parameters, mass spectrometry quantification, and data analysis

Raw mass spectrometry data files were processed with Mascot (Matrix Science; v2.6) and Proteome Discoverer (Thermo Fisher Scientific; v2.1). Data were searched against the predicted *P. aeruginosa* PA14 proteome (Pseudomonas.com), in addition to using the common repository of adventitious proteins (cRAP) v1.0. The data were searched with MS peptide tolerances of 10 ppm and MS/MS tolerance for identification of 100 millimass units (mmu) for precursor. Quantification was achieved by calculating the sum of centroided reporter ions within a ±2 mmu window around the expected m/z for each of the 4 TMT reporter ions. The results were then exported to Microsoft Excel for further data interpretation and statistical analysis. Proteins that were not detected in all channels were excluded and then log2 transformed. Differential protein expression was statistically determined using DEseq2 (Bioconductor). Proteins were considered as differentially increased or decreased in abundance only when their log2 fold change was >2 or <−2, respectively, and their adjusted p-value was <0.05. Venn diagrams were created with Venny (190).

Shape analysis: monotonic genes

To identify mRNA or proteins whose levels changed monotonically with growth rate, we evaluated the mRNA and protein dynamics across all four growth rates using two steps. First, for both mRNA and protein we selected genes that showed consistent positive or negative changes in levels using pairwise comparisons of the four generation times (3hr-6hr, 6hr-14hr and 14hr-25hr). Second, we performed differential gene expression between the fastest and slowest growth rates to ensure that the trend of increasing/decreasing levels is significant under these two extreme conditions using a padj ≤ 0.05.

Shape analysis: U-shape and upside-down U-shape

To identify mRNA and proteins with a U-shape and upside-down U-shape patterns, for each gene we performed four two-sample one-tailed t-tests comparing mRNA levels comparing four pairs of growth rates: 3hr-6hr, 3hr-14hr, 6hr-25hr, and 14hr-25hr. Genes were defined as U-shape if the following conditions were met for mRNA levels ($p \leq 0.05$): 3hr>6hr, 3hr>14hr, 6hr<25hr, and 14hr<25hr. Genes were selected as upside-down U-shape if the following conditions were met for mRNA levels ($p \leq 0.05$): 3hr<6hr, 3hr<14hr, 6hr>25hr, and 14hr>25hr.

RTP conversion factor calculation and application to human P. aeruginosa transcriptomes

In addition to our chemostat data, transcriptome and proteome data from 703 genes identified in Kwon et al (168) was also used for RTP conversion factor calculation. All data were normalized to identical depths for analysis.

To identify conserved metrics to predict protein levels from mRNA levels, we calculated an inconsistency factor as previously described (168) to identify genes with consistent RNA levels, protein levels, and PTR ratios. The inconsistency factor was calculated as the sum of positive distances between data points of individual datasets and mean value divided by the number of conditions.

$$inconsistency\ factor = \frac{\sum |\ x_i - \overline{x}|}{n}$$

We identified 309 genes that showed consistent RNA levels, protein levels, and PTR ratios using an inconsistency factor cutoff of 0.6 across all experimental conditions (**Figure 12**). For these genes, an RTP conversion factor was calculated as the average PTR ratio across all conditions.

To assess the robustness of the RTP conversion factors, we performed two analyses. First, We calculated the predicted protein level for each of the 309 genes by multiplying the mRNA level of each gene by its RTP conversion factor. Then we calculated the Spearman rank correlation coefficient of the predicted protein levels compared to the observed protein levels. Second, we performed a leave-one-out-analysis for all six experimental conditions. In each case, five experimental conditions were used for calculation of RTP conversion factors. These were then applied to the data from the left out experimental condition by multiplying the mRNA level of each gene by the gene-specific conversion factor. Predicted protein levels were then compared to the measured protein levels for the left out experimental condition using a Wilcoxon rank sum test.

RTP conversion factors were applied to *P. aeruginosa* mRNA levels quantified previously from human-derived infection samples from the cystic fibrosis lung infections (139, 146). The corrected mRNAs are calculated based on the predicted protein levels and the best fit regression line that linearly associated genome-wide mRNA and protein levels. Only 308 of the 309 RTP conversions factors were used for this analysis, as mRNA for one gene was not detected in the cystic fibrosis lung infection transcriptomes. Pathways predicted from the *P. aeruginosa* CF transcriptomes and corrected with the RTP conversion factor were constructed as previously described (129, 170, 191).

Data availability

RNA-seq raw data are available in the NCBI Sequence Read Archive under accession number PRJNA867727 and proteomics raw data are deposited to the Mass Spectrometry Interactive Virtual Environment (MassIVE) database with the identifier MSV000090126.

CHAPTER 3. MRNA-PROTEIN RELATIONSHIP ACROSS DOMAINS[3]

3.1 Background and significance

A major challenge in understanding the biology of native microbial communities is the lack of robust functional data for the microbes within these ecosystems. In many cases, this has hampered the development of biologically relevant in vitro and in vivo models that can provide fundamental, functional insights into the biology of native microbial communities. One way to tackle this problem is to quantify gene expression in native communities and use this data to infer microbial function. This transcriptomic approach has been highly successful in assessing the function of in situ microbial communities, including the human microbiome during health and disease (5, 6, 8, 192), marine environments (1, 2), soil (3, 4), and even during space flight (193-195). Although RNA-seq has been widely used to study bacterial physiology during infection, a critical concern arises regarding whether mRNA levels accurately predict protein levels, which are the primary functional units of a cell (56, 196). While previous studies have established a positive correlation between mRNA and protein levels, for certain genes mRNA is not a

[3] Parts of this subsection were adapted from the following reference: Zhang M, Zhang C, Ramos A, Whitaker R, Whiteley M 2024. Using conserved protein to mRNA ratios across domains to enhance functional discovery in microbial metatranscriptomes (in prep) Copyright © 2024 Zhang et al. © 2024 Zhang et al. Reused with permission. I was the primary co-first author of this work.

predictor of protein levels. However, performing comprehensive quantitative proteomics of microbes in natural ecosystems is incredibly challenging (197-200). Although this limitation may be overcome in the future with recent technological innovations (201-203), RNA-seq will remain an economical and powerful technique for assessing microbial function in complex environment. Thus, there is a pressing need to bridge the gap between mRNA levels and protein prediction.

To address this gap, large datasets have been generated to assess the protein to RNA ratio (PTR) of each gene, asking the question: Is the PTR highly conserved, or is it impacted by microbial strain and environmental factors? Surprisingly, these studies have shown that the PTR for some genes are highly conserved between bacterial strains in multiple growth environments (168, 204-206). This conservation in PTR has allowed for the calculation of an RTP conversion factor for these genes, which is a gene-specific conversion factor that can be applied to mRNA measurements to improve predictions of protein levels (204, 207). This approach has enhanced our understanding of microbial functions in complex natural environments, providing a straightforward approach for correcting functional predictions in transcriptomic data.

While RTP conversion factors have the potential to greatly enhance our understanding of microbial function in natural ecosystems, their calculation requires large proteomic and transcriptomic datasets, which are not available for most microbes and can be extremely costly to obtain. One potential solution is to use comprehensive datasets in a limited number of microbes to calculate RTP conversion factors that can be applied broadly to multiple microbes. The primary challenge to this approach is identifying orthologous genes in multiple microbes that have consistent PTRs. Here, we addressed this challenge by calculating PTRs for thousands of genes using comprehensive transcriptome and proteome datasets from Gram-negative bacteria, Gram-positive bacteria, and an archaea. We discovered that the overall correlation of mRNA and protein is similar across diverse prokaryotes, with mRNA and protein positively correlated. In addition, we identified orthologous genes in which mRNA was not predictive of protein, but their associated PTR ratios for these genes were similar across the diverse microbes tested. Calculation of RTP conversion factors for these genes from highly-related microbes significantly improved the accuracy of protein prediction from mRNA of distantly related microbes, indicating a core subset of genes in which RTP conversions factors can be broadly applied.

Significance statement

Understanding the biology of native microbial communities is crucial but challenging due to the lack of robust functional data. This gap hinders the development of relevant in vitro and in vivo models necessary for fundamental biological insights. Transcriptomics, by quantifying gene expression to infer microbial function, has been effective across various environments. However, mRNA levels do not always predict protein levels, complicating functional predictions. While RNA-seq is economical and powerful, comprehensive quantitative proteomics remains difficult. Recent studies show that protein-to-RNA (PTR) ratios are often conserved across strains and environments, enabling the calculation of RNA-to-protein (RTP) conversion factors. These factors enhance functional predictions from mRNA data. Despite requiring large datasets, this approach shows promise by identifying conserved orthologous genes with consistent PTRs across microbes. This advancement allows for broader application of RTP conversion factors, significantly improving protein prediction accuracy and enhancing our understanding of microbial functions in complex ecosystems.

3.2 mRNA-protein relationships across diverse taxa and gene subsets

3.2.1 mRNA and protein levels are positively correlated.

We first assessed the correlation of mRNA and protein for each microbe, restricting our analysis to the genes in which both the mRNA and protein were detected. The number of genes used varied between the microbes, with the highest percentage in *A. actinomycetemcomitans* (78%) and the lowest in *S. aureus* (30%, Table 4). For all microbes, we quantified the correlation between protein and mRNA abundances by calculating the Spearman rank correlation coefficient (ρ) of the log2-transformed data (**Figure 15**, **Figure 16**). Positive correlations between mRNA and protein abundances were observed across all the prokaryotes tested with correlation coefficients ranging from 0.46 to 0.62. The coefficient of determination (R^2) between mRNA and protein abundances revealed that, on average, 27% (range of 18% to 38%) of the variability in protein levels can be explained by the variability in mRNA levels. The positive correlation between mRNA and protein levels observed here agrees with previous studies using both bacteria and eukaryotes.

3.2.2 mRNA and protein abundances of essential genes are more highly correlated than non-essential genes

We previously reported that *P. aeruginosa* essential genes have: (i) higher mRNA and protein abundances than non-essential genes; (ii) less variance in in mRNA protein abundance; and (iii) a higher correlation of mRNA and protein than non-essential genes (204). We re-examined these findings with a more refined list of *P. aeruginosa* essential genes, identified using additional transposon sequencing data (208). This analysis identified 302 *P. aeruginosa* essential genes shared across diverse growth conditions, with 294 of these genes present in both our mRNA and protein data sets (170, 208). Examination mRNA-protein abundances and correlations revealed that as previously reported (204), *P. aeruginosa* essential genes are expressed at higher levels (**Figure 17A** and **B**), have reduced variance (**Figure 17**C and D), and have higher correlation coefficients compared to non-essential genes (**Figure 17E** and **F**).

To determine whether this phenomenon is restricted to *P. aeruginosa* essential genes, we next examined mRNA-protein abundances and correlations of both essential and non-essential *S. aureus* genes. We chose to focus on *S. aureus*, as similar to *P. aeruginosa*, the essential genome is well-defined from genome-wide transposon sequencing and RNA interference data (209, 210). 302 essential genes were reported using RNA interference (211) and 351 genes using transposon sequencing (212), with 166 genes shared between these two conditions. Of these 166 shared genes, we focused on 137 genes that were present in both our mRNA and protein datasets. As in *P. aeruginosa*, we discovered that: (i) mRNA and protein abundances were higher for essential genes compared to non-essential genes (**Figure 17A** and **B**); (ii) despite increased abundances of both mRNA and protein, essential genes had significantly lower variance compared to non-essential genes (**Figure 17C** and **D**); and (iii) correlation between mRNA and protein levels of *S. aureus* essential genes was higher (0.58) than that of nonessential genes (0.38) (**Figure 17E** and **F**). These data provide further evidence that mRNA and proteins from essential genes are produced at higher, and less variable levels than non-essential genes, and have higher mRNA-protein correlations compared to non-essential genes.

3.2.3 Identification of genes that have extreme protein-to-mRNA ratios across bacteria

We next asked the question: Are the genes in which mRNA is not predictive of protein levels unique in each prokaryote, or are they orthologous genes shared between the microbes? To answer this question, we first identified 'outlier' genes that have extreme protein-mRNA (PTR) ratios in each microbe. We did not utilize data from *S. islandicus* for this analysis as there were insufficient orthologues between this archaea and bacteria for a robust evaluation. To identify outliers, we used standardized residuals, which calculates the extent of deviation of each gene from the best-fit regression line of the genome-wide mRNA and protein levels. We defined outlier genes as those in which the absolute values of PTR ratios were ≥ 2 standard deviations away from the linear regression line between mRNA and protein abundances. We then categorized these outlier genes into two groups based on the sign of standardized residuals: those with a positive sign indicate higher protein levels than expected (termed high PTR) and those with a negative sign indicating lower protein levels than expected (termed low PTR). Significantly higher numbers of genes with low PTRs compared to high PTRs were identified in all microbes examined (Table 5), p=0.002, Wilcoxon signed rank test). Furthermore, the distribution of PTR ratios is 'left-skewed', with a notable prevalence of genes showing low PTRs over high PTRs (**Figure 23**), aligning with our identification of outlier genes using standardized residuals (Table 5).

To evaluate whether these orthologous outlier genes are bacterium-specific or shared among bacteria, we first averaged the standardized residuals for each outlier gene from all bacteria. When calculating this average, we set the maximum standardized residual at an absolute value of 2 to limit the impact of single values on the average. We then used this average standardized residual to rank all genes, then the top (**Figure 18A**) and bottom (**Figure 18C**) 15 genes were selected for further analysis. Examination of the variability in PTR ratios revealed that the top 15 outlier genes exhibited consistent and high PTR ratios among all bacteria (**Figure 18B**), while the bottom 15 genes displayed more diverse PTR ratios (**Figure 18D**). Notably, we identified an outlier gene, elongation factor Ts (tsf), that consistently displayed extremely high protein levels with moderate mRNA levels across all bacteria except *P. gingivalis* (**Figure 18A** and **B**). These data reveal the existence of orthologous outlier genes that show similar PTRs amongst several, but not all of the microbes tested.

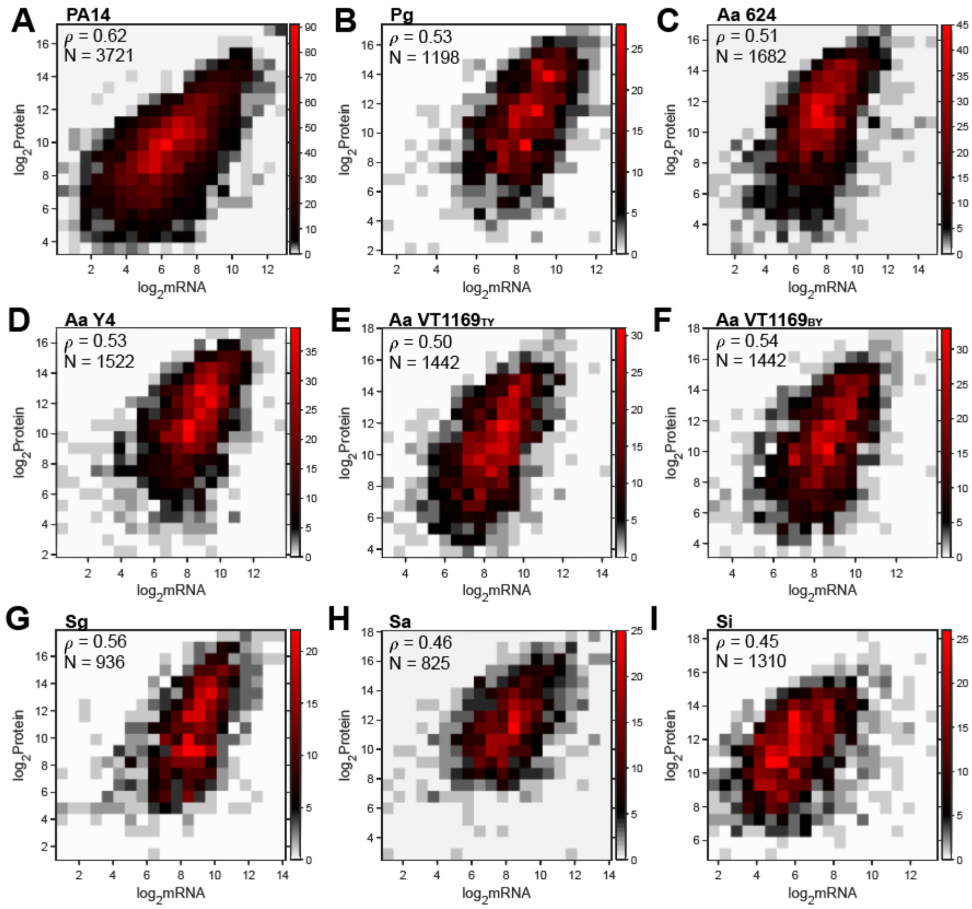


Figure 15. mRNA and protein are positively correlated across diverse prokaryotes. Binned scatterplots showing the relationship between measured protein abundances and mRNA abundances across a diverse range of species, including Gram-negative (A-F), Gram-positive (G-H), and archaea (I). Spearman rank correlation coefficients (ρ) were used to the quantify that higher mRNA levels were associated with higher protein levels. The correlation ρ ranged from 0.45 to 0.62, demonstrating that indicate a moderate correlation between mRNA and protein abundances. The analyses included genes detected in both transcriptome and proteome datasets, with gene counts varying from hundreds to thousands (N_{min}=825 in *S. aureus*, N_{max}=3721 in *P. aeruginosa*). The color scale on the right ordinate represents the number of genes at specific x-y coordinates.

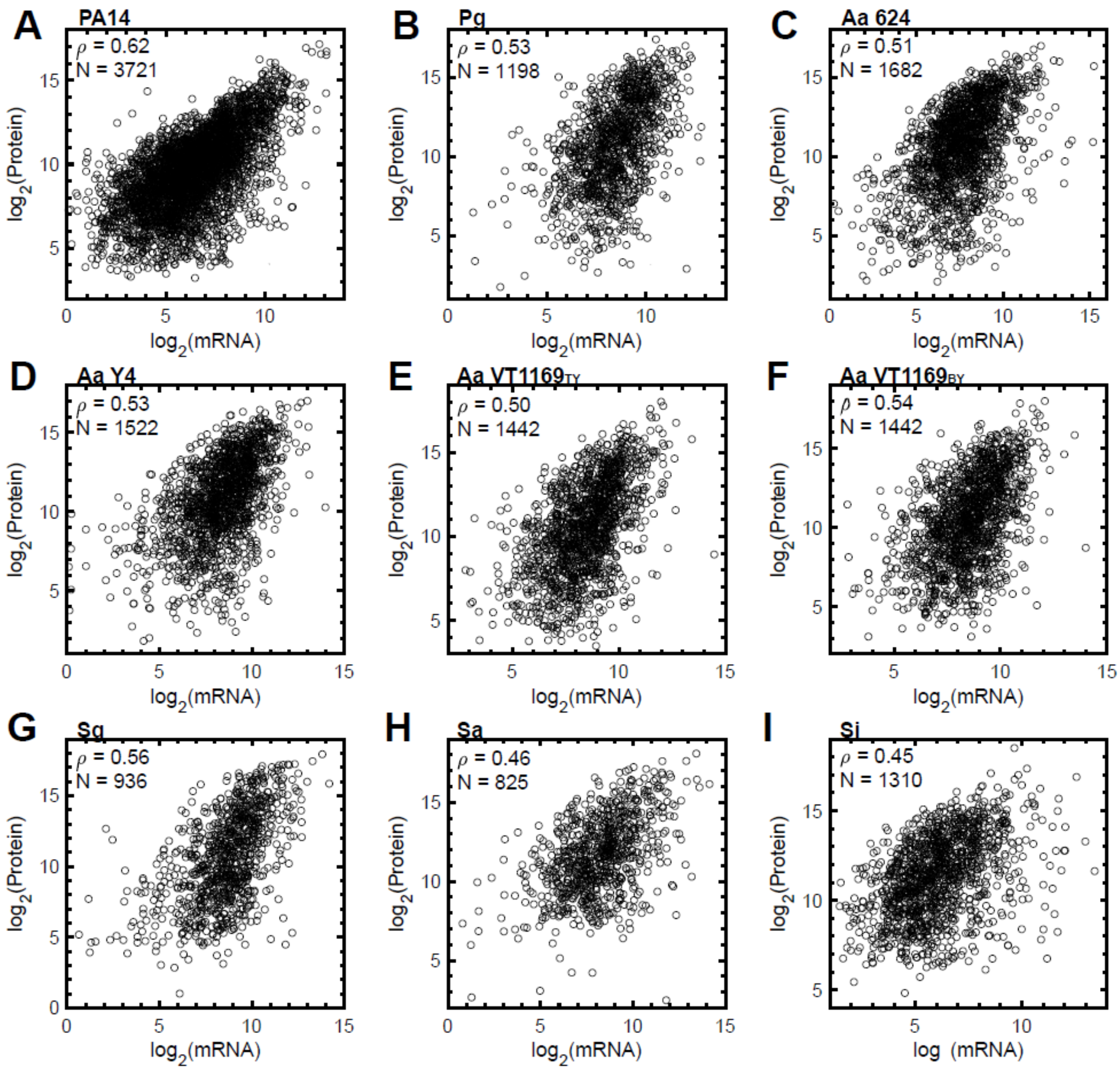


Figure 16. Scatterplots of mRNA and protein levels across a wide array of prokaryotic species.

Scatterplots show the relationship between protein and mRNA abundances across diverse taxa, encompassing Gram-negative bacteria (A-F), Gram-positive bacteria (G-H) and archaea (I). Spearman rank correlation coefficients (ρ) were employed to assess the association. Genes plotted here were detected in both transcriptome and proteome datasets, with gene counts varying from several hundred to thousands.

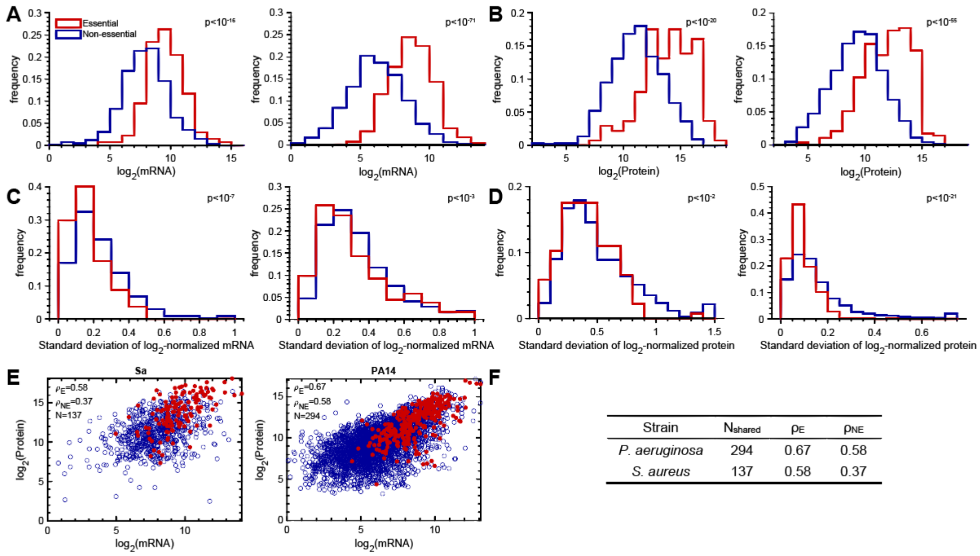


Figure 17. Conditional-independent essential genes are highly expressed and have increased mRNA-protein correlations in both *P. aeruginosa* and *S. aureus*

(A-B) Distributions of mRNA (A) and protein (B) abundances for essential and nonessential genes shared at diverse growth conditions for *P. aeruginosa* (right) and *S. aureus* (left), revealing elevated expression levels of mRNA and protein for essential genes. (C-D) Histograms of the standard deviations of mRNA (C) and protein (D) abundances for essential and nonessential genes, demonstrating lower variance for essential genes compared to nonessential genes. Histograms were constructed with a bin size of 0.1 using the variation in log2-normalized mRNA and a bin size of 0.05 using protein levels from all bacteria. The standard deviation which bigger than the range showed in the x-coordinate is included to the biggest bin. (E-F) Scatterplot with associated Spearman rank correlation coefficient (ρ) for essential genes (red circles) (n_E = 137 for *S. aureus*, n_E = 294 for *P. aeruginosa*) and nonessential genes (blue circles). All p-values were calculated using a Wilcoxon rank sum test. (F) Spearman rank correlation coefficients for condition-independent essential and nonessential genes for *P. aeruginosa* and *S. aureus*, showing that essential genes have higher mRNA-protein correlations in Gram-positive and Gram-negative bacteria.

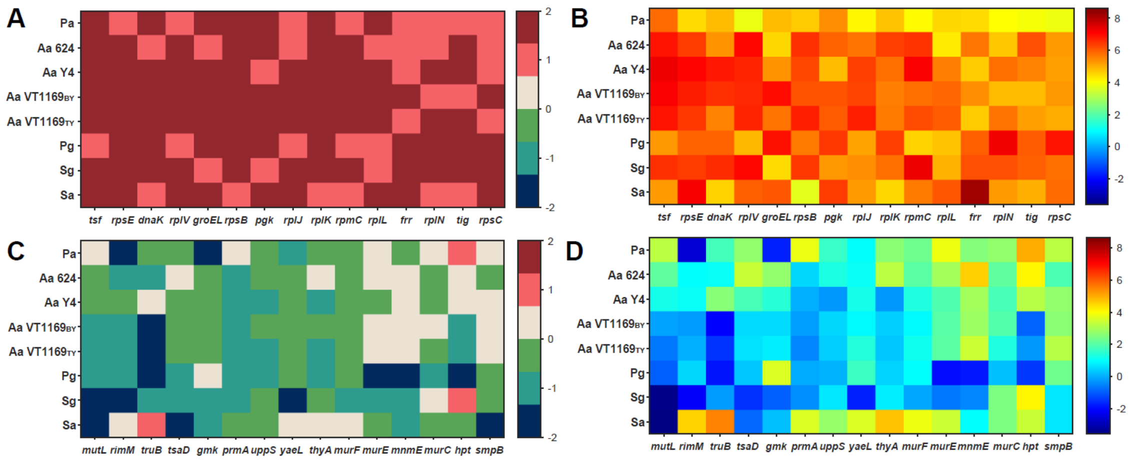


Figure 18. Outlier genes whose mRNAs were least predictive of protein levels are identified by standardized residuals and their associated protein-to-mRNA (PTR) ratios across diverse bacteria

(A-D) The top (A) and bottom 15 genes (C) are ranked by the ranked average of standardized residuals, with maximum values set at 2 or -2 for the maximum values. (B) The PTR ratios of the top 15 outlier genes are largely similar with a higher number of species exhibiting high PTRs (red and orange) and no genes displaying low PTRs (indicated in blue). (C-D) Bottom 15 genes, characterized by the lowest mean standardized residuals, show varied PTR ratios among all bacteria species. Bottom 15 genes contain many genes with low PTRs (denoted by blue and cyan) and some genes displaying high PTR (orange) in some species and displayed large dynamics range of the PTR ratios. Assessment of the PTR ratios for outlier genes highlights the variability in the top and bottom genes.

3.3 Protein levels from orthologous genes have higher correlation than the corresponding mRNAs

Previous studies in multiple eukaryotes have demonstrated a higher correlation in the abundances of orthologous proteins compared to their corresponding mRNA levels (213-215). To determine if this was also observed in bacteria, we quantified protein-protein correlations by calculating the Spearman rank correlation of orthologous proteins within each pair of bacteria (**Figure 19**C) and similarly assessed the rank correlations for their respective mRNAs (**Figure 19**B). Of the 36 pairs considered, the correlation in protein abundances was greater than that of mRNAs in 34 cases (**Figure 19**), highlighting a significantly higher conservation of protein abundances compared to the corresponding mRNAs ($p < 10^{-6}$, Wilcoxon signed rank test). One caveat of this study is that there are larger number of *A. actinomycetemcomitans* species compared to other species in our dataset, highlighting higher correlations and more orthologous genes. We investigated whether observed trend is still true by subsampling our datasets to include only one *A. actinomycetemcomitans* species: the same trend was observed and highly statistically significant in in all four scenarios (**Figure 20**, $p < 10^{-4}$, Wilcoxon signed rank test).

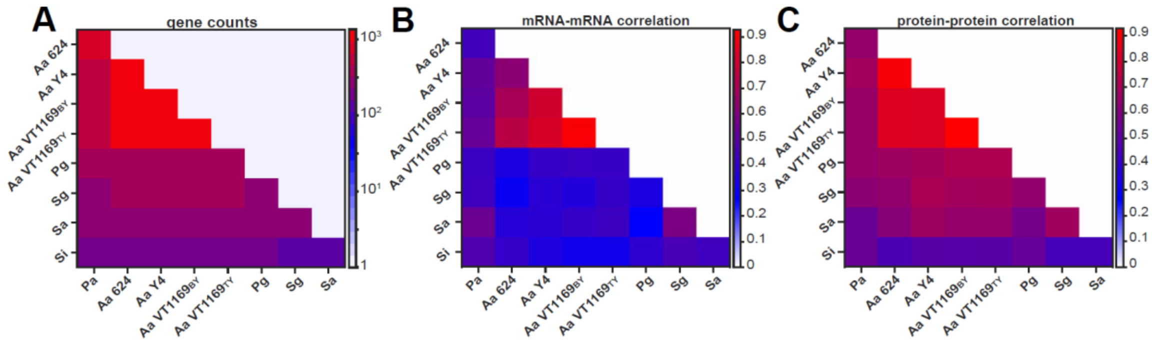


Figure 19. The correlation between orthologous protein abundances is stronger than that of corresponding mRNA levels

(A) Each pairwise group of two species were associated with different sets of orthologous genes among species on the x-axis and y-axis, with a total of 36 combination of all pairwise groups of prokaryotic species. (B-C) The pairwise correlations between mRNA or protein levels from one species (x-axis) and those from another (y-axis) were quantified using Spearman rank correlation coefficients (ρ). The color gradient, ranging from blue (weaker correlation) to red (stronger correlation), visually represents the correlation among each pairwise group. The analysis of mRNA-mRNA (B) and protein-protein (C) correlations across various prokaryotic species demonstrates a stronger correlation for protein abundances ($p < 10^{-6}$). P-values were calculated using Wilcoxon signed rank test.

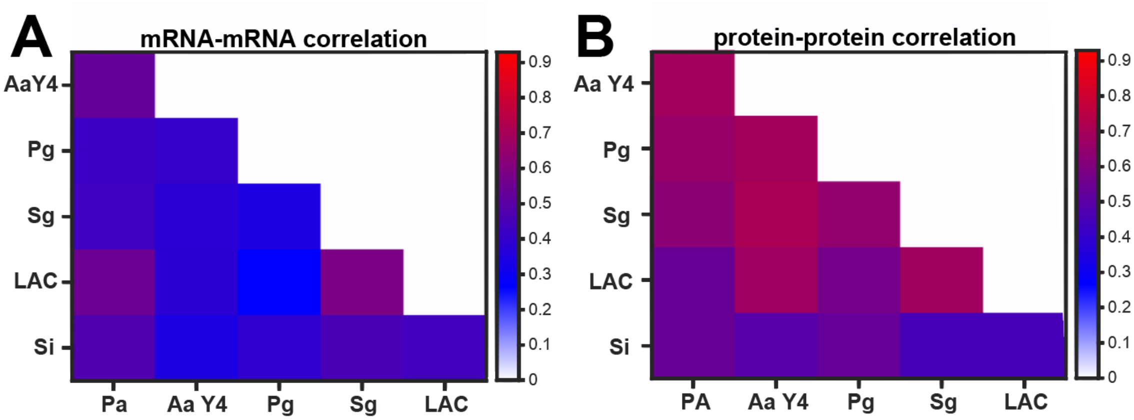


Figure 20. Subsampling analysis including only one *A. actinomycetemcomitans* species *Aa* Y4

Examining the pairwise correlations between mRNA or protein levels using subsampled datasets with only one *A. actinomycetemcomitans* species Aa Y4. P-values were calculated using Wilcoxon signed rank test ($p < 10^{-4}$).

Table 3. Strain information and growth conditions in lab

Species	Strain code	Growth lifestyle	Growth conditions
Aggregatibacter actinomycetemcomitans 624	Aa 624	biofilm on membrane (exponential growth)	Autoclaved TSAYE [a] plates, 37ºC, oxic (5% CO_2)
Aggregatibacter actinomycetemcomitans Y4	Aa Y4	planktonic	TSBYE [b] broth, filter-sterilized, 37ºC, oxic (5% CO2), 165rpm shaking
Aggregatibacter actinomycetemcomitans VT1169	Aa $VT1169_{TY}$	planktonic	TSBYE broth, filter-sterilized, 37ºC, oxic (5% CO2), 165rpm shaking
Aggregatibacter actinomycetemcomitans VT1169	Aa $VT1169_{BY}$	planktonic	BHIYE [c] broth, filter-sterilized, 37ºC, oxic (5% CO2), 165rpm shaking
Porphyromonas gingivalis ATCC 33277	Pg	planktonic	TSBYE broth + 5 ug/ml hemin + 1 ug/ml menadione; filter-sterilized, 37ºC anoxic (anaerobic chamber), 110rpm shaking
Streptococcus gordonii DL1.1	Sg	planktonic	TSBYE broth, filter-sterilized, 37ºC, oxic (5% CO2)
Staphylococcus aureus USA300_FPR3757	Sa	planktonic	Autoclaved BHI, 37ºC, oxic, 220rpm shaking
Pseudomonas aeruginosa UCBPP-PA14	PA14	planktonic	Autoclaved MOPS-succinate, 37ºC, oxic, stiring
Sulfolobus islandicus M.16.4	Si	planktonic	DT medium [d], 76-78ºC, oxic, no shaking

[a] TSAYE: autoclaved tryptic soy agar (BD Difco) supplemented with 0.5% yeast extract (Fluka) (TSAYE)

[b] TSBYE: tryptic soy broth (BD Difco) supplemented with 0.5% yeast extract (Fluka) (TSBYE)

[c] BHIYE: Brain Heart Infusion Broth (BD Difco) supplemented with 0.5% yeast extract (Fluka) (BHIYE)

[d] DT medium, which comprised a basal salt medium supplemented with 0.2% w/v Dextrin (Sigma-Aldrich, USA) and 0.1% w/v Ezmix™ N-Z-Amine® A (Sigma-Aldrich, USA)

Table 4. Statistics on transcriptome and proteome and reference genome

Strain code	Genome size (Mbp)	Detected transcriptome	Detected Proteome*	Paired transcriptome and proteome	Reference transcriptome	Proteome ID
PA14	2.6	2,806	1,314	1,310	GCF_000022445.1	UP000001479
Sa	6.5	5,894	3,721	3,721	GCF_000014625.1	UP000000653
Sg	2.9	2,779	825	825	GCF_000013465.1	UP000001939
Pg	2.2	2,031	936	936	GCF_000017005.1	UP000001131
Aa VT1169	2.4	1,998	1,202	1,198	GCF_000010505.1	UP000008842
Aa VT1169	2.1	2,027	1,442	1,442	GCF_001594245.1	GCF_001594245.1
Aa Y4	2.2	1,862	1,616	1,522	GCF_000318155.1	GCF_000318155.1
Aa 624	2.4	2,276	1,682	1,682	GCF_001594265.1	GCF_001594265.1

3.4 Using a gene-specific RNA-to-protein conversion factor to increase the predictability of protein from mRNA levels

Based on the lack of paired transcriptomic and proteomic data for most microbes, a primary goal of this work was to determine if we can improve protein predictability from mRNA for a microbe of interest using RTP conversion factors calculated from other microbes. The basic approach is: (i) to identify orthologous genes that are present in all, or a subset of the microbes tested; (ii) calculate an RTP conversion factor for these genes using data from one or a subset of microbes; and (iii) apply these RTP conversion factors to mRNA measurements from a microbe(s) not used in their calculation to determine if they improve the ability to predict protein levels. While we previously showed that RTP conversion factors calculated from one *P. aeruginosa* strain can improve the mRNA-protein prediction in a second *P. aeruginosa* strain (168), whether this methodology would be successful using data from more distantly related microbes was unknown. To assess the robustness of this approach, we tested how RTP conversion factors improved mRNA-protein predictions within a bacterial species, between bacterial species, and between bacteria and an archaea.

3.4.1 From within one bacterial species to within all bacterial species

Within a bacterial species

We first used the mRNA-protein datasets for three *A. actinomycetemcomitans* strains to assess how RTP conversion factors calculated for one strain improved the mRNA-protein correlation in other strains of the same species or in the same strain grown in a different medium. These results demonstrated a significant increase in mRNA-protein correlation in all cases, with the correlation coefficient increasing on average from 0.51

(range 0.50 to 0.55) to 0.95 (range 0.93 to 0.96) (**Figure 21A** and **Figure 22**, n = 4, $p<10^{-4}$, t-test and p = 0.0143, Wilcoxon rank sum test, p= 0.0625, Wilcoxon signed rank test). These results reveal that, similar to *P. aeruginosa* (204), *A. actinomycetemcomitans* RTP conversion factors calculated from one bacterial strain can significantly improve mRNA-protein correlations in other *A. actinomycetemcomitans* strains, even if different media are used for growth.

Within Gram-negative bacteria

We next used the mRNA-protein datasets from the three strains of *A. actinomycetemcomitans* to assess how application of the RTP conversion factors from this Gram-negative bacterium impacted mRNA-protein correlation in the Gram negative bacteria *P. aeruginosa* and *P. gingivalis*. Application of the *A. actinomycetemcomitans* RTP conversion factors significantly improved mRNA-protein correlation coefficients for orthologous outlier genes in both bacteria; increasing from 0.54 to 0.76 for *P. gingivalis* (**Figure 21C)** and 0.73 to 0.80 for *P. aeruginosa* (**Figure 21B)**. These data indicate that RTP conversion factors can improve protein prediction from mRNA for other Gram-negative bacteria.

Between Gram-negative and Gram-positive bacteria

Next, we assessed the impact of RTP conversion factors calculated from all the Gram-negative bacteria examined in this study (*A. actinomycetemcomitans, P. aeruginosa*, and *P. gingivalis*) on the mRNA-protein correlation coefficients in Gram-positive bacteria. The application of Gram-negative RTP conversion factors to the mRNA levels of Gram-positive bacteria resulted in substantial improvements in the mRNA-protein correlation coefficient for *S. gordonii* from 0.55 to 0.76 (**Figure 21E**) and *S. aureus* from 0.56 to 0.67 (**Figure 21B**). These data indicate that the RTP conversion factors from one bacterial group can improve protein level predictivity from mRNA measurements for distantly-related bacteria.

<u>Within all bacteria</u>

To assess how RTP conversion factors calculated with data from a subset of bacteria impacted mRNA-protein correlation coefficients in other bacteria, we performed leave-one-out analysis. This approach involves excluding one bacterial dataset from our RTP conversion factor calculations, then applying these conversion factors to the left-out transcriptome to determine the predictive accuracy of the observed protein levels (Table 6). This approach consistently demonstrated that the correlations among the predicted and observed protein levels are significantly higher than the original mRNA-to-protein correlations (p=0.0039, Wilcoxon signed rank test, Table 6) under all conditions. These data further support the robustness of RTP conversion factors for improving protein prediction in bacteria (**Figure 21,** p=0.002, Wilcoxon signed rank test).

3.4.2 Between bacteria and archaea

To determine if RTP conversions factors could be used to improve mRNA-protein predictivity across domains of life, we applied RTP conversion factors calculated with data from all bacteria to protein measurements from *S. islandicus*. These data revealed a significant improvement in protein prediction in *S. islandicus*, with an increased correlation coefficient of 0.38 to 0.52 (**Figure 21F**). It should be noted that the number of orthologous genes used in this comparison is low (n = 58), limited by the number of shared outlier genes in bacteria and *S. islandicus*.

Table 5. Genes with high and low PTRs identified using standardized residuals (SR).

Bacteria	# Low PTR	# High PTR
Pa	151	17
Aa Y4	70	4
Aa 624	71	0
Aa VT1169^{TY}	45	10
Aa VT1169^{BY}	43	14
Pg	39	7
Sa	23	8
Sg	20	9
Si	40	5

[a] *Pa* (*P. aeruginosa*), *Aa* Y4 (*A. actinomycetemcomitans* strain Y4), *Aa* 624 (*A. actinomycetemcomitans* strain 624), *Aa* VT1169^{TY} (*A. actinomycetemcomitans* strain VT1169 grown in TSBYE), *Aa* VT1169^{BY} (*A. actinomycetemcomitans* strain VT1169 grown in BHIYE), *Pg* (*P. gingivalis*), *Sa* (*S. aureus*), *Sg* (*S. gordonii*).

[b] Number of high and low PTR genes. High PTR gene are defined by SR>2 and low PTR genes are defined by SR < -2.

Table 6. Leave-one-out model validation of RTP conversion factors across bacteria

Bacteria[a]	ρ_s[b]	$\rho_{corrected}$[c]
Pa	0.73	0.81
Aa Y4	0.37	0.83
Aa 624	0.53	0.83
Aa VT1169^{TY}	0.49	0.88
Aa VT1169^{BY}	0.41	0.83
Pg	0.48	0.77
Sa	0.56	0.68
Sg	0.55	0.77

[a] Acronyms defined in Table 3.

[b] mRNA-protein correlation coefficient of orthologous outlier

genes before application of RTP conversion factors calculated by leaving data from one bacterium out

[c] mRNA-protein correlation coefficient after applying the RTP conversion factor.

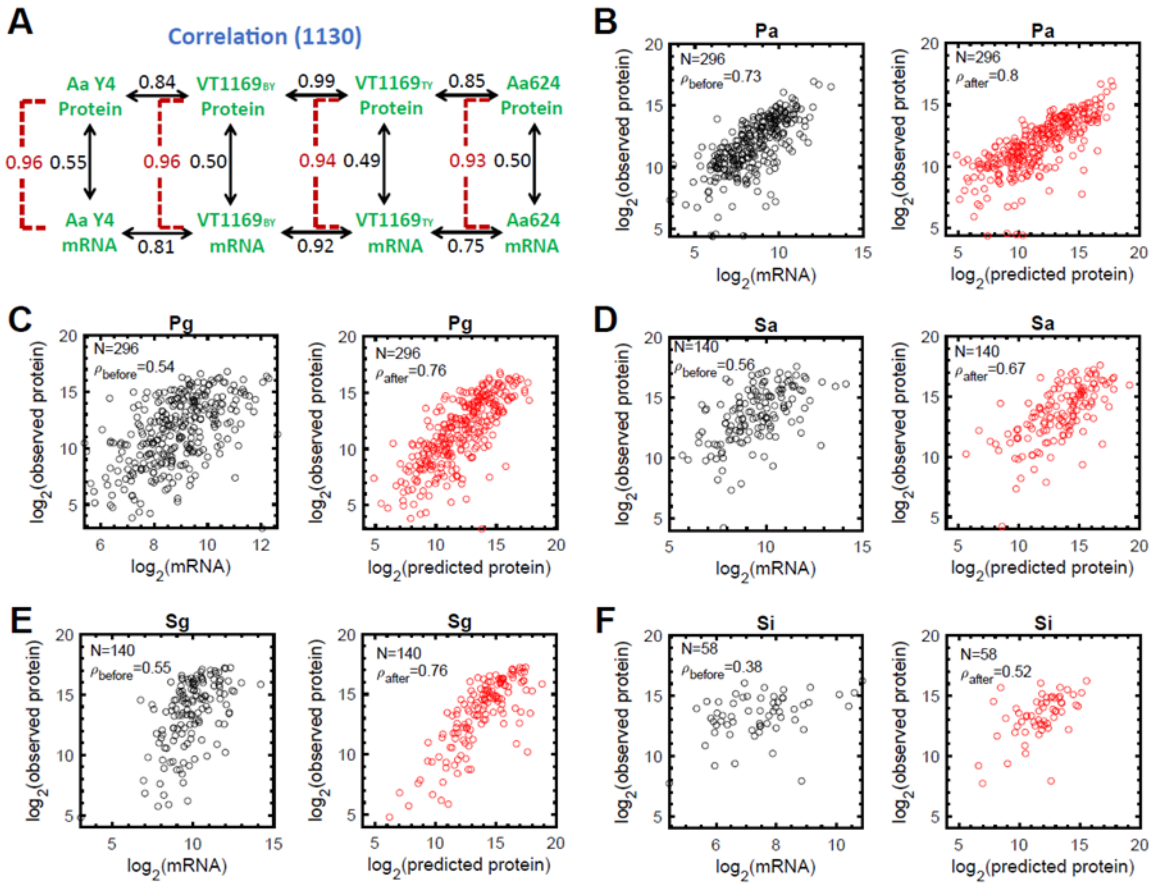


Figure 21. Improvement of predictivity of protein levels from respective mRNA abundances using gene-specific RTP conversion factor across strain levels and across species levels, including Gram-negative, Gram-positive bacteria and archaeon.

(A-C) Implementation of RTP conversion factors derived from *A. actinomycetemcomitans* (*Aa*) improves the predictivity of protein levels from mRNA abundances within various strain backgrounds (A) and across diverse Gram-negative bacterial species (B, C). (A) For strain of three strains *Aa* 624, *Aa* VT1169 and *Aa* Y4, the correlation before correlation ranges from 0.49 to 0.55, while the correlation after correlation ranges from 0.93 to 0.96. (A) In three *A. actinomycetemcomitans* strains (*Aa* 624, *Aa* VT1169, and *Aa* Y4), the predictivity correlation before RTP application ranged from 0.49 to 0.55 and increased to

0.93 to 0.96 after application of conversion factor. (B-C) The application of the RTP conversion factors calculated for *A. actinomycetemcomitans* increases the predictivity of protein based on mRNA levels across other Gram-negative bacteria *P. aeruginosa* (B) increased from and *P. gingivalis* (C). (D-E) For Gram-positive bacteria, *S. aureus* (D) and *S. gordonii* (E), the correlations between predicted and observed protein levels surpass the original mRNA-protein correlations when using conversion factors calculated from Gram-negative bacteria. (F) The application of bacterial RTP conversion factors yields a higher correlation coefficient ρ in archaea *S. islandicus*, outperforming the original mRNA-protein correlation benchmarks. P values were calculated using a Wilcoxon signed rank test and comparing the correlation before and after application of RTP conversion factor ($p=0.002$).

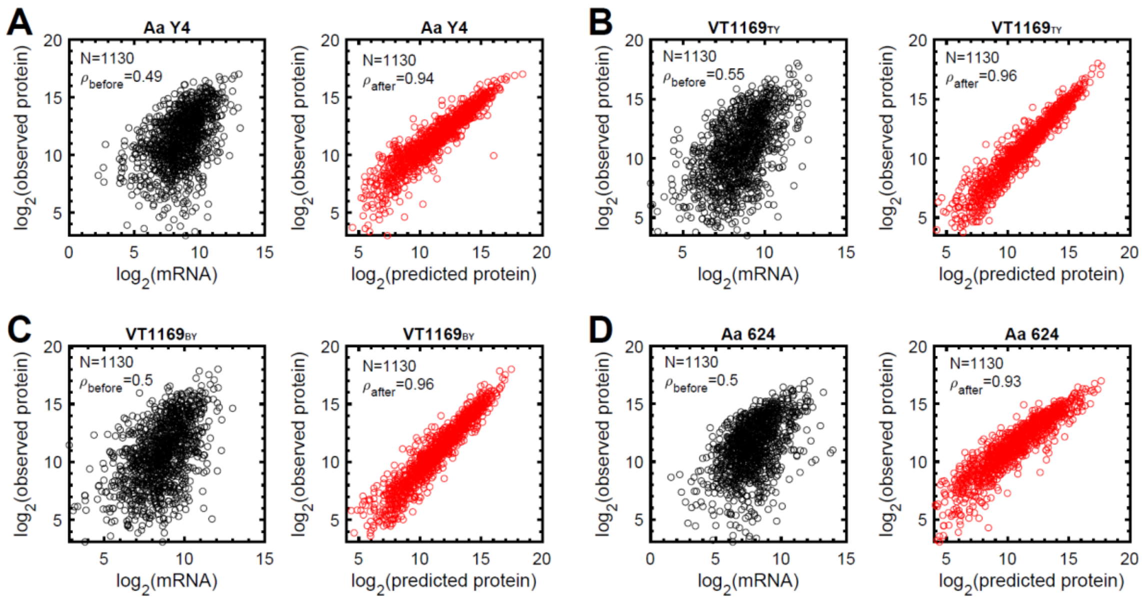


Figure 22. RTP conversion factors calculated from all *A. actinomycetemcomitans* significantly enhances the predictivity of protein levels from mRNA abundances, independent of strain backgrounds

(A-D) Scatterplots of mRNAs and proteins versus predicted proteins and proteins with associated Spearman rank correlation coefficients (ρ). The mRNA-protein relationships significantly improved after the application of RTP conversion factors in different strains, including Y4 (A), VT1169 (B-C), and Aa 624 (D). Among all those A. *actinomycetemcomitans* strains, the correlation coefficients before applying the conversion factor ranged from 0.49 to 0.55, while post-application, the correlations increased to a substantial range of 0.93 to 0.96, highlighting the improvement of correlation coefficients after the application of the RTP conversion factors.

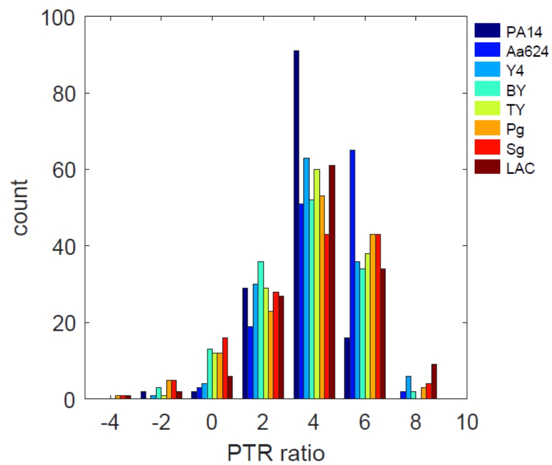


Figure 23. Distribution of the PTR ratios across prokaryotes

The PTR ratios displayed ‘left-skewed’ distribution for all species, indicating a prevalence of lower PTR values compared to higher ones.

3.5 Discussions

The goal of this work was two-fold: (i) to explore mRNA-protein relationship across diverse prokaryotes; and (ii) determine if mRNA-protein ratios from a specific bacterium or a group of bacteria can be used to improve protein predictivity from mRNA levels in other prokaryotes. Our results reveal that RTP conversion factors, while limited to orthologous genes, are highly effective at improving the prediction of protein levels within species, within bacteria, and between bacteria and an archaea. This approach has the potential to substantially improve our understanding of microbial function in natural environments where transcriptomic data exists, particularly for microbes that are difficult to culture or in ones in which paired proteomic-transcriptomic data does not exist. The data generated in this study will be particularly impactful in microbiomes dominated by bacteria that are highly related to those studied here, including the human oral microbiome and the microbiomes of chronically infected CF lungs and chronic wounds. While the *S. islandicus* paired transcriptome-proteome data provided a means of comparing the effectiveness of RTP conversion factors across domains of life, we anticipate that this data will have even greater utility for aiding in protein prediction from mRNA in highly related archaea.

The data generated in this study also allowed us to test whether our previous observation, that *P. aeruginosa* essential genes are more predictive of protein levels than non-essential genes, is observed in other bacteria. Our discovery that essential gene mRNA is highly predictive of protein levels in S. aureus (**Figure 17**) suggests that this phenomenon may be more widespread in bacteria. We focused on *S. aureus* for this analysis as the essential genome has been well-defined, and in the future we plan to perform this analysis on the other bacteria in this study once the essential genome is well defined in these microbes. We did attempt to perform the analysis with a ‘core’ bacterial essential genome, but there were not sufficient genes in this core gene list with sufficient transcriptome/proteome data for analysis. In addition to higher mRNA-protein correlations, *S. aureus* essential genes also had higher mRNA and protein levels than non-essential genes, consistent with studies in *P. aeruginosa* (204) and in yeast and human cell lines (184, 185). Indeed, only 10 out of 294 genes in *P. aeruginosa* and 2 out of 137 genes in *S. aureus* showed mRNA levels below 2^6 (**Figure 17**E). As we previously cussed (204), we do not think that the high protein predictability of essential gene mRNA is due to detection bias, as non-essential genes with similar protein and mRNA levels have lower correlation (**Figure 17**). Instead, it is likely that essential genes act as hubs in protein-protein interaction networks, similar to that observed in human cells (184), and are thus less impacted by post-transcriptional regulatory mechanisms.

Like many other studies, we found that mRNA and protein are generally positively correlated, with the highest Spearman rank correlation coefficients observed in *P. aeruginosa*, and the lowest in *S. islandicus* (**Figure 15; Figure 16**). Correlations were similar across domains, despite different metabolic capacities, genome size and ecological niches. Our data complement the many studies in eukaryotes, spanning from unicellular yeast to multicellular organisms, including tissue-specific mRNA-protein correlations that are within a similar range (**Figure 15; Figure 16**). Similar to the improvement observed when using bacterial RTPs to improve protein predictivity in *S. islandicus*, the data generated in this study will likely not significantly impact protein predictivity from mRNA in eukaryotes, as the number of shared outlier orthologous genes will be minimal.

3.6 Materials and methods

Strains, media, and growth conditions

The strains, media, and growth conditions used for routine cultures and specific experiments are described in Table 3. Bacteria were routinely grown on autoclaved tryptic soy agar (BD Difco) supplemented with 0.5% yeast extract (Fluka) (TSAYE) or specific conditions (described in

Table 3 3). Anaerobic growth was achieved in an anaerobic chamber (Coy) in an 85% N_2, 10% CO_2, 5% H2 atmosphere at 37ºC with shaking at 100 rpm. All *A. actinomycetemcomitans* strains were grown in filter-sterilized tryptic soy broth (BD Difco) supplemented with 0.5% yeast extract (Fluka) (TSBYE) shaking at 165 rpm or membrane-embedded TSAYE agar plates in a 5% CO_2 atmosphere at 37ºC. *S. gordonii* was routinely grown in TSBYE in a 5% CO_2 atmosphere at 37ºC without shaking. Overnight cultures were grown and diluted to an OD_{600} of 0.02-0.05 in a 20 ml volume of prewarmed corresponding media. Cell culture growth was monitored by measuring the OD_{600} of 0.75 ml aliquots and 5ml of samples were collected at OD_{600} of 0.5-0.7.

The hyperthermophilic archaeon *Sulfolobus. islandicus* M.16.4 was cultivated aerobically in Tissue Culture Flasks (Fisher Scientific, USA) at 76-78 °C and pH 3.5 without shaking in DT medium, which comprised a basal salt medium supplemented with 0.2% w/v Dextrin (Sigma-Aldrich, USA) and 0.1% w/v Ezmix™ N-Z-Amine® A (Sigma-Aldrich, USA) (216). Exponentially growing cells were diluted to an OD_{600} of 0.01 in a 130 ml DT medium and their growth was monitored over an 8-day incubation period by measuring the optical density at 600 nm. The experiments were conducted with 4 biological replicates.

Sample collection, library preparation for RNA-seq

A 5 ml sample of bacterial culture was collected and mixed with 20mRNAlater, which was incubated at 4°C for less than 24hrs and then stored at –80°C. All samples were performed in at least two separate experiments with 4 biological replicates in total. RNA-seq libraries were prepared as previously described for *in vitro* samples (204, 217) except for archaea. For the archaeon *S. islandicus* M.16.4, approximately 2 x 10^9 cells taken from mid-log and stationary phases were collected by centrifugation at 4,000 rpm for 20 min at 4 °C. The resulting cell pellets are flash-frozen in a mixture of dry ice and absolute ethanol, and stored in a -80°C freezer until needed. Total RNA extraction was performed using the *mir*Van™ miRNA Isolation Kit (Invitrogen, USA), followed by DNase I (Thermo Fisher Scientific, USA) treatment, and RNA recovery using an RNA Clean & Concentrator-25 Kit (Zymo Research, USA), according to the manufacturer's instructions. The concentration of RNA was quantified using Qubit™ RNA HS Assay Kit (Invitrogen, USA). The constructed RNA-seq Libraries were sequenced on an IlluminaNextSeq500 in the Molecular Evolution Core using 75-bp single-end runs at the Georgia Institute of Technology.

Bioinformatic analysis for RNA-seq

All RNA-seq reads were trimmed by cutadapt (version 3.2 with Python 3.8.6) with optimal length 22bp (218), mapped using bowtie2 (version 2.4.2) and analyzed as previously described by the Whiteley group (217). Gene counts were tallied to the coding sequences (CDS) using the R package featureCounts (Subread v1.6.3) using the strand-specific option. For correlational analyses, raw data were normalized using Reads per million mapped reads (RPM). Individual genes detected >0.1 million on average were discarded due to hotspot problem. Specifically, only four genes were discarded in archaea *Si* while no genes hit hotspot in *Aa Y4*. All figures were produced in MATLAB using function histogram, scatter, plot, imagesc, and PlotPub.

Sample collection for proteomics mass spectrometry

For each sample, 5 ml of flask culture was collected on ice, was centrifugated at 4ºC at 14,000g for 5 min and washed once in 1 ml of ice-cold sterile PBS buffer. Cell pellets were stored at -80ºC followed by centrifugation and removal of supernatants. Samples were then processed by the Emory Integrated Proteomics Core at Emory University.

Sample processing, Tandem mass tag (TMT) labelling and High pH fractionation

Samples were prep with a modified EasyPep digestion protocol (ThermoFisher Cat# A45733, A40006). Briefly, pellets were suspended in 200ul of EasyPep lysis buffer with Universal Nuclease (Thermo Scientific Pierce) and followed with 3 rounds of sonication on ice (5 min in total). An aliquot (40 µg) of each sample was reduced and alkylated using 10mM TCEP (Thermo Scientific Pierce) and 40mM (chloroacetamide) CAA. Samples were incubated at 95°C for 10 mins and then cool to room temperature. Digestion was carried out overnight with 4µg Trypsin/Lys-C Protease Mix (Thermo Scientific Pierce). Samples were desalted with EasyPep cleanup columns following EasyPep clean-up protocol (ThermoFisher Cat# A40006) and dried down by SpeedVac concentrator (LabConco). For each sample, peptides were labelled with TMTPro isobaric tags (ThermoFisher Cat# A44520) as previously described protocols (219). High pH fractionation was performed essentially based on a slightly modified protocol described here (220).

Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS)

LC-MS/MS were carried out on a Dionex Ultimate 3000 rapid separation LC (RSLC)nano system (Thermo Fisher Scientific) and Lumos Orbitrap mass spectrometer (Thermo Fisher Scientific). Peptides were separated on an in-house assembled capillary column (15 cm × 150 μm internal diameter, Waters, Cat# 186008814) packed with 1.9 mm Reprosil-Pur C18 beads (Dr. Maisch, Ammerbuch, Germany) and eluted over a 20 min gradient using a Dionex Ultimate 3000 rapid separation LC (RSLC)nano system (Thermo Fisher Scientific). Mass spectrometry was performed with a high-field asymmetric waveform ion mobility spectrometry (FAIMS)pro equipped Orbitrap Eclipse mass spectrometer (Thermo Fisher Scientific) in positive ion mode using data-dependent acquisition with 2 x 1.5 s top speed cycles. Each cycle consisted of one full MS scan followed by as many MS/MS events that could fit within the given 1.5 s cycle time limit. MS scans were collected at a resolution of 120,000 (410–1600 m/z range, 4×10^5 AGC, 50 ms maximum ion injection time, FAIMS compensation voltage of -45 and -65). All higher energy collision-induced dissociation (HCD) MS/MS spectra were acquired at a resolution of 30,000 (0.7 m/z isolation width, 38% collision energy,100% AGC target, 54 ms maximum ion time, TurboTMT on). Dynamic exclusion was set to exclude previously sequenced peaks for 30 s within a 10-ppm isolation window.

Protein search parameters, mass spectrometry quantification, and data analysis

Raw mass spectrometry data files were processed with Proteome Discoverer (Thermo Fisher Scientific v3.0) and search engine Sequest HT. The spectra were searched against the proteome database available in Uniprot or NCBI (Table 4), and a contaminant database called the common repository of adventitious proteins (cRAP release 2018). The data were searched with MS peptide tolerances of 20 ppm and MS/MS tolerance for identification of 0.05Da for precursor. Quantification was achieved by calculating the sum of most confident centroided reporter ions within a tolerance of 20ppm. Trypsin was selected as the enzyme and 2 missed cleavage sites were allowed. Database searches were performed with following modifications: static modifications: carbamidomethylation of cysteine; variable modifications: oxidation of methionine, and acetylation of N-terminal protein, loss of methylation of methionine, loss of methylation and acetylation of methionine. Peptides with minimum length of six amino acid length and labeled as “IsMasterProtein” were considered for identification. Proteins were only considered identified when observed in four replicates of one sample group. The results were then exported to MS Excel for further data interpretation and statistical analysis. For correlational analyses, raw protein abundance data were normalized to the same depth of 10 million (median of total protein abundance) and ensure standardized comparison across diverse species.

Orthology analysis

To get orthologous genes across diverse bacteria and archaea, we performed orthology analysis using software OrthoFinder2 (version 2.5.4) (221, 222). All the proteome sequence for all prokaryotes (Table 4) were included in the same folder as the input for OrthoFinder. Multiple Sequence Alignment (MSA) was chosen as orthology inference method. BLAST (version 2.5) was chosen as the search method to identify orthologs, which increase accuracy by 1-2% over default method DIAMOND and runtime were optimized using Georgia Tech PACE (223, 224). The output files with protein accessions were then exported to Matlab and mapped with paired transcriptome and proteome for further data analysis.

Pairwise correlation analysis among all bacteria

To calculate all pairwise correlations, we first listed all 36 possible pairwise combinations across all 9 conditions. Specifically, the pairwise combination of PA14-Aa624 and Aa624-PA14 was treated as the same pairwise combination. Each pairwise combination contains a different number of gene numbers due to distinct evolutionary distance between two species (**Figure 19. The correlation between orthologous protein abundances is stronger than that of corresponding mRNA levels**). For example, the pair Aa624 and Aa Y4 contains 1341 orthologous genes with both mRNA and protein detected but the pair Aa624 and PA14 only contains 644 orthologous genes. For each pairwise combination, we calculated the mRNA-mRNA correlation among two bacteria using Spearman rank correlation and repeat this for all 36 possible pairwise groups. The same process was then repeated to calculate protein-protein correlation for all pairwise correlations. All the p-values for the mRNA-mRNA correlation and protein-protein correlations were calculated. The graph was created in Matlab using "imagesc" function, and the colorbar was generated by "cptcmap-pkg" function.

Outlier analysis

We first fitted a linear regression model for each species and then calculated standardized residuals (SR) for all genes. To identify outlier genes in each species, we adopted 2 standards deviation (SD) as a threshold for SR. Thus, we categorized genes with SR>2 as high PTR genes and genes with SR<-2 as low PTR genes. For example, *Pg* has 39 low PTR genes and 7 high PTR genes out of 1198 detected genes (Table 5). As a stringent threshold for outlier genes, 4% of genes in Pg were determined as outlier genes.

Instead of employing one single threshold, we considered SR values as a spectrum and categorized all genes from each species into distinct bins based on their SR values. Specifically, outlier genes with extremely high PTR and low PTR were color-coded in purple and red, respectively. To identify outlier genes shared across all bacteria, we then focused on the 140 orthologous genes and arranged them based on associated mean SR values in a descending order. Top 15 genes were listed along with their associated SR values across all bacteria and provided with gene names of *P. aeruginosa*.

RTP conversion factor calculation and leave-one-out validation and subsampling methods

To predict protein level from mRNA dataset, we identify conditions for both training (RTP_j) and testing datasets ($P_{i,j}^{pred}$) and their associated number of orthologous genes, and calculated RTP conversion factors and performed leave-one-out validation. The gene-specific RTP conversion factor was calculated as described in the formula:

$$RTP_j = \frac{\sum_{i=1}^{N=species} P_{i,j} - R_{i,j}}{N} \text{ for gene j}$$

The calculation of the predicted protein level is described as follows:

$$P_{i,j}^{pred} = RTP_j + R_{i,j} \text{ for gene j}$$

To perform leave-one-out validation, we calculated the left-out conversion factor (LF) as follows:

$$LF_j = \frac{\sum_{i=1}^{N=species-1} P_{i,j} - R_{i,j}}{N-1} \text{ for gene j}$$

$P_{i,j}$ and $R_{i,j}$ represent for the log2-transformed protein and mRNA levels for orthologous gene j and species i. The number of j equals the orthologous genes present in the training datasets and i denotes the number of species included in the training dataset. For instance, the calculation of bacterial RTP conversion factors RTP_j involves 140 orthologous genes (j = 1, 2, ….,140) and 8 conditions (N=8) including all Gram-positive and Gram-negative bacteria. When applying bacterial RTP conversion factors to archaea, we focused on 58 orthologous genes present in both bacteria and archaea (j= 1, 2, ….,58). For validation analysis, the calculation of left-out conversion factors LF_j involves the same number of orthologous genes (N_{gene}=140) but with different combinations of all possible 7 conditions out of 8 conditions by leaving one species out each time. When utilizing left-out conversion factors to predict for left-out transcriptome, we focused on 140 orthologous genes present in all bacteria to predict for the left-out transcriptome (j= 1, 2, ….,140).

To perform subsampling and exclude the possibility of such effect caused by stochastic events, we calculated the randomized conversion factor as described in the formula:

$$RTP_j^{random} = \frac{\sum_{i=1}^{N=species} P_{i,k} - R_{i,h}}{N} \text{ for gene j}$$

$P_{i,k}$ denotes the log2-transformed protein level of a randomized non-orthologous gene k in species I, while $R_{i,h}$ represents the log2-transformed mRNA levels of a non-orthologous gene h in species i. Briefly, the calculation of RTP_j^{random} for gene j involved randomly selected protein level of non-orthologous gene k and mRNA levels of gene h. Since the combination of extremely high protein level and low mRNA levels doesn't exist, we need to restrict the degree of freedoms and avoid some combinations of mRNA and protein levels. Thus, we limited our selection of genes with paired mRNA and protein levels (k=h) to maintain a biologically relevant PTR ratio from each species.

To quantify the changes before and after applying the conversion factors, we calculated Spearman rank correlation between the observed mRNA and protein levels $\rho_{observed}$ and correlation among the predicted proteins and observed proteins $\rho_{corrected}$. The significance levels were determined using a paired non-parametric rank test called Wilcoxon signed-rank test.

CHAPTER 4. CONCLUSIONS AND FUTURE DIRECTIONS

4.1 Impact of growth rate on the Protein-mRNA Ratio in *P. aeruginosa*

4.1.1 Summary of results

Our understanding of how bacterial pathogens colonize and persist during human infection has been hampered by the limited characterization of bacterial physiology during infection and a research bias toward *in vitro*, fast-growing bacteria. Recent research has begun to address these gaps in knowledge by directly quantifying bacterial mRNA levels during human infection, with the goal of assessing microbial community function at the infection site. However, mRNA levels are not always predictive of protein expression, which are the primary functional units of a cell. Here, we present a system-level study to disentangle two hypotheses, 1) Is mRNA is a good proxy to predict protein abundance; and 2) How growth rate impacts bacterial physiology at the global scales using transcriptome and proteome as tools.

We used carefully controlled chemostat experiments to examine the relationship between mRNA and protein levels across four growth rates that cover *in situ* growth rates during chronic human infection in the bacterial pathogen *P. aeruginosa*. We found a genome-wide positive correlation between mRNA and protein abundances across all growth rates (ρ = 0.55 to 0.63, **Figure 8**), indicating that the growth rate has little impact on this correlation. We also demonstrated that genes required for *P. aeruginosa* PA14 viability are highly expressed, are more stable in their expression levels, and have stronger mRNA-protein correlations than nonessential genes across a range of growth rates (Figure 9). To identify genes whose mRNA abundances poorly predict protein abundances, we used a statistical measure called the standardized residual to identify outlier genes that have extreme protein-to-mRNA ratios. Based on the distribution of PTR ratios, we found it is more common to have genes with high mRNA levels but low protein levels than to have genes with low mRNA and high protein levels. Lastly, we calculated an RTP conversion factor to improve mRNA predictions of protein levels. The application of the RTP conversion factor to publicly available transcriptome data sets was highly robust, enabling the more accurate prediction of *P. aeruginosa* protein levels across strains and growth conditions (**Figure 12**A and B). Finally, the RTP conversion factor was applied to *P. aeruginosa* human CF infection transcriptomes to provide greater insights into the functionality of this bacterium in the CF lung (**Figure 12**C and D). This study addresses a critical problem in infection microbiology by providing a framework for enhancing the functional interpretation of bacterial human infection transcriptome data.

4.1.2 Future directions

<u>Prediction of growth rates using transcriptomic features</u>

The identification of a subset of genes whose expression levels increase or decrease with growth rates has the potential to serve as transcriptional signatures to predict for growth rates, particularly in complex bacterial communities where direct measurement of bacterial growth rates is challenging. One follow-up study to employ machine learning algorithms like LASSO and Ridge regression to identify the most predictive transcriptomic features across a diverse range of growth rates (225). Given the versatile metabolic capacities of *P. aeruginosa,* one important aspect of this study is to generate a comprehensive training dataset that covers a spectrum of growth environments and strain diversity. A chemostat study in yeast has provided an elegant way to tackle this by growing yeast under six nutrient limitations (Glucose, Sulfur, Leucine, Phosphate, Nitrogen, and Uracil) and six diverse growth rates (0.05 hr^{-1} through 0.3 hr^{-1})) (226). However, this study presented a limitation in the variability of transcriptome measurement using only one replicate for each of 36 conditions. In our work, we addressed this limitation by generating four replicates for each condition for a more robust analysis and focused on four growth rates under the same growth media.

In the future, more efforts and resources are needed to measure the *P. aeruginosa* transcriptome under diverse nutrient limitations and across diverse strains as demonstrated in this context. This will facilitate the identification of gene expression signatures using machine learning models that reliably select transcriptomic features to predict growth rates that are independent of nutrient-specific responses, and thus pave the way for the application of transcriptome data to measure growth rates in complex bacterial ecosystems.

Novel gene expression patterns across growth rates

Assessment of gene expression across a diverse range of growth rates revealed diverse patterns, including simple patterns, such as monotonically increasing and decreasing with the growth rates, and more complicated patterns, such as with U-shape or upside-down U-shape expression patterns. Here we used a statistical way to identify those patterns based on their pairwise comparisons of the four generation times. However, we can adopt another strategy with no prior assumptions about the interested gene shapes and conducted machine learning methods such as k-means clustering on those datasets to search for interesting shapes. I anticipate that the clustering of another type of patterns such as L-shape and upside-down L-shape in the k-means clustering analysis. The L-shape conveys an interesting biology that those genes are maintained at high expression levels at slower growth rates but have low expression levels at faster growth rates. The change from high expression levels to low expression level happen suddenly and function like a “On” and “Off” switch.

After identifying of genes associated with complex expression patterns, it is also important to uncover the underlying biology on how genes response to diverse growth rates. Genes involved in siderophore biosynthesis displayed U-shape patterns, displaying high expression levels at both the highest and lowest growth rates but had low gene expression levels at the moderate growth rates. During iron starvation, RhyB promotes siderophore production and down-regulates genes that encode nonessential iron acquisition proteins (227, 228). Under iron rich conditions, siderophore production is repressed at the transcriptional levels by Fur as the master regulator in *E.coli* (229). It may partially explain why siderophore biosynthesis genes showed a growth-rate dependent patterns in response to environmental factors such as diverse growth rates and are regulated at transcriptional levels and post-transcriptional level.

<u>*Utilizing in vivo proteome to validate our model predictions*</u>

The primary motivation for the predictive framework using RTP conversion factors is due to the challenges to assess proteomes *in vivo*, especially in the clinical CF sputum samples. It proposes a solution that we can utilize mRNA datasets to infer protein abundances. The challenge lies in these aspects 1) low bacterial load in the human tissues or CF sputum; 2) MS-based proteomics method is a “nonamplification” process and pose limitations on detection ranges; 3) detection bias towards some membrane proteins, lowly abundant proteins, poorly ionized proteins. With the advancement of MS-based technologies and proteome extraction protocols, especially single-cell proteome, we will be able to measure the bacterial specimen *ex vivo* with high coverage and fidelity of proteomes in the future. In this aspect, those data are valuable to validate our *in vitro* predictions and refine our framework to enable the prediction of proteins *ex vivo* using transcriptome datasets (Figure 12 and Figure 13).

Furthermore, *in vivo* proteome work allows comprehensive comparisons among the proteomes *in vitro* and *in vivo* with high precision. Comparative studies, including differential expression analysis, impose constraints on the protein coverage because the protein levels need to be detected in both *in vitro* and *in vivo* studies. For example, the proteome coverage for *P. aeruginosa in vitro* is around 78% but appears to be ~20% *in vivo* (1304 proteome) with bacteria-enriched CF sputum samples (230). However, it is unknown whether those disparity in protein coverages was due to the lowly abundant protein below the detection limit or due to unexpressed genes during infection. Despite all those challenges, the *in vivo* proteome is the valuable to understand bacterial physiology during infection by providing a direct measure of bacterial function.

4.2 Using conserved protein to mRNA ratios across domains to enhance functional discovery across prokaryotes

4.2.1 Summary of results

Understanding the biology of native microbial communities is hindered by the lack of robust functional data for the microbes within these communities, impeding the development of relevant in vitro and in vivo models. Quantifying gene expression via transcriptomics to infer microbial function has proven successful in diverse environments, such as the human microbiome, marine ecosystems, and soil. However, the reliability of mRNA levels as predictors of protein levels, which are crucial for understanding microbial function, remains uncertain. Recent studies have generated large datasets to assess the PTR ratios of genes, revealing that PTRs are often conserved across microbial strains and environments. This has enabled the calculation of RTP conversion factors, enhancing functional predictions from transcriptomic data. Despite the potential of RTP conversion factors, their calculation requires extensive datasets, which are costly and not widely available. We addressed this by analyzing comprehensive datasets from diverse prokaryotes and one archaea, identifying conserved PTRs and orthologous genes with consistent RTP ratios. This approach allows for the broad application of RTP conversion factors, improving protein prediction accuracy from mRNA in distantly related microbes, and advancing our understanding of microbial functions in complex ecosystems.

4.2.2 Future directions

<u>Integration of mRNA half-life dynamics to the mRNA-protein relationships in P. aeruginosa and A. actinomycetemcomitans</u>

Gene expression dynamics are influenced by the intricate interplay between the synthesis and degradation of mRNAs and proteins. The half-life of mRNAs is one of critical factors in controlling the overall expression levels in bacteria. Numerous computational studies have employed ordinary differential equations (ODEs) to unravel the complex interplay between mRNA stability, transcription, translation, and protein degradation and predict time-dependent changes in mRNA and protein levels. For instance, Elowitz and Leibler constructed a synthetic oscillatory network in *E. coli*, and captured the oscillation dynamics through an ODE model of transcriptional regulation with negative feedback loops including mRNA and protein half-lives (231). In addition to computational analysis, researchers have employed experimental approaches to measure the mRNA half-lives across genome-wide level in some model systems, including some prokaryotes, such as *E.coli* (232), *S. aureus* (233), archaea (234) and some eukaryotes including yeasts (235), insects (236), and human cell lines (237).

We propose to assess the impact of mRNA half-life dynamics to mRNA-protein relationships in the bacteria, highlighting the improvement of the predictability of protein levels from the abundances and the half-lives of mRNAs across two Gram-negative species. While extensive research has been conducted to quantify genome-wide mRNA half-lives in some bacterial model systems (232, 233, 238), it remains an unexplored area for *P. aeruginosa*. Here we propose that we can integrate the half-lives and abundances of mRNA to improve the predictivity of protein. We will include two species of interest, fast growers *P. aeruginosa* and slow growers *A. actinomycetemcomitans*, to investigate their mRNA half-life dynamics and evaluate bacterium-specific mRNA decay. This comparative analysis will provide insights into the regulatory mechanisms that govern gene expression among fast and slow growing bacteria and pave the way to the development of more accurate models for prediction of protein from mRNA levels.

We aim to determine RNA half-lives on a genome-wide scale for fast growers *P. aeruginosa* and slow growers *A. actinomycetemcomitans.* We first determine the minimum inhibitory concentration (MIC) of rifampicin against two *P. aeruginosa* lab strains, PAO1 and PA14. Based on our preliminary data, PA14 is resistant to rifampicin with a MIC >125ng/ml and Pao1 is sensitive to MIC (~16ng/ml, data not shown). In addition, we also tested the MIC in the slow grower *A. actinomycetemcomitans* strains and found *Aa* VT1169 is resistant to rifampicin, and other rifampicin derivatives, including rifabutin, rifapentine and rifaxinin. To narrow down our focus, we limited the antibiotics to rifampicin rather than its derivatives and identified the strains demonstrating sensitivity to rifampicin. *A. actinomycetemcomitans* Y4 and *P. aeruginosa* PAO1 turns of the model system in this study and we subsequently performed MIC tests in 96-well plates for precise evaluation. Once the MIC is established, we then use engineered fluorescent strains as a live reporter to visualize if any mRNA is newly synthesized and translated into fluorescent protein at the concentration of 100-fold MIC. This step is crucial, as it ensures that the concentration of rifampicin is sufficient to inhibit the mRNA synthesis and that the subsequent time-series RNA-seq experiments will not be confounded by incomplete transcriptional termination for some genes. After confirming the adequate rifampicin concentration, a time-series RNA-seq was conducted to measure the half-life of all bacterial mRNAs. It is critical to determine a rational range of time points and harvest those samples in the same batch of bacterial culture. In *B. subtilis*, a previous study has determined that around 80% of detected mRNAs exhibit half-lives < 7 minutes (239). This provides an approximate range of mRNA half-life distribution in some fast-growing bacteria but the half-life distribution in slow-growing bacteria is still largely unknown. Thus, we plan to harvest samples at multiple time points for *Aa* Y4 with a maximum time point of 60mins (1, 3, 5, 7, 9, 17, 40 and 60 minutes) and PAO1 with a maximum time point of 20mins (1, 3, 5, 7, 9, 20mins). This approach will allow us to capture the time-series dynamics of mRNA degradation and preserve some samples than needed, especially around the 1-hour mark in *Aa* Y4.

<u>mRNA-protein relationship with a balanced composition of bacterial model systems</u>

In this work, we described the mRNA-protein relationships across kingdoms, highlighting the application of RTP conversion factors to increase the predictability of protein from mRNA levels across diverse prokaryotes. However, the composition of bacterial species is biased towards Gram-negative bacteria, specifically including *P. aeruginosa, P. gingivalis, and A. actinomycetemcomitans*, with only two Gram-positive representatives (*S. gordonii, S. aureus*) and limited strain diversity from the same species. To address this, prospective research will expand the investigation of mRNA-protein relationships across a broader range of prokaryotes, ensuring a balanced representation of both Gram-negative and Gram-positive species. Alternatively, rather than a balanced composition of bacterial species, follow-up study employing this framework can shift focus to Gram-positive bacteria and emphasize in-depth exploration of the unique physiological aspects of Gram-positive bacteria rather than Gram-negative bacteria.

4.3 Future perspectives on polymicrobial interactions of oral microbiota

Over a century of work has studied the ecology of oral microbiota, demonstrating how spatiotemporal dynamics lead to oral diseases. Despite this rich history of investigation, there remains much to learn about the emergent properties of oral biofilms and how microbe-microbe and microbe-host interactions influence disease. With diverse microenvironments influenced by dramatic nutrient and temperature changes, regular disturbances (such as hygiene practices), and host interactions, the oral cavity is a valuable model system to study biofilm development, ecological succession, and the eco-evolutionary dynamics that shape an organism's niche. Given the enormous complexity of oral communities, we suggest that both bottom-up and top-down approaches are important for the further investigation of microbial interactions and the dynamics of oral biofilms (**Figure 3**; Table 1). Bottom-up approaches using simplified experimental models provide an easy-access system to evaluate biofilm development and the outcome of clinical interventions. For example, the 10-species "Zurich model" has been used to study topics such as biofilm development, virulence factors, and microbial interactions (240, 241). The advancement of sequencing technologies and microscopy techniques offers further opportunities to design and customize *in vitro* models to match the oral community composition, physiology, and spatial patterning (128, 135, 139), and there are significant opportunities to use these model systems to probe diverse questions (Table 1). In contrast, top-down approaches that directly analyze human patient samples will continue to be essential for understanding the functional and spatial dynamics of oral microbial communities.

Research on oral microbes has been critical for our overall understanding of microbial interactions, biofilm development, and spatial patterning and has driven the advancement of technologies for studying microbial communities. Undoubtedly, the oral microbiome provides an accessible and deeply studied system for further exploration of microbial interactions and the role of our microbiota in human health. The future unraveling of the complex dynamics leading to health and disease in the oral cavity will continue to result in new discoveries at the intersection of microbial ecology and pathogenesis.

4.4 Final Discussions

In this thesis, we have expanded our understanding of the mRNA-protein relationship in *P. aeruginosa,* as detailed in Chapter 2, which has inspired an inquiry into whether such framework is exclusive to this bacterium or can extend to other bacterial model systems. Motivated by this, we have proposed the second project, as described in Chapter 3, to assess mRNA-protein relationships across diverse prokaryotes. Moving forward, we propose to explore the impact of mRNA half-life dynamics on mRNA-protein relationships, highlighting the potential to enhance the predictive accuracy of protein levels based on mRNA abundance and half-life measurements.

REFERENCES

1. Shi, Y., G.W. Tyson, and E.F. DeLong, *Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column.* Nature, 2009. **459**(7244): p. 266-9.

2. Frias-Lopez, J., et al., *Microbial community gene expression in ocean surface waters.* Proc Natl Acad Sci U S A, 2008. **105**(10): p. 3805-10.

3. Nuccio, E.E., et al., *Community RNA-Seq: multi-kingdom responses to living versus decaying roots in soil.* ISME Commun, 2021. **1**(1): p. 72.

4. Bang-Andreasen, T., et al., *Total RNA sequencing reveals multilevel microbial community changes and functional responses to wood ash application in agricultural and forest soil.* FEMS Microbiol Ecol, 2020. **96**(3).

5. Cornforth, D., et al., *Pseudomonas aeruginosa transcriptome during human infection.* Proc Natl Acad Sci USA, 2018. **115**: p. E5125-E5134.

6. Duran-Pinedo, A., et al., *Long-term dynamics of the human oral microbiome during clinical disease progression.* BMC Biol, 2021. **19**(1): p. 240.

7. Duran-Pinedo, A.E., et al., *Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis.* ISME J, 2014. **8**(8): p. 1659-1672.

8. Duran-Pinedo, A.E., et al., *Subgingival host-microbiome metatranscriptomic changes following scaling and root planing in grade II/III periodontitis.* J Clin Periodontol, 2023. **50**(3): p. 316-330.

9. Goldberg, J.B., *'Pseudomonas' 99, The Seventh International Congress on Pseudomonas: Biotechnology and Pathogenesis', organized by the American Society for Microbiology, was held in Maui, HI, USA, 1–5 September 1999.* Trends in microbiology, 2000. **8**(2): p. 55-57.

10. Han, K., et al., *Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu.* Scientific reports, 2013. **3**(1): p. 2101.

11. Stover, C.K., et al., *Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen.* Nature, 2000. **406**(6799): p. 959-964.

12. Kung, V.L., E.A. Ozer, and A.R. Hauser, *The accessory genome of Pseudomonas aeruginosa.* Microbiology and molecular biology reviews, 2010. **74**(4): p. 621-641.

13. Klockgether, J., et al., *Pseudomonas aeruginosa genomic structure and diversity.* Frontiers in microbiology, 2011. **2**: p. 10805.

14. Valot, B., et al., *What it takes to be a Pseudomonas aeruginosa? The core genome of the opportunistic pathogen updated.* PloS one, 2015. **10**(5): p. e0126468.

15. Tariq, M.A., et al., *Temperate bacteriophages from chronic Pseudomonas aeruginosa lung infections show disease-specific changes in host range and modulate antimicrobial susceptibility.* MSystems, 2019. **4**(4): p. 10-1128.

16. Diggle, S.P. and M. Whiteley, *Microbe Profile: Pseudomonas aeruginosa: opportunistic pathogen and lab rat.* Microbiology, 2020. **166**(1): p. 30-33.

17. Horcajada, J.P., et al., *Epidemiology and treatment of multidrug-resistant and extensively drug-resistant Pseudomonas aeruginosa infections.* Clinical microbiology reviews, 2019. **32**(4): p. 10-1128.

18. Sousa, A.M. and M.O. Pereira, *Pseudomonas aeruginosa diversification during infection development in cystic fibrosis lungs—a review.* Pathogens, 2014. **3**(3): p. 680-703.

19. Thöming, J.G. and S. Häussler, *Transcriptional profiling of Pseudomonas aeruginosa infections.* Pseudomonas aeruginosa: Biology, Pathogenesis and Control Strategies, 2022: p. 303-323.

20. Govan, J.R.W., D.W. Martin, and V.P. Deretic, *Mucoid Pseudomonas aeruginosa and cystic fibrosis: the role of mutations in muc loci.* FEMS microbiology letters, 1992. **100**(1-3): p. 323-329.

21. Haussler, S., *Biofilm formation by the small colony variant phenotype of Pseudomonas aeruginosa.* Environmental Microbiology, 2004. **6**(6): p. 546-551.

22. Bridier, A., et al., *Resistance of bacterial biofilms to disinfectants: a review.* Biofouling, 2011. **27**(9): p. 1017-1032.

23. Mayer-Hamblett, N., et al., *Pseudomonas aeruginosa in vitro phenotypes distinguish cystic fibrosis infection stages and outcomes.* American journal of respiratory and critical care medicine, 2014. **190**(3): p. 289-297.

24. Nyström, T., *Stationary-phase physiology.* Annu. Rev. Microbiol., 2004. **58**: p. 161-181.

25. Hoehler, T.M. and B.B. Jørgensen, *Microbial life under extreme energy limitation.* Nature Reviews Microbiology, 2013. **11**(2): p. 83-94.

26. Wang, S., et al., *Strategies for zinc uptake in Pseudomonas aeruginosa at the host–pathogen interface.* Frontiers in Microbiology, 2021. **12**: p. 741873.

27. Barer, M.R., *Bacterial growth, physiology and death.* Medical Microbiology. A Guide to Microbial Infections: Pathogenesis, Immunity, Laboratory Diagnosis and Control, 2012: p. 39-53.

28. Narang, A. and S.S. Pilyugin, *Bacterial gene regulation in diauxic and non-diauxic growth.* Journal of theoretical biology, 2007. **244**(2): p. 326-348.

29. Kremling, A., et al., *The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose.* Metabolic engineering, 2001. **3**(4): p. 362-379.

30. Smith, H.L. and P. Waltman, *The theory of the chemostat: dynamics of microbial competition*. Vol. 13. 1995: Cambridge university press.

31. Panikov, N.S., *Microbial growth kinetics.* 1995.

32. Kolpen, M., et al., *Bacterial biofilms predominate in both acute and chronic human lung infections.* Thorax, 2022.

33. Kragh, K.N., et al., *Polymorphonuclear leukocytes restrict growth of Pseudomonas aeruginosa in the lungs of cystic fibrosis patients.* Infect Immun, 2014. **82**(11): p. 4477-86.

34. Weissman, J.L., S. Hou, and J.A. Fuhrman, *Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns.* Proc Natl Acad Sci U S A, 2021. **118**(12).

35. Yang, L., et al., *In situ growth rates and biofilm development of Pseudomonas aeruginosa populations in chronic lung infections*. 2008, Am Soc Microbiol.

36. Palmer, K.L., L.M. Aye, and M. Whiteley, *Nutritional cues control Pseudomonas aeruginosa multicellular behavior in cystic fibrosis sputum.* J Bacteriol, 2007. **189**(22): p. 8079-87.

37. Gonzalez, M.R., et al., *Transcriptome analysis of Pseudomonas aeruginosa cultured in human burn wound exudates.* Frontiers in cellular and infection microbiology, 2018. **8**: p. 39.

38. Wheeler, K.M., et al., *Mucin glycans attenuate the virulence of Pseudomonas aeruginosa in infection.* Nature microbiology, 2019. **4**(12): p. 2146-2154.

39. Cattoir, V., et al., *Transcriptional response of mucoid Pseudomonas aeruginosa to human respiratory mucus.* MBio, 2012. **3**(6): p. 10-1128.

40. Korgaonkar, A., et al., *Community surveillance enhances Pseudomonas aeruginosa virulence during polymicrobial infection.* Proceedings of the National Academy of Sciences, 2013. **110**(3): p. 1059-1064.

41. Doing, G., et al., *Conditional antagonism in co-cultures of Pseudomonas aeruginosa and Candida albicans: An intersection of ethanol and phosphate signaling distilled from dual-seq transcriptomics.* PLoS genetics, 2020. **16**(8): p. e1008783.

42. Pan, X., et al., *PvrA is a novel regulator that contributes to Pseudomonas aeruginosa pathogenesis by controlling bacterial utilization of long chain fatty acids.* Nucleic Acids Research, 2020. **48**(11): p. 5967-5985.

43. Bielecki, P., et al., *Ex vivo transcriptional profiling reveals a common set of genes important for the adaptation of P seudomonas aeruginosa to chronically infected host sites.* Environmental microbiology, 2013. **15**(2): p. 570-587.

44. Turner, K.H., et al., *Requirements for Pseudomonas aeruginosa acute burn and chronic surgical wound infection.* PLoS genetics, 2014. **10**(7): p. e1004518.

45. Harrington, N.E., J.L. Littler, and F. Harrison, *Transcriptome analysis of Pseudomonas aeruginosa biofilm infection in an ex vivo pig model of the cystic fibrosis lung.* Applied and Environmental Microbiology, 2022. **88**(3): p. e01789-21.

46. Dötsch, A., et al., *The Pseudomonas aeruginosa transcriptome in planktonic cultures and static biofilms using RNA sequencing.* PloS one, 2012. **7**(2): p. e31092.

47. Patell, S., et al., *Comparative microarray analysis reveals that the core biofilm-associated transcriptome of Pseudomonas aeruginosa comprises relatively few genes.* Environmental microbiology reports, 2010. **2**(3): p. 440-448.

48. Dötsch, A., et al., *The Pseudomonas aeruginosa transcriptional landscape is shaped by environmental heterogeneity and genetic variation.* MBio, 2015. **6**(4): p. 10-1128.

49. Schlichting, C. and M. Pigliucci, *Phenotypic evolution: a reaction norm perspective.* 1998.

50. Bielecki, P., et al., *In vivo mRNA profiling of uropathogenic Escherichia coli from diverse phylogroups reveals common and group-specific gene expression profiles.* MBio, 2014. **5**(4): p. 10-1128.

51. Cornforth, D.M., et al., *Pseudomonas aeruginosa transcriptome during human infection.* Proceedings of the National Academy of Sciences, 2018. **115**(22): p. E5125-E5134.

52. Rossi, E., et al., *High-resolution in situ transcriptomics of Pseudomonas aeruginosa unveils genotype independent patho-phenotypes in cystic fibrosis lungs.* Nature communications, 2018. **9**(1): p. 3459.

53. Kordes, A., et al., *Genetically diverse Pseudomonas aeruginosa populations display similar transcriptomic profiles in a cystic fibrosis explanted lung.* Nature communications, 2019. **10**(1): p. 3397.

54. Valli, R.X.E., M. Lyng, and C.L. Kirkpatrick, *There is no hiding if you Seq: recent breakthroughs in Pseudomonas aeruginosa research revealed by genomic and transcriptomic next-generation sequencing.* Journal of medical microbiology, 2020. **69**(2): p. 162-175.

55. Cahill, D.J., et al., *Proteomics of microorganisms: fundamental aspects and application*. Vol. 83. 2003: Springer Science & Business Media.

56. Jungblut, P.R. and M. Hecker, *Proteomics of Microbial Pathogens.* 2007.

57. Jaffe, J.D., H.C. Berg, and G.M. Church, *Proteogenomic mapping as a complementary method to perform genome annotation.* Proteomics, 2004. **4**(1): p. 59-77.

58. Sulaiman, J.E. and H. Lam, *Proteomic investigation of tolerant Escherichia coli populations from cyclic antibiotic treatment.* Journal of proteome research, 2020. **19**(2): p. 900-913.

59. Schmidt, A., et al., *The quantitative and condition-dependent Escherichia coli proteome.* Nature biotechnology, 2016. **34**(1): p. 104-110.

60. Eymann, C., et al., *A comprehensive proteome map of growing Bacillus subtilis cells.* Proteomics, 2004. **4**(10): p. 2849-2876.

61. Kühner, S., et al., *Proteome Organization in a Genome-Reduced Bacterium.* Science, 2009. **326**(5957): p. 1235-1240.

62. Ahrné, E., et al., *Critical assessment of proteome‐wide label‐free absolute abundance estimation strategies.* Proteomics, 2013. **13**(17): p. 2567-2578.

63. Becher, D., et al., *A proteomic view of an important human pathogen–towards the quantification of the entire Staphylococcus aureus proteome.* PloS one, 2009. **4**(12): p. e8176.

64. Kunnath-Velayudhan, S., et al., *Dynamic antibody responses to the Mycobacterium tuberculosis proteome.* Proceedings of the National Academy of Sciences, 2010. **107**(33): p. 14703-14708.

65. Klinger, J.D., et al., *Antibodies to proteases and exotoxin A of Pseudomonas aeruginosa in patients with cystic fibrosis: demonstration by radioimmunoassay.* Journal of Infectious Diseases, 1978. **138**(1): p. 49-58.

66. Arevalo‐Ferro, C., et al., *Identification of quorum‐sensing regulated proteins in the opportunistic pathogen Pseudomonas aeruginosa by proteomics.* Environmental Microbiology, 2003. **5**(12): p. 1350-1369.

67. Chandramouli, K. and P.-Y. Qian, *Proteomics: challenges, techniques and possibilities to overcome biological sample complexity.* Human genomics and proteomics: HGP, 2009. **2009**.

68. Hecker, M. and U. Völker, *General stress proteins in Bacillus subtilis.* FEMS Microbiology Letters, 1990. **74**(2-3): p. 197-213.

69. Maaβ, S., et al., *Highly precise quantification of protein molecules per cell during stress and starvation responses in Bacillus subtilis.* Molecular & Cellular Proteomics, 2014. **13**(9): p. 2260-2276.

70. Klumpp, S. and T. Hwa, *Bacterial growth: global effects on gene expression, growth feedback and proteome partition.* Current opinion in biotechnology, 2014. **28**: p. 96-102.

71. Meirelles, L.A. and D.K. Newman, *Both toxic and beneficial effects of pyocyanin contribute to the lifecycle of Pseudomonas aeruginosa.* Molecular microbiology, 2018. **110**(6): p. 995-1010.

72. Cahan, R., et al., *A secreted aminopeptidase of Pseudomonas aeruginosa: identification, primary structure, and relationship to other aminopeptidases.* Journal of Biological Chemistry, 2001. **276**(47): p. 43645-43652.

73. Upritchard, H.G., S.J. Cordwell, and I.L. Lamont, *Immunoproteomics to examine cystic fibrosis host interactions with extracellular Pseudomonas aeruginosa proteins.* Infection and immunity, 2008. **76**(10): p. 4624-4632.

74. Smalley, N.E., et al., *Evolution of the quorum sensing regulon in cooperating populations of Pseudomonas aeruginosa.* MBio, 2022. **13**(1): p. e00161-22.

75. Whiteley, M., K.M. Lee, and E.P. Greenberg, *Identification of genes controlled by quorum sensing in Pseudomonas aeruginosa.* Proceedings of the National Academy of Sciences, 1999. **96**(24): p. 13904-13909.

76. Fuqua, C. and E.P. Greenberg, *Self perception in bacteria: quorum sensing with acylated homoserine lactones.* Current opinion in microbiology, 1998. **1**(2): p. 183-189.

77. Schey, K.L., A.C. Grey, and J.J. Nicklay, *Mass spectrometry of membrane proteins: a focus on aquaporins.* Biochemistry, 2013. **52**(22): p. 3807-3817.

78. Nouwens, A.S., B.J. Walsh, and S.J. Cordwell, *Application of proteomics to Pseudomonas aeruginosa.* Proteomics of Microorganisms: Fundamental Aspects and Application, 2003: p. 117-140.

79. Ramos, J.-L., *Pseudomonas: volume 1 genomics, life style and molecular architecture*. 2011: Springer Science & Business Media.

80. Michéa-Hamzehpour, M., et al., *Two-dimensional polyacrylamide gel electrophoresis isolation and microsequencing of Pseudomonas aeruginosa proteins.* Enzyme and Protein, 1993. **47**(1): p. 1-8.

81. Cowell, B.A., et al., *Effect of nutrient limitation on adhesion characteristics of Pseudomonas aeruginosa.* Journal of applied microbiology, 1999. **86**(6): p. 944-954.

82. Lee, N.-G., et al., *Immunization of burn-patients with a Pseudomonas aeruginosa outer membrane protein vaccine elicits antibodies with protective efficacy.* Vaccine, 2000. **18**(18): p. 1952-1961.

83. von Specht, B.-U., et al., *Outer membrane proteins of Pseudomonas aeruginosa as vaccine candidates.* Journal of biotechnology, 1996. **44**(1-3): p. 145-153.

84. Meyer, J.-M., et al., *Pyoverdin is essential for virulence of Pseudomonas aeruginosa.* Infection and immunity, 1996. **64**(2): p. 518-523.

85. Jimenez, P.N., et al., *Role of PvdQ in Pseudomonas aeruginosa virulence under iron-limiting conditions.* Microbiology, 2010. **156**(1): p. 49-59.

86. Hoch, J.A., *Two-component and phosphorelay signal transduction.* Current opinion in microbiology, 2000. **3**(2): p. 165-170.

87. Ouidir, T., et al., *Potential of liquid-isoelectric-focusing protein fractionation to improve phosphoprotein characterization of Pseudomonas aeruginosa PA14.* Analytical and bioanalytical chemistry, 2014. **406**: p. 6297-6309.

88. Hitchen, P.G. and A. Dell, *Bacterial glycoproteomics.* Microbiology, 2006. **152**(6): p. 1575-1580.

89. Schirm, M., et al., *Structural and genetic characterization of glycosylation of type a flagellin in Pseudomonas aeruginosa.* Journal of bacteriology, 2004. **186**(9): p. 2523-2531.

90. Verma, A., et al., *Glycosylation of b-Type flagellin of Pseudomonas aeruginosa: structural and genetic basis.* Journal of bacteriology, 2006. **188**(12): p. 4395-4403.

91. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, *Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.* Lancet, 2018. **392**(10159): p. 1789-1858.

92. Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS), *National Health and Nutrition Examination Survey Questionnaire 2017-2018.* U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Hyattsville, MD, 2018.

93. Stacy, A., et al., *Bacterial fight-and-flight responses enhance virulence in a polymicrobial infection.* Proc Natl Acad Sci U S A, 2014. **111**(21): p. 7819-7824.

94. Kuboniwa, M., et al., *Metabolic crosstalk regulates Porphyromonas gingivalis colonization and virulence during oral polymicrobial infection.* Nat Microbiol, 2017. **2**(11): p. 1493-1499.

95. Chipashvili, O., et al., *Episymbiotic Saccharibacteria suppresses gingival inflammation and bone loss in mice through host bacterial modulation.* Cell Host Microbe, 2021. **29**(11): p. 1649-1662.

96. Kim, D., et al., *Spatial mapping of polymicrobial communities reveals a precise biogeography associated with human dental caries.* Proc Natl Acad Sci U S A, 2020. **117**(22): p. 12375-12386.

97. Tal, M., *Periodontal disease and oral hygiene. Described by Antoni van Leeuwenhoek.* J Periodontol, 1980. **51**(11): p. 668-669.

98. Gest, H., *The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society.* Notes Rec R Soc Lond, 2004. **58**(2): p. 187-201.

99. Gibbons, R.J., *Bacteriology of dental caries.* J Dent Res, 1964. **43**(6): p. 1021-1028.

100. Miller, W.D., *The micro-organisms of the human mouth: the local and general diseases which are caused by them*. 1890, Philadelphia: The S. S. White Dental Mfg. Co.

101. Clarke, J.K., *On the bacterial factor in the aetiology of dental caries.* Br J Exp Pathol, 1924. **5**(3): p. 141-147.

102. Socransky, S.S., *Microbiology of periodontal disease -- present status and future considerations.* J Periodontol, 1977. **48**(9): p. 497-504.

103. Gibbons, R.J. and J.V. Houte, *Bacterial adherence in oral microbial ecology.* Annu Rev Microbiol, 1975. **29**(1): p. 19-42.

104. Gibbons, R.J. and M. Nygaard, *Interbacterial aggregation of plaque bacteria.* Arch Oral Biol, 1970. **15**(12): p. 1397-IN39.

105. Stevens, M.R.E., et al., *Coaggregation occurs between microorganisms isolated from different environments.* FEMS Microbiol Ecol, 2015. **91**(11).

106. Costerton, J.W., G.G. Geesey, and K.J. Cheng, *How bacteria stick.* Sci Am, 1978. **238**(1): p. 86-95.

107. Kolenbrander, P.E., et al., *Oral multispecies biofilm development and the key role of cell-cell distance.* Nat Rev Microbiol, 2010. **8**(7): p. 471-480.

108. Kolenbrander, P.E., et al., *Communication among oral bacteria.* Microbiol Mol Biol Rev, 2002. **66**(3): p. 486-505.

109. Kolenbrander, P.E. and J. London, *Adhere today, here tomorrow: oral bacterial adherence.* J Bacteriol, 1993. **175**(11): p. 3247-3252.

110. Kolenbrander, P.E., et al., *Bacterial interactions and successions during plaque development.* Periodontol 2000, 2006. **42**(1): p. 47-79.

111. Kolenbrander, P.E., *Intergeneric coaggregation among human oral bacteria and ecology of dental plaque.* Annu Rev Microbiol, 1988. **42**(1): p. 627-656.

112. Mark Welch, J.L., F.E. Dewhirst, and G.G. Borisy, *Biogeography of the oral microbiome: the site-specialist hypothesis.* Annu Rev Microbiol, 2019. **73**: p. 335-358.

113. Henrici, A.T., *Studies of freshwater bacteria: I. A direct microscopic technique.* J Bacteriol, 1933. **25**(3): p. 277-287.

114. Zobell, C.E. and E.C. Allen, *The significance of marine bacteria in the fouling of submerged surfaces.* J Bacteriol, 1935. **29**(3): p. 239-251.

115. Socransky, S.S., et al., *Microbial complexes in subgingival plaque.* J Clin Periodontol, 1998. **25**(2): p. 134-144.

116. Armitage, G.C., *Learned and unlearned concepts in periodontal diagnostics: a 50-year perspective.* Periodontol 2000, 2013. **62**(1): p. 20-36.

117. Haffajee, A.D., et al., *Microbial complexes in supragingival plaque.* Oral Microbiol Immunol, 2008. **23**(3): p. 196-205.

118. Hajishengallis, G. and R.J. Lamont, *Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology.* Mol Oral Microbiol, 2012. **27**(6): p. 409-419.

119. Cross, K.L., et al., *Targeted isolation and cultivation of uncultivated bacteria by reverse genomics.* Nat Biotechnol, 2019. **37**(11): p. 1314-1321.

120. Escapa, I.F., et al., *New insights into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for the microbiome of the human aerodigestive tract.* mSystems, 2018. **3**(6).

121. Thompson, H., et al., *In vitro culture of previously uncultured oral bacterial phylotypes.* Appl Environ Microbiol, 2015. **81**(24): p. 8307-14.

122. Rheims, H., F.A. Rainey, and E. Stackebrandt, *A molecular approach to search for diversity among bacteria in the environment.* J Ind Microbiol, 1996. **17**(3): p. 159-169.

123. Soro, V., et al., *Axenic culture of a candidate division TM7 bacterium from the human oral cavity and biofilm interactions with other oral bacteria.* Appl Environ Microbiol, 2014. **80**(20): p. 6480-6489.

124. He, X., et al., *Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle.* Proc Natl Acad Sci U S A, 2015. **112**(1): p. 244-249.

125. Solbiati, J. and J. Frias-Lopez, *Metatranscriptome of the oral microbiome in health and disease.* J Dent Res, 2018. **97**(5): p. 492-500.

126. Relman, D.A., *The human microbiome: ecosystem resilience and health.* Nutr Rev, 2012. **70**(Suppl 1): p. S2-S9.

127. Yost, S., et al., *Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis.* Genome Med, 2015. **7**(1): p. 1-19.

128. Lewin, G.R., et al., *A quantitative framework reveals traditional laboratory growth is a highly accurate model of human oral infection.* Proc Natl Acad Sci U S A, 2022. **119**(2).

129. Jorth, P., et al., *Metatranscriptomics of the human oral microbiome during health and disease.* mBio, 2014. **5**(2): p. e01012-14.

130. Nowicki, E.M., et al., *Microbiota and metatranscriptome changes accompanying the onset of gingivitis.* mBio, 2018. **9**(2): p. e00575-18.

131. Amann, R.I., L. Krumholz, and D.A. Stahl, *Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology.* J Bacteriol, 1990. **172**(2): p. 762-770.

132. Valm, A.M., et al., *Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging.* Proc Natl Acad Sci U S A, 2011. **108**(10): p. 4152-4157.

133. Mark Welch, J.L., et al., *Biogeography of a human oral microbiome at the micron scale.* Proc Natl Acad Sci U S A, 2016. **113**(6): p. E791-E800.

134. Wilbert, S.A., J.L. Mark Welch, and G.G. Borisy, *Spatial ecology of the human tongue dorsum microbiome.* Cell Rep, 2020. **30**(12): p. 4003-4015.

135. Azimi, S., G.R. Lewin, and M. Whiteley, *The biogeography of infection revisited.* Nat Rev Microbiol, 2022.

136. Shi, H., et al., *Highly multiplexed spatial mapping of microbial communities.* Nature, 2020. **588**(7839): p. 676-681.

137. Jorth, P., et al., *Quantitative visualization of gene expression in mucoid and nonmucoid Pseudomonas aeruginosa aggregates reveals localized peak expression of alginate in the hypoxic zone.* mBio, 2019. **10**(6): p. e02622-19.

138. Dar, D., et al., *Spatial transcriptomics of planktonic and sessile bacterial populations at single-cell resolution.* Science, 2021. **373**(6556): p. eabi4882.

139. Cornforth, D.M., et al., *Quantitative framework for model evaluation in microbiology research using Pseudomonas aeruginosa and cystic fibrosis infection as a test case.* mBio, 2020. **11**(1): p. e03042-19.

140. Shapiro, S., E. Giertsen, and B. Guggenheim, *An in vitro oral biofilm model for comparing the efficacy of antimicrobial mouthrinses.* Caries Res, 2002. **36**(2): p. 93-100.

141. Carrouel, F., et al., *Quantitative molecular detection of 19 major pathogens in the interdental biofilm of periodontally healthy young adults.* Front Microbiol, 2016. **7**: p. 840.

142. Socransky, S.S., *Dental biofilms: difficult therapeutic targets.* Periodontol 2000, 2002. **28**: p. 12-55.

143. Socransky, S.S. and A.D. Haffajee, *Periodontal microbial ecology.* Periodontol 2000, 2005. **38**(1): p. 135-187.

144. Palmer Jr, R.J., et al., *Retrieval of biofilms from the oral cavity.* Methods Enzymol, 2001. **337**: p. 393-403.

145. Foster, J.S., R.J. Palmer Jr, and P.E. Kolenbrander, *Human oral cavity as a model for the study of genome-genome interactions.* Biol. Bull., 2003. **204**: p. 200-204.

146. Cornforth, D.M., et al., *Pseudomonas aeruginosa transcriptome during human infection.* Proc Natl Acad Sci U S A, 2018. **115**(22): p. E5125-E5134.

147. Rossi, E., et al., *High-resolution in situ transcriptomics of Pseudomonas aeruginosa unveils genotype independent patho-phenotypes in cystic fibrosis lungs.* Nat Commun, 2018. **9**(1): p. 3459.

148. Subashchandrabose, S., et al., *Host-specific induction of Escherichia coli fitness genes during human urinary tract infection.* Proceedings of the National Academy of Sciences, 2014. **111**(51): p. 18327-18332.

149. Becker, K., et al., *Quantifying post-transcriptional regulation in the development of Drosophila melanogaster.* Nat Commun, 2018. **9**(1): p. 4970.

150. Caglar, M.U., et al., *The E. coli molecular phenotype under different growth conditions.* Sci Rep, 2017. **7**: p. 45303.

151. Chen, W.H., et al., *Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances.* Nucleic Acids Res, 2016. **44**(3): p. 1192-202.

152. Choi, Y.W., et al., *Analysis of growth phase-dependent proteome profiles reveals differential regulation of mRNA and protein in Helicobacter pylori.* Proteomics, 2008. **8**(13): p. 2665-75.

153. Corbin, R.W., et al., *Toward a protein profile of Escherichia coli: comparison to its transcription profile.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9232-7.

154. Edfors, F., et al., *Gene-specific correlation of RNA and protein levels in human cells and tissues.* Molecular systems biology, 2016. **12**(10): p. 883.

155. Gygi, S.P., et al., *Correlation between protein and mRNA abundance in yeast.* Molecular and cellular biology, 1999. **19**(3): p. 1720-1730.

156. Jayapal, K.P., et al., *Uncovering genes with divergent mRNA-protein dynamics in Streptomyces coelicolor.* PLoS One, 2008. **3**(5): p. e2097.

157. Jeacock, L., J. Faria, and D. Horn, *Codon usage bias controls mRNA and protein abundance in trypanosomatids.* Elife, 2018. **7**.

158. Lu, P., et al., *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.* Nat Biotechnol, 2007. **25**(1): p. 117-24.

159. Maier, T., et al., *Quantification of mRNA and protein and integration with protein turnover in a bacterium.* Mol Syst Biol, 2011. **7**: p. 511.

160. Marguerat, S., et al., *Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells.* Cell, 2012. **151**(3): p. 671-83.

161. Ponnala, L., et al., *Correlation of mRNA and protein abundance in the developing maize leaf.* Plant J, 2014. **78**(3): p. 424-40.

162. Riba, A., et al., *Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates.* Proc Natl Acad Sci U S A, 2019. **116**(30): p. 15023-15032.

163. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.

164. Johansson, H.J., et al., *Breast cancer quantitative proteome and proteogenomic landscape.* Nature communications, 2019. **10**(1): p. 1-14.

165. Nicolet, B.P. and M.C. Wolkers, *Limited but gene-class specific correlation of mRNA and protein expression in human CD8+ T cells.* BioRxiv, 2020.

166. Ly, T., et al., *A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells.* Elife, 2014. **3**: p. e01630.

167. Greenbaum, D., et al., *Comparing protein abundance and mRNA expression levels on a genomic scale.* Genome biology, 2003. **4**(9): p. 1-8.

168. Kwon, T., et al., *Protein-to-mRNA ratios are conserved between Pseudomonas aeruginosa strains.* Journal of proteome research, 2014. **13**(5): p. 2370-2380.

169. Erdmann, J., et al., *Environment-driven changes of mRNA and protein levels in Pseudomonas aeruginosa.* Environ Microbiol, 2018. **20**(11): p. 3952-3963.

170. Turner, K.H., et al., *Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum.* Proceedings of the National Academy of Sciences, 2015. **112**(13): p. 4110-4115.

171. Haft, D.H., et al., *TIGRFAMs and genome properties in 2013.* Nucleic acids research, 2012. **41**(D1): p. D387-D395.

172. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-15.

173. Murray, J.L., et al., *Intrinsic Antimicrobial Resistance Determinants in the Superbug Pseudomonas aeruginosa.* mBio, 2015. **6**(6): p. e01603-15.

174. Belsley, D.A., E. Kuh, and R.E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*. 2005: John Wiley & Sons.

175. Saito, K., R. Green, and A.R. Buskirk, *Translational initiation in E. coli occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing.* Elife, 2020. **9**.

176. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms for molecular biology, 2011. **6**(1): p. 1-14.

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

178. Scott, M., et al., *Interdependence of cell growth and gene expression: origins and consequences.* Science, 2010. **330**(6007): p. 1099-1102.

179. Scott, M., et al., *Emergence of robust growth laws from optimal regulation of ribosome synthesis.* Mol Syst Biol, 2014. **10**: p. 747.

180. Ahrens, C.H., et al., *A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry.* J Bacteriol, 2022. **204**(1): p. e0035321.

181. Bartel, J., et al., *Optimized Proteomics Workflow for the Detection of Small Proteins.* J Proteome Res, 2020. **19**(10): p. 4004-4018.

182. Dincer, A.B., et al., *Reducing Peptide Sequence Bias in Quantitative Mass Spectrometry Data with Machine Learning.* J Proteome Res, 2022. **21**(7): p. 1771-1782.

183. Mallick, P., et al., *Computational prediction of proteotypic peptides for quantitative proteomics.* Nat Biotechnol, 2007. **25**(1): p. 125-31.

184. Chen, H., et al., *New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform.* Briefings in bioinformatics, 2020. **21**(4): p. 1397-1410.

185. Deutschbauer, A.M., et al., *Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast.* Genetics, 2005. **169**(4): p. 1915-1925.

186. Ibberson, C.B. and M. Whiteley, *The Staphylococcus aureus Transcriptome during Cystic Fibrosis Lung Infection.* mBio, 2019. **10**(6).

187. Rahme, L.G., et al., *Common virulence factors for bacterial pathogenicity in plants and animals.* Science, 1995. **268**(5219): p. 1899-902.

188. Schroth MN, C.J., Green SK, Kominos SD. , *Epidemiology of Pseudomonas aeruginosa in agricultural areas*, in *Pseudomonas Aeruginosa: Ecological Aspects and Patient Colonization* Y. VM, Editor. 1977, Raven Press: New York.

189. Whiteley, M., et al., *Gene expression in Pseudomonas aeruginosa biofilms.* Nature, 2001. **413**(6858): p. 860-4.

190. JC, O., *Venny. An interactive tool for comparing lists with Venn's diagrams [online]*. 2015: Available at https://bioinfogp.cnb.csic.es/tools/venny/.

191. Turner, K.H., et al., *Requirements for Pseudomonas aeruginosa acute burn and chronic surgical wound infection.* PLoS Genet, 2014. **10**(7): p. e1004518.

192. Duran-Pinedo, A.E., et al., *Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis.* ISME J, 2014. **8**(8): p. 1659-72.

193. Zhao, X., et al., *Phenotypic, genomic, and transcriptomic changes in an Acinetobacter baumannii strain after spaceflight in China's Tiangong-2 space laboratory.* Braz J Microbiol, 2022. **53**(3): p. 1447-1464.

194. Ott, E., et al., *Molecular repertoire of Deinococcus radiodurans after 1 year of exposure outside the International Space Station within the Tanpopo mission.* Microbiome, 2020. **8**(1): p. 150.

195. Crabbe, A., et al., *Transcriptional and proteomic responses of Pseudomonas aeruginosa PAO1 to spaceflight conditions involve Hfq regulation and reveal a role for oxygen.* Appl Environ Microbiol, 2011. **77**(4): p. 1221-30.

196. Liang, X., C.J. Martyniuk, and D.B.D. Simmons, *Are we forgetting the “proteomics” in multi-omics ecotoxicology?* Comparative Biochemistry and Physiology Part D: Genomics and Proteomics, 2020. **36**: p. 100751.

197. Zhang, Y., et al., *Protein analysis by shotgun/bottom-up proteomics.* Chemical reviews, 2013. **113**(4): p. 2343-2394.

198. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present.* Analytical and bioanalytical chemistry, 2012. **404**: p. 939-965.

199. Liessi, N., et al., *Proteomics and metabolomics for cystic fibrosis research.* International Journal of Molecular Sciences, 2020. **21**(15): p. 5439.

200. Pineau, C. *Looking for missing proteins in the testicular germ cell lineage: new insights into normal and pathological spermatogenesis*.

201. Ohayon, S., et al., *Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification.* PLoS Comput Biol, 2019. **15**(5): p. e1007067.

202. Palmblad, M., *Theoretical Considerations for Next-Generation Proteomics.* J Proteome Res, 2021. **20**(6): p. 3395-3399.

203. Swaminathan, J., et al., *Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures.* Nat Biotechnol, 2018.

204. Zhang, M., et al., *Impact of Growth Rate on the Protein-mRNA Ratio in Pseudomonas aeruginosa.* mBio, 2022. **14**(1): p. e03067-22.

205. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-587.

206. Franks, A., E. Airoldi, and N. Slavov, *Post-transcriptional regulation across human tissues.* PLoS computational biology, 2017. **13**(5): p. e1005535.

207. Edfors, F., et al., *Gene-specific correlation of RNA and protein levels in human cells and tissues.* Mol Syst Biol, 2016. **12**(10): p. 883.

208. Poulsen, B.E., et al., *Defining the core essential genome of Pseudomonas aeruginosa.* Proceedings of the National Academy of Sciences, 2019. **116**(20): p. 10072-10080.

209. Luo, H., et al., *DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools.* Nucleic Acids Research, 2021. **49**(D1): p. D677-D686.

210. Zhang, R., H.Y. Ou, and C.T. Zhang, *DEG: a database of essential genes.* Nucleic Acids Research, 2004. **32**(suppl_1): p. D271-D272.

211. Ji, Y., et al., *Identification of Critical Staphylococcal Genes Using Conditional Phenotypes Generated by Antisense RNA.* Science, 2001. **293**(5538): p. 2266-2269.

212. Chaudhuri, R.R., et al., *Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH).* BMC Genomics, 2009. **10**(1): p. 291.

213. Laurent, J.M., et al., *Protein abundances are more conserved than mRNA abundances across diverse taxa.* PROTEOMICS, 2010. **10**(23): p. 4209-4212.

214. Schrimpf, S.P., et al., *Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes.* PLOS Biology, 2009. **7**(3): p. e1000048.

215. Vogel, C. and E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.* Nature Reviews Genetics, 2012. **13**(4): p. 227-232.

216. Zhang, C., et al., *Augmenting the genetic toolbox for Sulfolobus islandicus with a stringent positive selectable marker for agmatine prototrophy.* Applied and environmental microbiology, 2013. **79**(18): p. 5539-5549.

217. Cornforth Daniel, M., et al., *Quantitative Framework for Model Evaluation in Microbiology Research Using Pseudomonas aeruginosa and Cystic Fibrosis Infection as a Test Case.* mBio, 2020. **11**(1): p. 10.1128/mbio.03042-19.

218. Lewin, G.R., et al., *A quantitative framework reveals traditional laboratory growth is a highly accurate model of human oral infection.* Proceedings of the National Academy of Sciences, 2022. **119**(2): p. e2116637119.

219. Higginbotham, L., et al., *Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic Alzheimer's disease.* Science Advances. **6**(43): p. eaaz9360.

220. Zlatic, S.A., et al., *Convergent cerebrospinal fluid proteomes and metabolic ontologies in humans and animal models of Rett syndrome.* iScience, 2022. **25**(9): p. 104966.

221. Emms, D.M. and S. Kelly, *OrthoFinder: phylogenetic orthology inference for comparative genomics.* Genome biology, 2019. **20**: p. 1-14.

222. Emms, D.M. and S. Kelly, *OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy.* Genome biology, 2015. **16**: p. 1-14.

223. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.

224. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature methods, 2015. **12**(1): p. 59-60.

225. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996. **58**(1): p. 267-288.

226. Airoldi, E.M., et al., *Predicting cellular growth from gene expression signatures.* PLoS computational biology, 2009. **5**(1): p. e1000257.

227. Miethke, M. and M.A. Marahiel, *Siderophore-based iron acquisition and pathogen control.* Microbiology and molecular biology reviews, 2007. **71**(3): p. 413-451.

228. Salvail, H., et al., *A small RNA promotes siderophore production through transcriptional and metabolic remodeling.* Proceedings of the National Academy of Sciences, 2010. **107**(34): p. 15223-15228.

229. Salvail, H. and E. Massé, *Regulating iron storage and metabolism with RNA: an overview of posttranscriptional controls of intracellular iron homeostasis.* Wiley Interdisciplinary Reviews: RNA, 2012. **3**(1): p. 26-36.

230. Wu, X., et al., *In vivo proteome of Pseudomonas aeruginosa in airways of cystic fibrosis patients.* Journal of proteome research, 2019. **18**(6): p. 2601-2612.

231. Elowitz, M.B. and S. Leibler, *A synthetic oscillatory network of transcriptional regulators.* Nature, 2000. **403**(6767): p. 335-338.

232. Selinger, D.W., et al., *Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation.* Genome research, 2003. **13**(2): p. 216-223.

233. Roberts, C., et al., *Characterizing the effect of the Staphylococcus aureus virulence factor regulator, SarA, on log-phase mRNA half-lives.* Journal of bacteriology, 2006. **188**(7): p. 2593-2603.

234. Andersson, A.F., et al., *Global analysis of mRNA stability in the archaeon Sulfolobus.* Genome biology, 2006. **7**: p. 1-10.

235. Miller, C., et al., *Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast.* Molecular systems biology, 2011. **7**(1): p. 458.

236. Thomsen, S., et al., *Genome-wide analysis of mRNA decay patterns during early Drosophila development.* Genome biology, 2010. **11**: p. 1-27.

237. Chen, C.Y.A., N. Ezzeddine, and A.B. Shyu, *Messenger RNA half-life measurements in mammalian cells.* Methods in enzymology, 2008. **448**: p. 335-357.

238. Rustad, T.R., et al., *Global analysis of mRNA stability in Mycobacterium tuberculosis.* Nucleic acids research, 2013. **41**(1): p. 509-517.

239. Hambraeus, G., C. von Wachenfeldt, and L. Hederstedt, *Genome-wide survey of mRNA half-lives in Bacillus subtilis identifies extremely stable mRNAs.* Molecular Genetics and Genomics, 2003. **269**: p. 706-714.

240. Ammann, T.W., R. Gmür, and T. Thurnheer, *Advancement of the 10-species subgingival Zurich Biofilm model by examining different nutritional conditions and defining the structure of the in vitro biofilms.* BMC Microbiol, 2012. **12**(1): p. 1-13.

241. Bao, K., et al., *Role of Porphyromonas gingivalis gingipains in multi-species biofilm formation.* BMC Microbiol, 2014. **14**(1): p. 1-8.