Linking Self Versus Non-Self Recognition Genes and Their Function to Microbial Community Structure

Linking self versus non-self recognition genes and their function to microbial community structure

A dissertation presented

by

Denise Sirias

to

The Department of Molecular and Cellular Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biochemistry

Harvard University

Cambridge, Massachusetts

May 2019

Dissertation Advisor: Professor Karine A. Gibbs By: Denise Sirias

Linking self versus non-self recognition genes and their function to microbial community structure

Abstract

Many bacterial species reside in dense communities, such as the gut and oral microbiomes, which retain a structural organization where genetically similar bacterial populations are physically separated. Key candidates for the organization of such structures are contact-dependent interactions, because these structures involve cell contact. These interactions can be competitive, as bacteria will be competing for the same resources. Bacteria use various secretion systems to deliver toxins or effector proteins into neighboring cells in a contact-dependent manner. Cells producing toxins have a competitive advantage by inhibiting growth of neighboring cells. The production of a strain-specific immunity protein by the recipient cell can counteract the effects of the toxin, making these interactions a form of self-recognition.

Using biochemical and genetic approaches, I elucidated the target of the *Proteus mirabilis* BB2000-derived toxin, IdrD, and its immunity protein, IdrE. The C-terminal domain of IdrD functions as a DNase; IdrE counteracts this activity and causes loss of toxin signal by an unknown mechanism. The same molecular function was observed in the homologous toxin-immunity pair in the bacterium *Rothia aeria* C6B, which is distantly related to *P. mirabilis*. The molecular characterization was combined with metagenomic analysis. We are able to probe for the abundance of these toxin-immunity pairs in different communities, and even detect subdomains within the toxin. Combining molecular characterization with metagenomic analysis provides a way to study toxin-immunity pairs in the context of a microbial community.

Table of Contents

Acknowledgments

"It is our choices, Harry, that show what we truly are, far more than our abilities"

–Albus Dumbledore in *Harry Potter and the Chamber of Secrets* by J.K. Rowling

I consider graduate school one of the many life-changing choices that shaped who I am today. The people in my life, both in and out of graduate school, have largely shaped my experiences.

First and foremost, I would like to thank my advisor Dr. Karine A. Gibbs. Her passion for science was apparent from the first time I met her (as an undergrad at the IDEAS). After every data meeting with her, I felt more excited and inspired to continue working on my project. She has been the best possible mentor I could have asked for.

Karine was able to develop an amazing lab culture. I want to thank members of the Gibbs lab, Dr. Murray Tipping, Dr. Christina Saak, Dr. Kristin Little, Dr. Martha Zepeda Rivera, Achala Chittor, and Jacob Austerman, for both helpful scientific discussion and for being a fun group of people to be around. In addition to being great colleagues, Christina and Martha have also become two of my best friends. Thank you for being the Hufflepuff and Slytherin to my Ravenclaw and for all the McDonald's runs/orders over the years. I also want to thank my mentees, Sajal Akkipeddi and Emma Keteku, for their enthusiasm in taking on a tough project. It has been an inspiration to watch you grow as scientists.

Thank you my thesis committee Dr. Cassandra Extavour, Dr. Rich Losick and Dr. Dan Kahne for providing valuable feedback throughout the years. And a special thank you to Dr. Rich Losick whose IDEAs program put me in a lab my freshman year of college. This experience was crucial to my development as a scientist and I am forever grateful for the opportunity.

Thank you to the Britton lab which not only gave me the skills I needed to get into graduate school, but also a great mentor (Dr. Rob Britton) and amazing group of friends (Kylie Farrell, Anthony Findley, Megha Gulati, Kar Mun Neoh).

To my MCO classmates, it's been an incredible journey full of problem sets, volleyball, and the best G3 skit. I hope we continue to be known as the last class to beat the faculty at the G1 vs. faculty volleyball match. In particular, thank you to Sisteine Club (Haneui Bae, Viktoria Betin, Brenda Marin-Rodriguez, Shristi Pandey, Stephanie Tsai, and Jenelle Wallace) for being an incredible support system and always being down for data club (and free food). Also to my MCO friends in other years.

Thank you to the Biolabs third floor beer hour crew for lively discussions every week. A big thank you to Dan Utter for collaborating with me on this project, and for teaching me about evolutionary biology.

Often times in graduate school, you forget that there is a life outside of a PhD. A big thank you goes to my friends who remind me of that:

- My roomies, Viktoria Betin, Jocelyn Fuentes and Stephanie Tsai, for being amazing friends and a much needed support system at home.
- Ally Bratzel, Julia Miller, and Steven Zhang, for our monthly Skype calls and your constant support both at MSU and in grad school.
- My Midland crew: Katie Frank, Libby Gorton, Lizzy Jacks, Annie Krueger, Rebecca Roper, Elizabeth Schaffert, Maggie Zemanek, and Sarah Zimmerman. We have come a long way since high school, and you guys have been there every step of the way.

To my extended family in Nicaragua, it's hard that we are so far away, but I love you with all my heart and could not have done this without your support. To my "adopted" extended

family, Uncle Mark and the Gortons, for all the support over the years and for inviting us to your homes for Thanksgiving/Christmas/Easter.

To my boyfriend, James Kremer, for being my partner-in-crime for the past two years. Thank you for always being in my corner, and pushing me to be spontaneous every once in awhile. I couldn't have survived the last two years without your love and support.

And last, but not least, thank you to my amazing family. Jessica, the moment you were born, I was given an automatic best friend for life. It has been amazing to be your big sister and I've learned so much from you. Thank you Mama and Daddy for being the kind of parents who would bring me warm milk when I was up late studying and call me in sick to school if I didn't get enough sleep. You are a constant inspiration to and I owe all my achievements to you.

As a side note, I also want to thank the following:

- My cat Mimi for being the sweetest pet ever.
- The Jonas Brothers, for getting back together, and reminding me of my high school days.
- J.K. Rowling for writing the Harry Potter series, and creating such an inspirational story with lessons that I carry with me always.

List of Figures

Chapter 2

Chapter 3

Appendix

List of Tables

I dedicate this thesis to my family, friends, and mentors who have helped me become the person I am today

Page intentionally left blank

Chapter 1

Polymorphic toxin systems and their role in bacterial interactions

Microbial communities, such as the oral microbiome, have been shown to retain structural organization [1-4]. Because such structures involve cell contact, key candidates for defining community structure are contact-dependent toxic proteins [5]. Polymorphic toxin systems are described family of multi-domain toxins that have been implicated in interbacterial competition in a contact-dependent manner [6-9]. Though their role in polymicrobial structures has not been elucidated, their presence in many Gram-negative and Gram-positive bacteria and toxin diversity makes them a potential candidate for defining community structure.

Polymorphic toxins are characterized by their modular structure. They contain a conserved N-terminal region involved in secretion or transport, followed by a divergent region in the C-terminal containing the toxic domains [6, 7, 10]. Toxic domains are often shared between distinct N-terminal domains indicating that multiple secretion mechanisms can deliver similar toxins [6, 7, 10]. This modularity suggests that novel effectors are acquired through horizontal gene transfer and recombination [6-9]. Though many toxin functions are still unknown, those identified include DNAses, RNases, pore-formers, deaminases, and peptidases [11]. PTS loci include an immunity gene that confers protection against the toxin [6-9]. Additionally, many loci also contain toxin-immunity pairs where the toxin is not attached to the N-terminal region required for secretion [8]. These pairs are termed orphan modules and are hypothesized to have been displaced after acquisition of a new toxin-immunity pair [8]. The mechanism of effector evolution after acquisition remains an open question in the field. In *Salmonella enterica* serovar typhirium, it has been shown that recombination can restore an orphan toxin to the N-terminal domain resulting in expression of the former orphan pair as a result of serial passaging [12]. Polymorphic toxin systems include the following well-characterized families: SUKH

superfamily, multiple adhesion family (Maf), colicins and S-type pyocins, contact dependent inhibition (CDI), and type VI secretion system-associated effectors [13].

SUKH Superfamily

Polymorphic toxin systems were defined by *in silico* analysis of the SUKH superfamily [7]. Members of the SUKH superfamily have a shared structural core, but often have low sequence similarity [7]. *In silico* analysis of the genomic neighborhood of genes containing SUKH domains showed that these proteins co-occurred with genes encoding different types of nucleases [7]. It was concluded that SUKH genes encoded immunity proteins to counteract the effect of their cognate toxins [7]. Similar analysis was done starting with genes containing nuclease domains to identify downstream immunity genes [11]. This analysis identified hundreds of novel toxins and immunity genes in the polymorphic toxin systems [11].

Multiple adhesion family (Maf)

The multiple adhesion family (*maf*) has been characterized in the *Neisseria* species [14, 15]. Interestingly, *maf* genomic islands (MGIs) are a found in pathogenic species, but absent in nonpathogenic species [14]. MGIs are in a conserved chromosomal location and contain *mafA*, encoding a predicted adhesion, *mafB*, encoding a polymorphic toxin, and *mafI*, encoding the corresponding immunity protein [16]. MafB proteins contain a conserved N-terminal domain of unknown function (DUF1020, PF06255), which is restricted to species in the *Neisseria* genus [14]. Additionally, MafB contains a signal peptide and has been detected, along with MafA, in outer membrane vesicles; however exact mechanism of secretion is still unknown [17].

Colicins and S-type pyocins

Colicins produced by *Escherichia coli* and soluble (S)-type pyocins are produced by the pseudomonads are well-studied bacteriocins. They are considered polymorphic toxin systems

because of their modular structure: an N-terminal translocation domain, a receptor binding domain, and a C-terminal toxic domain [18]. The colicin gene cluster also contains an immunity gene; often there are multiple immunity genes downstream of the colicin gene [18]. While colicins are encoded on plasmids, S-type pyocins are located on the chromosomes, similar to other polymorphic toxins [18, 19]. Group A colicins are encoded with a lysis gene which allows for release from the cells [18]. Group B colicins and S-type pyocins are not encoded with a lysis gene, and mechanism of release is unknown [18, 19]. An ecological role is suggested for colicins because they are produced by many *E. coli* strains isolated from the human gastrointestinal tract [20].

Contact-dependent growth inhibition

Contact-dependent inhibition (CDI) proteins were first discovered in *Escherichia coli* strain EC93 [21]. EC93 was found to be predominant in rat intestine and inhibit the growth of *E. coli* K-12 (laboratory strain) in a contact-dependent manner [21]. CDI proteins have now been found to be widespread in Gram-negative bacteria, and two major classes have been defined: *E. coli* and *Burkholderia*-type [6]. Both classes contain a polymorphic toxin (CdiA and BcpA, respectively), a two-partner secretion protein (CdiB and BcpB), and an immunity protein (CdiI and BcpI) [6]. The N- and C-terminal of the polymorphic toxin is delineated by a VENN (*E. coli* type) or Nx(E/Q)LYN (*Burkholderia*-type) motif [6, 22]. *Burkholderia*-type CDI systems also encode a small lipoprotein termed BcpO; however, its function remains unknown [22].

Role of CDI in biofilm architecture

Burkholderia thailandensis biofilm formation has been shown to require a CDI locus (*bcpAIOB*) [23]. Mutants lacking *bcpB* or the whole locus have less biomass and lack the pillar structure compared to wild-type biolfilms [23]. The exact molecular mechanism of how BcpA-

CTs modulates biofilm formation is unknown [23]. One hypothesis put forward is that BcpA-CTs, specifically those that target DNA, could have a role in modifying extracellular DNA (eDNA), which is an important component of the biofilm matrix [23]. Additionally, the *bcpAIOB* locus has been implicated in interbacterial competition within *B. thailandensis*; strains with different *bcpA-bcpI* pairs segregate within a biofilm [24].

Type VI secretion system-associated toxins/effectors

Rearrangement hotspot (RHS) family of toxins

Rhs proteins are found broadly in Gram-negative and Gram-positive bacteria, and are associated with bacterial competition [8, 9]. They are large filamentous proteins characterized by YD repeats in the N-terminal end and highly variable in the C-termini containing toxic domains [8-10, 25]. In Gram-negative bacteria, the toxins are secreted through T6SS; in Gram-positive bacteria, WapA, one of the identified RHS repeat proteins in *B. subtilis*, contains signal peptides for the general secretory pathway [9].

The Gibbs lab reported an *rhs*-containing locus, *idr*, and a functional type VI secretion system are necessary for competitions between *P. mirabilis* BB2000 and foreign *P. mirabilis* strains on a surface [26]. In surface competitions between two different clinical isolates of *P. mirabilis*, BB2000 and HI4320, BB2000 dominated at the leading edge [26]. Disruption in the *idr* or T6S components of BB2000 resulted HI4320 dominating the swarm; therefore, BB2000 lost its competitive advantage [26]. One gene in the *idr* locus, *idrD*, contains *rhs* sequences and contains the structure of a polymorphic toxin [26]. In Chapter 2, I identify the target of the encoded protein of *idrD*. In Chapter 3, I identify the downstream gene of *idrE* as the cognate immunity protein.

Additional type VI effectors (toxins)

Valine-glycine repeat protein G (VgrG) and hemolysin co-regulated protein (Hcp) are structural components of the T6SS that typically interact with T6S substrates [27-29]. Homologs of these proteins have been found fused to C-terminal toxic domains with activities that can target bacterial or eukaryotic cells, forming the modular structure that characterizes polymorphic toxins [30-33]. Additional T6S effectors have been classified by their activity: Tae (type VI secretion amidase effector), Tde (type VI secretion DNase effector), and Tle (type VI secretion lipase effector)[34-36]. Most of these effectors are not polymorphic, but some Tde effectors exhibit a modular structure [35].

Metagenomic analysis of T6S effectors

Metagenomic analysis has been used to study the potential role of T6SS in microbial communities [5, 37, 38]. For example, metagenomic analysis of *tle* (type VI secretion lipase effector) genes showed that they were more abundant in host associated niches, including human, arthropod, and rhizosphere metagenomes, than non-host associated niches, including bulk soil and aquatic environments [37]. Further, some niches showed variation in the frequency of different Tle families [37]. Combined, this data suggests that effectors may play a niche-specific role, defined by the selection pressures in a specific environment [37]. The continual molecular characterization of new T6S effectors and other polymorphic toxins provides an opportunity to probe the large amount of metagenomic data to understand their role within a community. In Chapter 2 and 3, I combine the molecular characterization and metagenomic analysis to investigate the abundance and specificity of the T6S, Rhs-associated toxin IdrD, its cognate immunity protein IdrE, and their homologs in other bacterial species.

Functions of the PD-(D/E)XK superfamily

Recently, effectors of the contact-dependent inhibition (CDI) system in *Enterobacter cloacae* ATCC 13047 and two tRNases in *Burkholderia pseudomallei* (isolates E479 and 1026b), called CdiA, were found to contain a fold similar to the PD-(D/E)XK superfamily and degrade RNA [39]. The type II restriction endonucleases are the most well-studied group of PD-(D/E)XK phosphodiesterases and were previously thought to be the only members of this family [40]. Studies of different type II restriction endonucleases identified the common core of a four-stranded mixed β-sheet flanked by two α-helices [41, 42]. Though the core is conserved in PD-(D/E)XK enzymes, little sequence homology was observed [40]. Diversity in structure is observed in subdomains that are important for substrate binding or dimerization of the enzymes [40, 43]. Two of the β-strands on within this core contain the amino acid residues required for catalysis of DNA cleavage: two carboxylates (one aspartate and one glutamate or aspartate) and one lysine residue [41, 42]. The two carboxylates are predicted to bind Mg^{2+}, a cofactor known to be essential for activity of many of these enzymes [41]. The exact catalytic mechanism of the cleavage of the phosphodiester bond by these enzymes is still unknown.

The great sequence divergence of PD-(D/E)XK enzymes has made it difficult to identify new superfamily members. In addition to DNA restriction, recently added members of this superfamily are involved in DNA recombination, transposon excision, Holliday junction resolving, DNA repair, Pol II termination, DNA binding, tRNA splicing, and bacterial competition (CDI) [40]. These various functions all involve nucleic acids, but specificity and type of nucleic acid targeted differ among members of this superfamily [40]. Understanding how the common structural core can recognize different nucleic acid targets and functionally characterizing new members of the PD-(D/E)XK superfamily are currently open questions in the field. In Chapter 2, I report that the C-terminal domain of IdrD contains a PD-(D/E)XK domain

that is required for the observed DNase activity. Characterization of homologs in *Rothia aeria* showed us that in addition to the PD-(D/E)XK active site, the C-terminal region is also necessary for DNase activity. This work identifies a new subfamily of the PD-(D/E)XK superfamily.

This thesis addresses the role of polymorphic toxin systems in communities using the toxin-immunity pair IdrD and IdrE from the bacterium *P. mirabilis* BB2000. Molecular characterization of IdrD and IdrE allows for identification of homologs found across the bacterial tree. Further, we combine molecular and metagenomic analyses to observe abundance of IdrD and IdrE in various microbial communities.

References

1. Lloyd-Price, J., et al., *Strains, functions and dynamics in the expanded Human Microbiome Project.* Nature, 2017. **550**(7674): p. 61-66.

2. Costea, P.I., et al., *Subspecies in the global human gut microbiome.* Mol Syst Biol, 2017. **13**(12): p. 960.

3. Eren, A.M., et al., *Oligotyping analysis of the human oral microbiome.* Proc Natl Acad Sci U S A, 2014. **111**(28): p. E2875-84.

4. Aas, J.A., et al., *Defining the normal bacterial flora of the oral cavity.* J Clin Microbiol, 2005. **43**(11): p. 5721-32.

5. Verster, A.J., et al., *The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition.* Cell Host Microbe, 2017. **22**(3): p. 411-419.e4.

6. Aoki, S.K., et al., *A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria.* Nature, 2010. **468**(7322): p. 439-42.

7. Zhang, D., L.M. Iyer, and L. Aravind, *A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems.* Nucleic Acids Res, 2011. **39**(11): p. 4532-52.

8. Poole, S.J., et al., *Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems.* PLoS Genet, 2011. **7**(8): p. e1002217.

9. Koskiniemi, S., et al., *Rhs proteins from diverse bacteria mediate intercellular competition.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 7032-7.

10. Jackson, A.P., et al., *Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement.* BMC Genomics, 2009. **10**: p. 584.

11. Zhang, D., et al., *Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics.* Biol Direct, 2012. **7**: p. 18.

12. Koskiniemi, S., et al., *Selection of orphan Rhs toxin expression in evolved Salmonella enterica serovar Typhimurium.* PLoS Genet, 2014. **10**(3): p. e1004255.

13. Jamet, A. and X. Nassif, *New players in the toxin field: polymorphic toxin systems in bacteria.* MBio, 2015. **6**(3): p. e00285-15.

14. Jamet, A., et al., *A new family of secreted toxins in pathogenic Neisseria species.* PLoS Pathog, 2015. **11**(1): p. e1004592.

15. Paruchuri, D.K., et al., *Identification and characterization of a Neisseria gonorrhoeae gene encoding a glycolipid-binding adhesin.* Proc Natl Acad Sci U S A, 1990. **87**(1): p. 333-7.

16. Parkhill, J., et al., *Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491.* Nature, 2000. **404**(6777): p. 502-6.

17. Zielke, R.A., et al., *Quantitative proteomics of the Neisseria gonorrhoeae cell envelope and membrane vesicles for the discovery of potential therapeutic targets.* Mol Cell Proteomics, 2014. **13**(5): p. 1299-317.

18. Cascales, E., et al., *Colicin biology.* Microbiol Mol Biol Rev, 2007. **71**(1): p. 158-229.

19. Ghequire, M.G. and R. De Mot, *Ribosomally encoded antibacterial proteins and peptides from Pseudomonas.* FEMS Microbiol Rev, 2014. **38**(4): p. 523-68.

20. Smajs, D., et al., *Bacteriocin synthesis in uropathogenic and commensal Escherichia coli: colicin E1 is a potential virulence factor.* BMC Microbiol, 2010. **10**: p. 288.

21. Aoki, S.K., et al., *Contact-dependent inhibition of growth in Escherichia coli.* Science, 2005. **309**(5738): p. 1245-8.

22. Anderson, M.S., E.C. Garcia, and P.A. Cotter, *The Burkholderia bcpAIOB genes define unique classes of two-partner secretion and contact dependent growth inhibition systems.* PLoS Genet, 2012. **8**(8): p. e1002877.

23. Garcia, E.C., et al., *Burkholderia BcpA mediates biofilm formation independently of interbacterial contact-dependent growth inhibition.* Mol Microbiol, 2013. **89**(6): p. 1213-25.

24. Anderson, M.S., E.C. Garcia, and P.A. Cotter, *Kind discrimination and competitive exclusion mediated by contact-dependent growth inhibition systems shape biofilm community structure.* PLoS Pathog, 2014. **10**(4): p. e1004076.

25. Ma, J., et al., *PAAR-Rhs proteins harbor various C-terminal toxins to diversify the antibacterial pathways of type VI secretion systems.* Environ Microbiol, 2017. **19**(1): p. 345-360.

26. Wenren, L.M., et al., *Two independent pathways for self-recognition in Proteus mirabilis are linked by type VI-dependent export.* MBio, 2013. **4**(4).

27. Mougous, J.D., et al., *A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus.* Science, 2006. **312**(5779): p. 1526-30.

28. Leiman, P.G., et al., *Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin.* Proc Natl Acad Sci U S A, 2009. **106**(11): p. 4154-9.

29. Silverman, J.M., et al., *Haemolysin coregulated protein is an exported receptor and chaperone of type VI secretion substrates.* Mol Cell, 2013. **51**(5): p. 584-93.

30. Blondel, C.J., et al., *Comparative genomic analysis uncovers 3 novel loci encoding type six secretion systems differentially distributed in Salmonella serotypes.* BMC Genomics, 2009. **10**: p. 354.

31. Brooks, T.M., et al., *Lytic activity of the Vibrio cholerae type VI secretion toxin VgrG-3 is inhibited by the antitoxin TsaB.* J Biol Chem, 2013. **288**(11): p. 7618-25.

32. Dong, T.G., et al., *Identification of T6SS-dependent effector and immunity proteins by Tn-seq in Vibrio cholerae.* Proc Natl Acad Sci U S A, 2013. **110**(7): p. 2623-8.

33. Pukatzki, S., et al., *Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin.* Proc Natl Acad Sci U S A, 2007. **104**(39): p. 15508-13.

34. Russell, A.B., et al., *A widespread bacterial type VI secretion effector superfamily identified using a heuristic approach.* Cell Host Microbe, 2012. **11**(5): p. 538-49.

35. Ma, L.S., et al., *Agrobacterium tumefaciens deploys a superfamily of type VI secretion DNase effectors as weapons for interbacterial competition in planta.* Cell Host Microbe, 2014. **16**(1): p. 94-104.

36. Russell, A.B., et al., *Diverse type VI secretion phospholipases are functionally plastic antibacterial effectors.* Nature, 2013. **496**(7446): p. 508-12.

37. Egan, F., F.J. Reen, and F. O'Gara, *Tle distribution and diversity in metagenomic datasets reveal niche specialization.* Environ Microbiol Rep, 2015. **7**(2): p. 194-203.

38. Bulgarelli, D., et al., *Structure and function of the bacterial root microbiota in wild and domesticated barley.* Cell Host Microbe, 2015. **17**(3): p. 392-403.

39. Johnson, P.M., et al., *Functional Diversity of Cytotoxic tRNase/Immunity Protein Complexes from Burkholderia pseudomallei.* J Biol Chem, 2016. **291**(37): p. 19387-400.

40. Steczkiewicz, K., et al., *Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily.* Nucleic Acids Res, 2012. **40**(15): p. 7016-45.

41. Selent, U., et al., *A site-directed mutagenesis study to identify amino acid residues involved in the catalytic function of the restriction endonuclease EcoRV.* Biochemistry, 1992. **31**(20): p. 4808-15.

42. Venclovas, C., A. Timinskas, and V. Siksnys, *Five-stranded beta-sheet sandwiched with two alpha-helices: a structural link between restriction endonucleases EcoRI and EcoRV.* Proteins, 1994. **20**(3): p. 279-82.

43. Kosinski, J., M. Feder, and J.M. Bujnicki, *The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function.* BMC Bioinformatics, 2005. **6**: p. 172.

Chapter 2

Identifying the functional role of IdrD, a novel nuclease, in microbial communities

All of the work presented in this chapter adapted from a manuscript in progress written by Denise Sirias, Dan Utter, and Dr. Karine A. Gibbs. Denise Sirias performed biochemical, genetic and physiological characterization, Dan Utter performed metagenomic analysis, and Dr. Karine A. Gibbs advised.

Abstract

Within complex microbial communities such as the gut and oral microbiomes, bacteria can establish physically separated colonies. This is partially achieved by killing competing bacteria. Polymorphic toxins, which are widely described within individual genera, are hypothesized to be a critical factor for such competitions. Here, we characterize a novel DNA-degrading polymorphic toxin and identify similar proteins across distantly-related bacteria. By searching human metagenomes for the toxin sequences, we found that though in low abundance, specific toxins are restricted to specific populations and habitats, bolstering methodology, combining molecular function and metagenomics analyses, serves as a model for other analyses to probe the abundance and role of sub-protein domains in complicated environments.

Introduction

Complex microbial communities have historically been defined by the identity of the bacteria within. Recent reports have shown that some communities, such as those within the oral microbiome, retain structural organization in which genetically similar bacterial populations are physically separated and inhabit distinct sub-regions [1-4]. However, identifying the processes by which populations maintain niche specificity and spatial separation remains an ongoing area of research. Potential key candidates for defining community structure are contact-dependent toxic proteins [5]. Bacteria can deploy various contact-dependent transport systems to deliver toxins or effector proteins into neighboring cells, often killing the recipient cell or causing long-term growth inhibition [6, 7]. Genetically identical siblings resist death due to the production of a protein to inhibit the delivered toxin [6-8]. Our ability to resolve the prevalence of such toxic proteins in microbiomes remains limited, partially due to their low abundance.

We have combined molecular and metagenomic analyses of a single toxin to address the role of toxins in community structure. In the bacterium *Proteus mirabilis*, which is a low-abundance member of the mammalian gut, one strain is able to physically and spatially exclude another [9]. Many molecular details governing this population separation have been described [10-18]. One critical factor is a previously undescribed Rhs toxin, encoded by the gene *idrD* [12]. Polymorphic toxin systems, such as the Rhs protein family, are widespread among Gram-negative and Gram-positive bacteria and are characterized by a modular organization [6-8, 19]. These toxins are often exported from the bacterial cell via a diverse set of mechanisms, such as Types IV, V, and VI secretion systems [20-22]. The majority of the protein, near the N-terminus in general, is required for secretion and possible horizontal gene transfer, while the C-terminal region, often ~ 140 amino acids in length, harbors toxic activity. These C-terminal domains are found associated with distinct N-terminal regions, suggesting that multiple secretion mechanisms can deliver apparently homologous toxins [7, 8, 19]. Several Rhs toxins have been identified as RNAses, DNases, pore-formers, deaminases, and peptidases [23]. However, the function of the *idrD*-encoded toxin was unknown.

Results

IdrD is a DNase part of a previously uncharacterized subfamily in the PD-(D/E)XK family

Based on amino acid similarity, we identified transport and Rhs subdomains in the predicted IdrD polypeptide, and from that, predicted a potential toxin domain in the C-terminal domain (CTD), which we termed "IdrD-CTD" (Figure 2.1A). The predicted IdrD-CTD polypeptide resembles many polymorphic toxins: it is physically transferred from one cell into its adjacent neighbor, is found only in a subset of *P. mirabilis* strains, is encoded within a gene cluster that varies in gene content between strains, and is encoded within a gene with a conserved

repetitive element (Rhs). Using IdrD-CTD of *P. mirabilis*, we set out to elucidate the function of this novel toxin and to elucidate the prevalence of this protein, as well as similar toxins, among human-associated microbial populations.

We engineered the nucleotide sequence for the predicted IdrD-CTD toxin domain into an anhydrotetracycline-inducible vector. As a negative control, we engineered the nucleotide sequence for Green Fluorescent Protein (GFP) into the parent vector. Vectors were introduced into a *P. mirabilis* strain in which *idrD* and the downstream genes are disrupted [12]. We then measured toxic activity upon protein production by measuring for the number of viable cells after growth on surfaces for 72 hours at room temperature. *P. mirabilis* producing GFP grew to a saturating density of 10^{10}, while the strain producing the predicted IdrD-CTD toxin grew only to ~10^{7} (Figure 2.1B). Additionally, swarm migration of *P. mirabilis* producing IdrD-CTD was inhibited compared to controls (Figure A.1B, A.1C). A decrease in viability was also observed when IdrD-CTD was produced in *Escherichia coli* (Figure A.1A). Therefore, IdrD-CTD causes lethality.

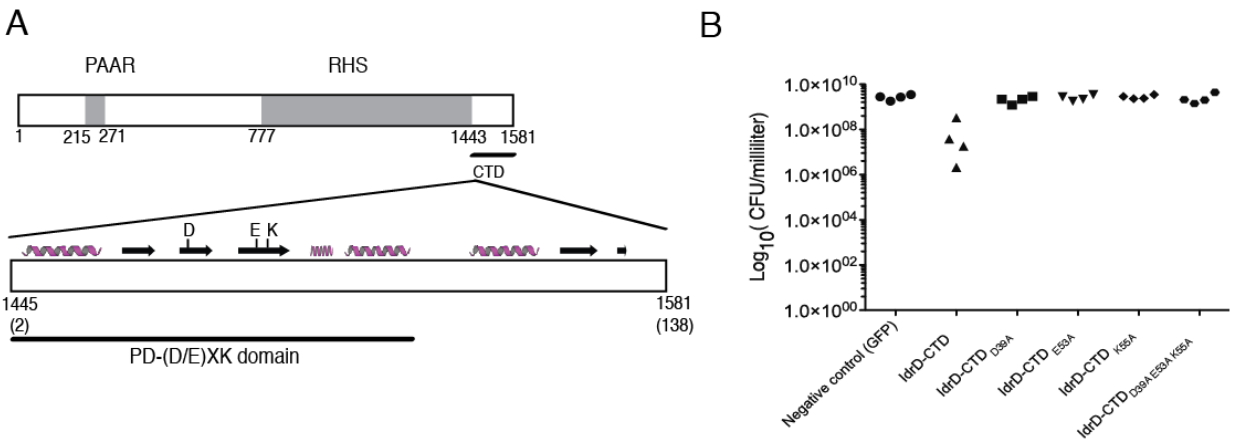
When we examined secondary structure predictions of IdrD-CTD, a structural domain found in the PD-(D/E)XK phosphodiesterase superfamily was revealed (Figs 1A). Members of the PD-(D/E)XK phosphodiesterase superfamily include functionally diverse nucleases involved in replication, restriction, DNA repair, and tRNA–intron splicing [24]. The catalytic core and essential residues for nuclease activity (Figure 2.1A) are known for this superfamily, having originally been characterized for type II restriction endonucleases [25, 26]. Furthermore, two contact dependent inhibition (CDI) toxins were found to belong to this superfamily: two tRNases in *Burkholderia pseudomallei* (isolates E479 and 1026b) [27]. However, IdrD-CTD itself, as well as similar proteins, comprised a not-yet identified and not-yet characterized subfamily.

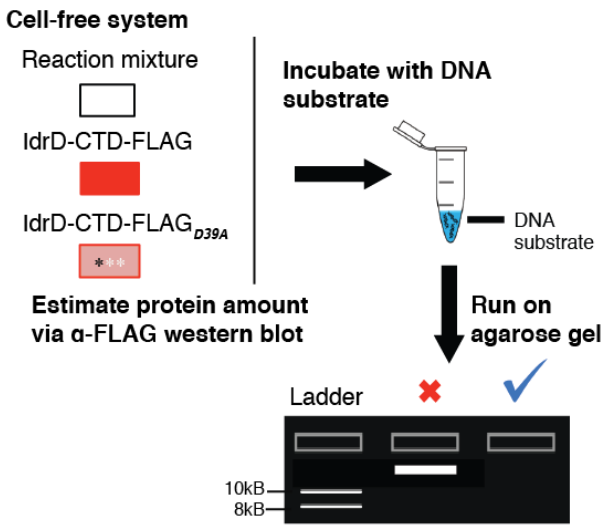
To ascertain whether IdrD-CTD is a member of the PD-(D/E)XK phosphodiesterase superfamily, we first introduced mutations to the essential residues of the predicted catalytic core. We individually replaced D39, E53, and K55 with an alanine residue in the vector-encoded IdrD-CTD; we also disrupted all three at once (Figs 2.1A, 2.1B). We measured for viable cells after 72 hours at room temperature as above. All mutant strains displayed increased viability and migration as compared to the wild-type IdrD-CTD (Figs 2.1B, A.1B, A.1C). We also observed no lethality when these mutant proteins were produced in *E. coli* (Figure A.1A). Therefore, IdrD-CTD contains a catalytic core consistent with the PD-(D/E)XK phosphodiesterase superfamily.

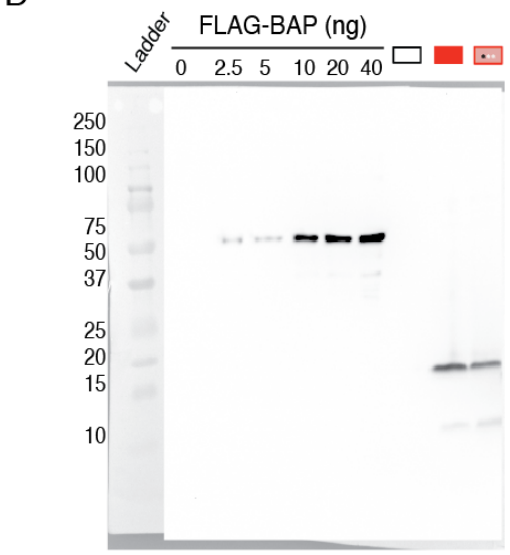
We reasoned that IdrD-CTD contains nuclease activity and set out to determine the nucleic acid target. We engineered DNA constructs to produce either IdrD-CTD or IdrD-CTD_{D39A}, each of which contained a C-terminal FLAG epitope tag, in a commercial PURExpress cell-free system (Figure 2.1C). We confirmed that each protein (~ 17 kDa) was produced by using electrophoresis across a Tris-tricine gel followed by Western blot analysis with an anti-FLAG antibody (Figure 2.1D). Protein yields of IdrD-CTD-FLAG were low, suggesting that either DNA, tRNA, or rRNA was self-limiting in reaction mixture.

As rRNA and tRNA are already present in the reaction mixture, we provided additional DNA and then assayed for nuclease activity. Samples were analyzed after 1 hour incubation at 37°C by electrophoresis on 1% agarose gels stained with ethidium bromide (Figure 2.1C). A reaction with no template DNA was used as a negative control in the nuclease activity assays. We examined the ability to cut methylated or unmethylated DNA, and if so, whether degradation was correlated with protein amount. We added increasing amounts of PURExpress reaction to produce increasing quantities of IdrD-CTD-FLAG or IdrD-CTD_{D39A}-FLAG, from 2.5 ng to 10

ng. We observed degradation of both methylated and unmethylated lambda DNA in the presence of IdrD-CTD-FLAG (Figure 2.1E). The band intensity of lambda DNA inversely correlated with the amount of IdrD-CTD protein present (Figure 2.1E). By contrast, degradation was not apparent in samples containing the negative control or IdrD-CTD$_{D39A}$-FLAG, even at 10 ng (Figure 2.1E; A.1D). Further, bands corresponding to RNA increased in intensity as the amount of PURExpress reaction mixture was increased across samples (Figure A.1D), suggesting that IdrD-CTD does not degrade RNA under these conditions. To examine whether IdrD-CTD was capable of endonuclease activity, we measured degradation of supercoiled or linearized plasmid DNA (~13,500 basepairs) using equivalent reaction conditions. We observed a single band at the expected size for samples containing the purified DNA, the negative control, or IdrD-CTD$_{D39A}$-FLAG but not for the sample containing IdrD-CTD-FLAG (Figure A.1E). Thus, IdrD-CTD is a novel endonuclease targeting DNA that belongs to the PD-(D/E)XK phosphodiesterase superfamily


A
PAAR
RHS
1
215 271
777
1443 1581
CTD
D
E K
1445
(2)
1581
(138)
PD-(D/E)XK domain
B
Log10(CFU/milliliter)
1.0×10^10
1.0×10^08
1.0×10^06
1.0×10^04
1.0×10^02
1.0×10^00
Negative control (GFP)
IdrD-CTD
IdrD-CTD D39A
IdrD-CTD E53A
IdrD-CTD K55A
IdrD-CTD D39A E53A K55A


C
Cell-free system
Reaction mixture
IdrD-CTD-FLAG
IdrD-CTD-FLAG D39A
Incubate with DNA substrate
DNA substrate
Estimate protein amount via α-FLAG western blot
Run on agarose gel
Ladder
10kB
8kB


D
Ladder
FLAG-BAP (ng)
0 2.5 5 10 20 40
250
150
100
75
50
37
25
20
15
10

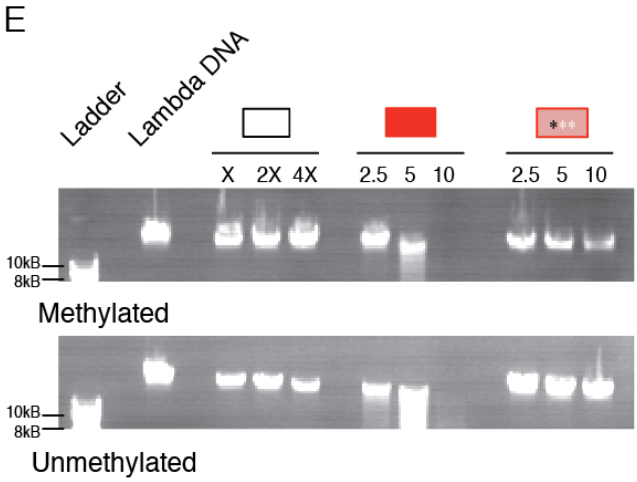

E
Ladder
Lambda DNA
X 2X 4X
2.5 5 10
2.5 5 10
10kB
8kB
Methylated
10kB
8kB
Unmethylated

Figure 2.1: *P. mirabilis* IdrD-CTD is a novel DNase in the PD-(D/E)XK superfamily. (A) Schematic of full-length IdrD and IdrD-CTD (drawn to scale). Amino acid numbers are indicated along the bottom, and gray boxes denote PAAR and RHS domains in the N-terminal region. Predicted secondary structure of IdrD-CTD is shown with catalytic residues labeled within PD-(D/E)XK domain (underlined). (B) Quantification of viable cells after overexpression of IdrD-CTD and active site mutants in swarmer cells. *P. mirabilis* BB2000 *idrD** swarmer cells containing expression vectors of IdrD-CTD, single and triple mutants assayed for colony-forming units per milliliter (plotted on a $\log_{10}$ scale) compared to negative protein production (GFP) control. (C) Schematic of *in vitro* DNase assay. IdrD-CTD-FLAG was produced in a cell-free system and added to a DNA substrate to test for DNase activity. (D) IdrD-CTD-FLAG production. α-FLAG western blot with gradient of known concentrations of FLAG-tagged *E. coli* bacterial alkaline phosphatase (FLAG-BAP) used to estimate amount of IdrD-CTD-FLAG (E) *In vitro* DNase assay of IdrD-CTD. Increasing concentrations of IdrD-CTD-FLAG and $IdrD_{D39A}$-CTD-FLAG (2.5, 5 and 10 ng) were added to methylated and unmethylated lambda DNA (48,502 bp) and run on an agarose gel with NEB 2-log DNA ladder. Bands running below 1kB are presumed to be rRNA and tRNA from PURExpress reaction.

Similar toxins in the genus Rothia *reveal the C-terminal region after the PD-(D/E)XK active site is required for DNase function*

Having found that IdrD-CTD represents a new DNase, we searched for this gene in other bacteria to ascertain its distribution by querying public sequence databases. We found that the closest predicted protein with sequence similarity to IdrD-CTD is found in *Rothia* spp., which are Gram-positive inhabitants of the normal flora of the human oral cavity and pharynx [4, 28, 29]. Interactions between *P. mirabilis* and *Rothia* have not been reported. The identified proteins were found in *R. aeria* F0184, *R. sp. Olga*, *R. aeria* C6B, and *R. aeria* C6D, the latter two containing two copies per genome (Figure 2.2A). All predicted peptides contained the critical residues of the catalytic core (Fig. 2.2A). To evaluate whether the *R. aeria* C6B_10599 protein contained DNA-degrading activity, we engineered DNA constructs to either produce IdrD^{Rothia}-CTD or IdrD$_{D39A}$^{Rothia}-CTD, each of which contained a C-terminal FLAG epitope tag, in a PURExpress cell-free reaction mixture (Figure 2.2B). Like IdrD$_{D39A}$^{Proteus}-CTD, the IdrD$_{D39A}$^{Rothia}-CTD construct contains a disruption in the catalytic core. The *Rothia*-originating peptides were subjected to equivalent analyses as the *P. mirabilis*-originating proteins (Figure 2.2B). We observed degradation of both methylated and unmethylated lambda DNA in the presence of IdrD^{Rothia}-CTD-FLAG, starting at ~ 2.5 ng of protein (Figure 2.2C, A.2A). By contrast, degradation was not apparent in samples containing the negative control or IdrD$_{D39A}$^{Rothia}-CTD-FLAG (Figure 2.2C, A.2A). Bands corresponding to rRNA increased in intensity with increasing volumes of PURExpress reaction (Figure A.2A), suggesting that IdrD^{Rothia}-CTD-FLAG is primarily targeting DNA in these reactions. Therefore, we concluded that IdrD^{Rothia}-CTD-FLAG is also a DNase belonging to the PD-(D/E)XK phosphodiesterase superfamily.

The sequences originating from *Rothia* strains differed in length, suggesting that each variant might function differently. The predicted proteins from *R. aeria* F0184 and *R sp. Olga* lack part of the N-terminal region where the first alpha helix of the catalytic core resides, while C6B_10582 and C6D_12695 encode peptides that lack the C-terminal region after the catalytic core (Figure 2A, S2B). We produced each of these truncated *Rothia* proteins using the *in vitro* translation system and assayed for degradation of lambda DNA as detailed above. We found that none of these proteins have DNase activity in spite of containing the three catalytic residues (Figure A.2B). We next examined whether either the equivalent N-terminal region or C-terminal region of IdrD^{Proteus}-CTD was required for the DNase activity. We engineered and produced proteins with individual deletions of each corresponding region in IdrD^{Proteus}-CTD-FLAG and assayed for DNase activity as detailed above. We found that deletion of either the N-terminal region or the C-terminal region in IdrD^{Proteus}-CTD resulted in no degradation of methylated and unmethylated lambda DNA (Figure 2.2D; A.2C). To test if the active site and 3'-regions of IdrD-CTD are modular, we constructed chimeric proteins: one with the active site of the *Proteus* IdrD and the C-terminal region of the *Rothia* toxin, and vice versa. Though these proteins come from significantly different organisms, both constructs exhibit DNase activity against lambda DNA (Figure A.2D). However, further optimization is required to obtain more stable constructs. Thus, the full length of IdrD^{Proteus}-CTD is required for DNA degradation activity.

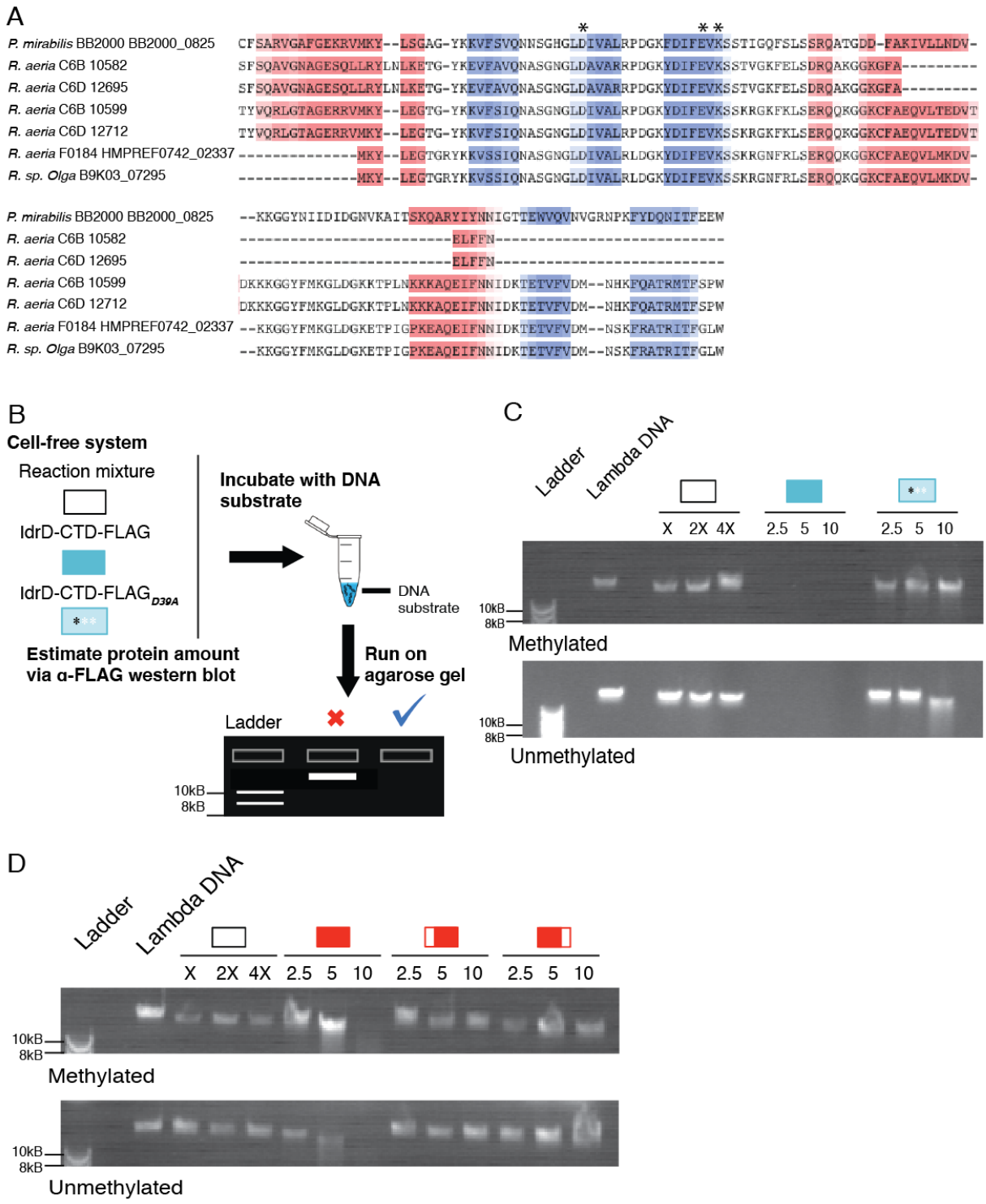

A
P. mirabilis BB2000 BB2000_0825
R. aeria C6B 10582
R. aeria C6D 12695
R. aeria C6B 10599
R. aeria C6D 12712
R. aeria F0184 HMPREF0742_02337
R. sp. Olga B9K03_07295
B
Cell-free system
Reaction mixture
IdrD-CTD-FLAG
IdrD-CTD-FLAG D39A
Estimate protein amount via α-FLAG western blot
Incubate with DNA substrate
DNA substrate
Run on agarose gel
Ladder
10kB
8kB
C
Ladder
Lambda DNA
X 2X 4X 2.5 5 10 2.5 5 10
Methylated
Unmethylated
D
Ladder
Lambda DNA
X 2X 4X 2.5 5 10 2.5 5 10 2.5 5 10
Methylated
Unmethylated

Figure 2.2: Similar IdrD-CTDs in *Rothia* show additional subdomain required for DNase function. (A) Alignment of IdrD-CTD and similar CTDs in three *Rothia* species (MUSCLE followed by Ali2D on MPI Bioinformatics toolkit) [30]. Predicted alpha helices are red and beta sheets are purple; darker color indicates higher confidence. (B) Schematic of *in vitro* DNase assay. IdrD^{Rothia}-CTD was produced in a cell-free system and added to a DNA substrate to test for DNase activity. (C) *In vitro* DNase assay of IdrD^{Rothia}-CTD. Increasing concentrations of IdrD^{Rothia}-CTD-FLAG and IdrD$_{D39A}{}^{Rothia}$-CTD-FLAG (2.5, 5 and 10 ng) were added to methylated and unmethylated lambda DNA (48,502 bp) and run on an agarose gel with NEB 2-log DNA ladder. (D) N- and C-terminal truncations from similar IdrD-CTDs in *R. aeria* in *P. mirabilis* IdrD-CTD. *In vitro* DNase assay as previously described.

Metagenomic analysis of IdrD-CTDs from different bacteria show abundance of subdomains in different communities

We propose that IdrD ^{Proteus}-CTD and IdrD^{Rothia}-CTD comprise a novel DNase subfamily of the PD-(D/E)XK phosphodiesterase superfamily. To identify additional potential members, we searched publicly accessible databases. Employing hmmer and tblastx on these IdrD-CTD sequences, we found 23 additional proteins that are full-length and conserve the critical catalytic core residues (Figure A.3A, A.3B) [30, 31]. These *idrD*-like sequences generally included an extended N-terminal region containing an VENN or Rhs or other conserved motif, and the encoding sequence for the IdrD-CTD-like domain was found within the C-terminal region [7, 32]. The predicted secondary structures are consistent across the length of the predicted proteins (Figure A.3A). Phylogenetic reconstruction of these predicted proteins showed that these proteins are found across the bacterial tree (Figure A.3B). Protein sequences within a genus or species are always more related than to those of other groups, but genera from the same phylum were not always most closely related (Figure A.3B). A species tree based on the full-length 16S rRNA gene (Figure A.3C) did not align with the protein-based tree (Figure A.3B), confirming that proteins from evolutionarily distant bacteria share more similarity than with more related bacteria. Yet, these bacteria are predominantly human-associated, with many being members of the human gut or oral microbiome and/or are in low abundance. Given the toxins' sequence diversity between genera, we hypothesized that the IdrD-CTD-like nucleotide sequences could be used to resolve the distribution of this DNase toxin in human-associated bacterial communities.

Therefore, we investigated the representation of *idrD*-like nucleotide sequences in the human microbiome as captured by metagenomes. Publically-available metagenomes, which are

deep, short-read sequencing of random genomic DNA in a sample, were screened with the PARTIE algorithm using full-length *idrD* sequences from *P. mirabilis* BB2000, *R. aeria* C6B, *Cronobacter turicensis* z3032, and *Prevetolla jejuni* CD3:33 (Figure 2.3A) [33]. We focused on 319 high-quality metagenomes that were annotated as originating from the human gut or the human oral cavity. We then filtered out irrelevant metagenomes for each *idrD*-like sequence by retaining only metagenomes covering at least half of that *idrD* nucleotide sequence. This mapping analysis reports the number of times a metagenomic read mapped to a position along the *idrD*-like sequence (Figure 2.3A), reflecting the abundance of that specific sequence in a metagenome. Smoothly-decreasing coverage at the terminal 5' and 3' ends of a gene remains an artifact of the mapping process. Diversity in *idrD*-like sequences, e.g., deletions or too many nucleotide polymorphisms to be mapped, manifests as changes in coverage along the length of the gene. Ultimately, the coverage of a given region within an *idrD*-like sequence reflects the proportional abundance of that region among in a single metagenomic dataset.

Our metagenome mapping revealed that *idrD*-like sequences are diverse and present in 45% of the human microbiome samples that we analyzed (Figure 2.3B). Each *idrD*-like sequence was abundant only in metagenomes known to host its parent organism (e.g., IdrD^{Proteus}-CTD was abundant in gut but not oral microbiomes), suggesting our mapping analysis can resolve IdrD-like sequences to at least the genus level. Across the entire gene sequence, coverage trends were highly variable, likely reflecting varying degrees of conservation and/or presence among subpopulations. For example, the Rhs core domain generally attracted much more coverage than the rest of the sequence (Figure 2.3B, positions marked with dark green bar). This coverage pattern likely represents multiple Rhs-like genes within a single genus given that the Rhs core sequence is relatively conserved. Of note, sufficient diversity likely exists at the

nucleotide level even within the Rhs domains; we found that individual genera exhibited distinct coverage patterns (Figure 2.3B). We also found that the C-terminal region encoding the rD-CTD-like domains (CTD, light green bar in Figure 2.3B) recruited substantially less but more even coverage and was present in just 8.5% of the metagenomes analyzed (Figure 2.3B, 2.3C). These results likely reflect that each IdrD-CTD variant is restricted to a narrower subpopulation. Thus, while relatively large populations contain *idrD*-like sequences, only a relatively small subpopulation appears to harbor an IdrD-CTD-like domain.

Yet even at this low abundance, the coverage patterns combined with the functional domain analysis of IdrD-CTD hinted that there was more complexity to the occurrence of IdrD-CTD proteins in the microbiomes. Specifically, the 5' region of the *P. jejuni* rD-CTD recruited reads abundantly from oral microbiomes, where the catalytic core resides (dark blue bar in Figure 2.3C), and comparably less in the 3' region (Figure 2.3C). To determine whether this coverage pattern reflected a true truncation, we mapped the same oral metagenomes against *idrD* sequences from *P. jejuni* CD3:33, *P.* sp. C561, *P. fusca* JCM 17724, and *P. denticola* NCTC13067 (Figure 2.3D). The entirety of the *P.* sp. C561 and *P. fusca* JCM 17724 sequences received even coverage with few SNPs, while the *P. jejuni* CD3:33 sequence still received little to no coverage across the 3' region. Thus, the 3' region of this gene contains species-level differences within the *Prevotella* genus, raising the possibility that such nucleotide diversity could be used to detect species-level differences in metagenomic datasets.

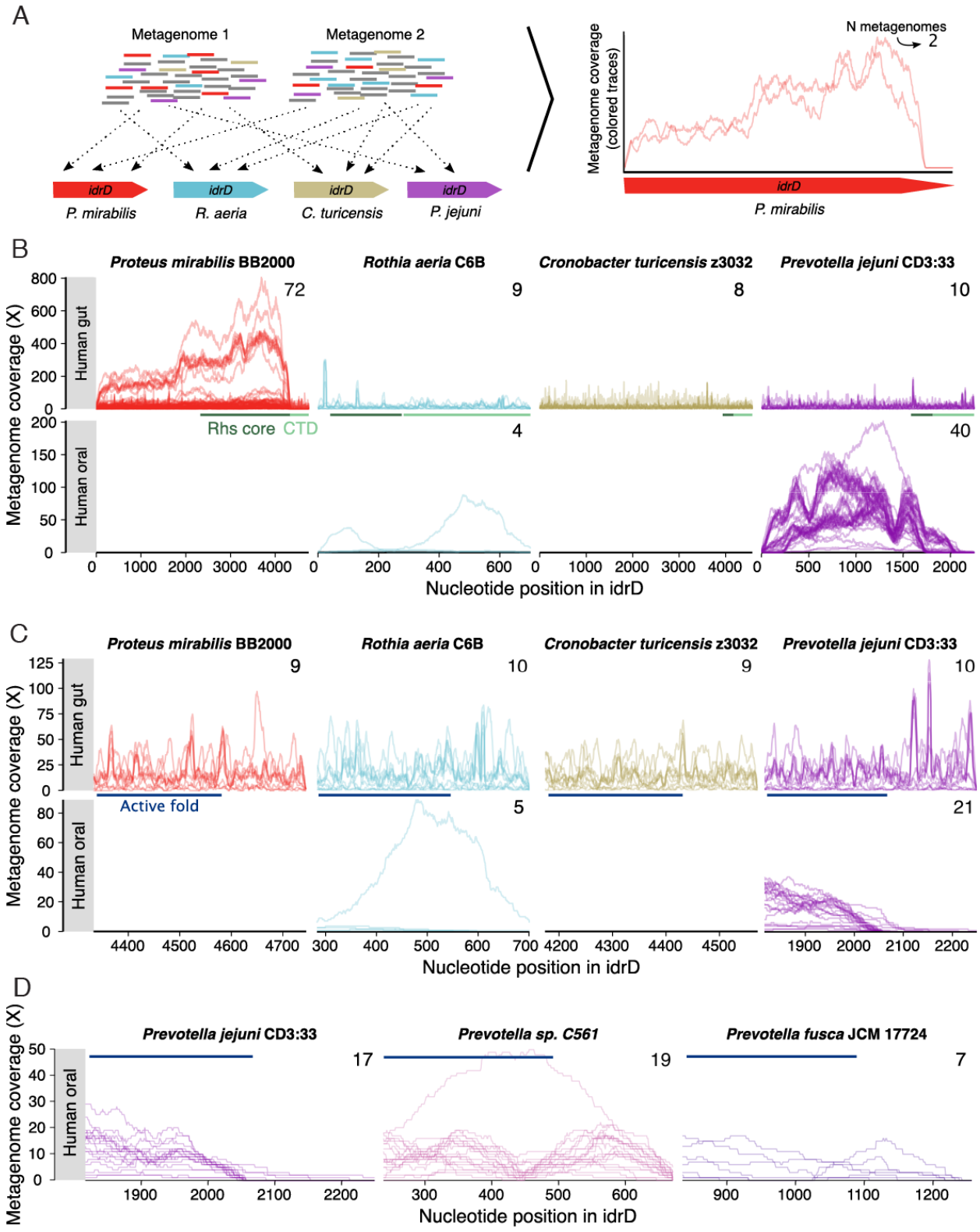

A
Metagenome 1
Metagenome 2
idrD
P. mirabilis
R. aeria
C. turicensis
P. jejuni
Metagenome coverage (colored traces)
N metagenomes
2
idrD
P. mirabilis
B
Proteus mirabilis BB2000
Rothia aeria C6B
Cronobacter turicensis z3032
Prevotella jejuni CD3:33
Metagenome coverage (X)
Human gut
Human oral
72
9
8
10
Rhs core
CTD
4
40
Nucleotide position in idrD
C
Proteus mirabilis BB2000
Rothia aeria C6B
Cronobacter turicensis z3032
Prevotella jejuni CD3:33
Metagenome coverage (X)
Human gut
Human oral
9
10
9
10
Active fold
5
21
Nucleotide position in idrD
D
Prevotella jejuni CD3:33
Prevotella sp. C561
Prevotella fusca JCM 17724
Metagenome coverage (X)
Human oral
17
19
7
Nucleotide position in idrD

Figure 2.3: Metagenome mapping reveals abundance patterns of *idrD* in the human microbiome**.** (A) Cartoon workflow shows how short reads from metagenomes, here depicting two, are mapped against a reference containing the four *idrD* sequences shown here. Based on the mapping results, each positions' coverage, i.e., the number of metagenomic reads mapping to that position, can be plotted for each metagenome (colored, partially transparent lines; each line is a different metagenome). (B) Coverage of each sequence (subpanel columns) by metagenomes originating from the human gut (top row of subpanels) or human oral cavity (bottom row), for metagenomes covering at least half of that *idrD* sequence. The number of metagenomes passing this filter is shown in the top right corner of each subpanel. The positions corresponding to the RHS core and CTD are annotated with dark and light green bars, respectively, between the two rows of subpanels. (C) The x-axis is zoomed in to show CTD coverages (region highlighted by the light green bar in B) by metagenomes covering at least half of that CTD sequence. (D) Coverage is shown for the CTD of three different *Prevotella* species' *idrD* sequences, from metagenomes covering at least half of that CTD sequence. In C and D, the positions corresponding to the critical active site cores are annotated with dark blue bars.

Discussion

Here, we show that IdrD ^{Proteus}–CTD and IdrD^{Rothia}-CTD are endonuclease enzymes targeting DNA; activity requires both the catalytic core and a 3'-subdomain (Figs 2.1 and 2.2). While we have yet to characterize the primary function of this 3'-subdomain, we hypothesize that it might contribute protein-binding to and/or possible DNA-protein interactions. Similar DNA-targeting proteins function as multimeric complexes, while similar toxin proteins can bind allele-specific immunity proteins [6-8, 34]. Characterized restriction endonucleases of the PD(D/E)XK-superfamily differ in the target sequence [35]. Differences in cutting patterns of lambda DNA between IdrD ^{Proteus}–CTD and IdrD^{Rothia}-CTD suggest different target sequences. Sequence specificity could be tested by excising and sequencing the smaller lambda DNA band observed when 5 ng of IdrD ^{Proteus}–CTD (Figure 2.1E). The predicted secondary structures of this region are consistent (Figure A.3A), but further structural studies are necessary to observe similarities and differences in the predicted tertiary structures. Though we were unable to purify IdrD-CTD^{Proteus} due to its toxicity, future structural studies using IdrD $_{D39A}$ ^{Proteus}-CTD as a tool are warranted. Determination and comparisons of IdrD-CTD protein structures would likely reveal features for how these proteins interact with the DNA target and protein partners.

We hypothesize that the IdrD-CTD proteins are indicative of a genetic element that has been mobile earlier in evolutionary history and has also been retained through some microbial lineages. The predominant hypothesis is that toxic C-terminal domains are exchanged through horizontal gene transfer because these toxins can occur in variable genomic hotspots [6-8, 19]. While we cannot conclusively determine this here, identifying the regions required for protein function can improve bioinformatics searches to obtain more similar toxins to address this

hypothesis. We identified a total of 25 members, which can be used to seed the definition of a new protein domain.

Finally, we found that the varying 3' region of the *idrD* nucleotide sequence is a candidate region to define sub-population identity within the *Prevotella* genus (Figure 3) This would allow for detection of species-level differences in metagenomic datasets. The 5' region contains the catalytic core and may be protected from mutations due to constraints in the three-dimensional space for enzymatic function [26, 36-40]. The 3'-region of IdrD^{Proteus}-CTD is essential and has more variation between amino acid sequences across phyla. Therefore, the 3'-region appears more flexible in sequence while still retaining function. If so, it might be more accessible for accumulating mutations allowing for each IdrD-CTD to become distinct.

Materials and Methods:

Bacterial strains and media:

Overnight cultures of all strains were grown aerobically at 37°C in LB broth. Swarm-permissive nutrient plates were made with CM55 blood agar base agar (Oxoid, Basing- stoke, England). *P. mirabilis* strains were maintained on low swarming agar (LSW^-). All swarm and growth media contained 35g/ml kanamycin for plasmid maintenance.

Strain construction:

P_{idrA}-*idrD*-CTD was constructed by PCR amplifying the last 416 bp of the *idrD* gene from BB2000 using primers AS174 and AS175 and cloning it into the SacI and AgeI sites of pAS1034, resulting in plasmid pAS1054. The inducible anhydrotetracycline promoter (Ptet) was introduced into the *idrD-CTD* expression vectors by generating gBlocks (gDS0005) of the promoter region with 29 bp overhangs for the plasmid, and using SLiCE to recombine into pAS1054. This resulted in the plasmid pDS0002 (*idrD-CTD*).

A C-terminal FLAG tag (GACTACAAGGACGACGATGACAAG) was added to *idrD* by using SLiCE to recombine the gBlock gDS0023 (FLAG tag with 49 bp overhang of *idrD-CTD* and 52 bp overhang of pDS0002) into pDS0002. This vector is pDS0034. The FLAG-tagged *idrD* active site mutants were generated by replacing *idrD-CTD-FLAG* in pDS0034 with the mutants sequences which are encoded in gBlocks gDS0025-28—resulting in plasmids pDS0048 (D1482A), pDS0049 (E1496A), pDS0050 (K1498A), and pDS0051 (triple mutant). The untagged versions of *idrD* were constructed by PCR amplifying the mutant *idrD* sequences from pDS0048-51 with oDS0137 and oDS0159 to remove FLAG tag, and performing a restriction digest with SacI and AgeI to insert into pDS0034 (pDS0058-61).

To generate pDS0062, a *gfp* expression vector under the control of Ptet, primers oDS0161 and oDS0162 were used to amplify *gfpmut2* from p*idsBB-idsE*-GFP, and put into pDS0034 through restriction digest with SacI and AgeI.

All plasmids were confirmed by Sanger sequencing (Genewiz). Plasmids with conjugative transfer elements, including all *idr* expression vectors, were moved into the *Escherichia coli* conjugative strain S17 which were then mated with recipient *P. mirabilis* strains. The presence of plasmids were confirmed in recipient strains by PCR using plasmid-specific primers.

***P. mirabilis* Swarm assay**:

Overnight cultures of BB2000 *idrD** carrying each expression vector were normalized to an optical density at 600nM (OD_{600}) of 1; swarm-permissive nutrient plates supplemented with kanamycin and 10nM anhydrotetraclycine were inoculated with 1uL of normalized culture. Plates were incubated for 48 hours at room temperature. The radii of the migrating swarms were measured. Swarms were then resuspended in 6 ml of LB broth; 20uL of this resuspension was used for a 10-fold dilution series (total of 8 dilutions). 10uL of each dilution was spotted onto LSW^- agar plates supplemented with kanamycin.

***E. coli* liquid viability assay**:

Overnight cultures of *E. coli* MG1655 carrying each expression vector were normalized to an optical density at 600nM (OD_{600}) of 1. 2 ul of normalized cultures was added to 198ul of LB broth supplemented with kanamycin and 10nM anhydrotetraclycine and grown at 37°C for 16 hours in a 96 well-plate. OD_{595} reading were taken every half hour.

***In vitro* DNase assay**:

IdrD-CTD and IdrD-CTD_{D1482A} with a C-terminal FLAG epitope tag was produced using the New England Biolabs PURExpress *In Vitro* Protein Synthesis Kit. Template DNA was amplified from pDS0034 (IdrD-CTD-FLAG) or pDS0048 (IdrD-CTD-FLAG D1482A) using primers with overhangs to add the required elements specified by the PURExpress kit. Reactions were performed with 250ng of template DNA (no template DNA added to negative control reaction) and incubated at 37°C for two hours. Protein amount was determined using an α–FLAG western blot with a known gradient of FLAG-BAP (2.5, 5, 10, 20 ng). Protein (2.5, 5, and 10 ng) was added to 0.5 ug of lambda DNA (methylated and unmethylated-), 5uL of New England Biolabs Buffer 3.1, and up to a final volume of 25uL. For plasmid DNAse assays, 10 ng of protein was added to 250 ng of circular or linear plasmid DNA (pidsBB). This reaction was incubated for one hour at 37°C, then Proteinase K (New England Biolabs) was added and incubated for 15 minutes at 37°C. Reaction was then run on a 1% agarose gel for analysis.

Western blotting:

In vitro translation reaction samples described above were run on a 12% Tris-Tricine polyacrylamide gel, transferred to a nitrocellulose membrane, probed with rabbit anti-FLAG (1:4,000; Sigma-Aldrich, St. Louis, MO), then goat anti-rabbit conjugated to horseradish peroxidase (HRP) (1:5,000; KPL, Inc., Gaithersburg, MD), and finally developed using Immun-Star HRP substrate kit (Bio-Rad Laboratories, Hercules, CA). Visualization of blots were done using a Chemidoc (Bio-Rad Laboratories, Hercules, CA) and TIFF files were used for analysis on Fiji (ImageJ, Madison, WI).

Phylogenetic reconstruction of *idrD* and species relationships

The amino acid sequences for the 25 *idrD*-CTD genes identified as described above were obtained and aligned with muscle [41]before removing positions with less than 70% occupancy with trimAl[42]. This alignment was passed to MrBayes v3.2.6 [43] to reconstruct the tree using the WAG substitution model [44] and gamma model of rate heterogeneity. Four independent runs, each run for 20 million generations sampled every 20,000 generations by four coupled chains heated at the default temperature of 0.2, were checked for convergence and then combined into a 50% majority rule tree after burning the initial 40% of the sampled trees.

The species tree was reconstructed using identical methods but with full-length 16S ribosomal RNA sequences obtained from the various genomes, choosing arbitrarily if multiple 16S rRNA copies were found in a genome. Specifically, alignment and MrBayes parameters were identical except for changing the model to GTR [45].

Maximum-likelihood trees were also generated with RAxML [46] using the appropriate WAG or GTR + gamma models and produced identical topologies, albeit with less support.

Metagenome selection and processing

A fully-reproducible workflow explaining and documenting all commands and scripts used to perform the metagenome analyses will be published online. Full-length *idrD* nucleotide sequences were obtained from *Acinetobacter baumannii* XH858, *Cronobacter turicensis* z3032, *Prevotella jejuni* CD3:33, *Proteus mirabilis* BB2000, *Pseudomonas fluorescens* F113, *Rothia aeria* C6B, and *Xanthomonas citri* pv. *malvacearum* XcmN1003. Publically-available metagenomes likely to represent populations with these genes were identified using the Search SRA portal (www.searchsra.org) that employs the PARTIE algorithm [33]. Briefly, PARTIE uses bowtie2 [47] to search a 100,000-read random subset from each of ~110,000 published

metagenomes. From this, 3,801 candidate metagenomes were identified that contributed non-zero coverage to at least one *idrD* sequence, from which we curated a list of 1,189 metagenomes by removing genome assemblies, transcriptomes, non-random library preparation methods, etc. 1,188 of these metagenomes were downloaded from available on NCBI's Short Read Archive (fastq-dump --split-3) for deeper analysis (1 metagenome could not be downloaded). A single bowtie2 database was created with all seven *idrD* nucleotide sequences, onto which bowtie2 mapped metagenomes using default parameters (--sensitive; [47]). Anvi'o, an analysis and visualization platform for 'omics data, managed the resultant data and subsequent analyses [48]. With Anvi'o, a contigs database was generated (anvi-gen-contigs-db command) from the seven *idrD* sequences and profiled with the merged results of the bowtie2 mapping (anvi-profile and anvi-merge commands, respectively). Per-nucleotide coverages and variability (i.e, single-nucleotide polymorphism (SNP) counts) were exported for all sequences with the Anvi'o anvi-get-split-coverages and anvi-gen-variability-profile commands, respectively.

To investigate the distribution of *idrD* coverage among different *Prevotella* in the human oral cavity, *idrD* sequences from the three additional *Prevotella* spp. (*P.* sp. C561, *P. fusca* JCM 17724, *P. denticola* NCTC 13067) along with the *P. jejuni* CD3:33 *idrD* sequence were mapped separately, using the same methods but mapping reads from only the 202 metagenomes from the human oral cavity.

Metagenome categorization

Metagenomes were binned into categories based on the provided "ScientificName" annotation. Metagenomes from the Human Microbiome Project (HMP) were listed as "human metagenome"; these were disaggregated into "HMP oral metagenome" and "HMP gut metagenome" based on the sampled site listed in the "Analyte_Type" column. From the 1,188

metagenomes, we focused on only the 324 human oral or human gut metagenomes, defining human oral as metagenomes labelled “human oral metagenome”, "HMP oral metagenome", or "Non-HMP oral metagenome" (n=42) and human gut as metagenomes labelled “human gut metagenome” or “human metagenome” (n=277). Metagenomes annotated as “human metagenome” (n=47) were categorized as human gut since though they originated from a variety of human body sites, the only three passing the filtration criteria (see next section) were from the human gut. Metagenomes annotated as “oral metagenome” (n=112) were excluded as they provided extremely low (single digit), noisy coverages (data not shown). In addition, five metagenomes were specifically discarded: SRR628272 was removed from all datasets as the original FASTQ had extremely low quality scores; SRR1779144 came from a diseased infant (https://www.ncbi.nlm.nih.gov/sra/?term=SRR1779144) and skewed the y-axis with an extremely high *C. turicensis* coverage (>1,000x); SRR2047620 came from a pediatric stem-cell treatment dataset with high antibiotic loads (https://www.ncbi.nlm.nih.gov/sra/?term=SRR2047620) with similarly extreme *C. turicensis* coverage; SRR1781983 came from a subgingival plaque of a patient with periodontitis (https://www.ncbi.nlm.nih.gov/biosample/SAMN03287617) and had extremely high *R. aeria* coverage that skewed the y-axis (>500x); and SRR1038387 came from an infant with necrotizing enterocolitis (https://www.ncbi.nlm.nih.gov/sra/?term=SRR1038387) with extremely high *P. mirabilis* coverage that skewed the y-axis (>2,000x). 1,183 metagenomes remained after discarding these metagenomes, of which 319 were human oral or human gut metagenomes. Moving forward, we focused exclusively on *P. mirabilis, R. aeria*, *C. turicensis*, and *P. jejuni* as the other taxa’s *idrD* sequences had little to no coverage from the human metagenomes of interest (Figure A.5).

Mapping filtration criteria

To minimize noisy coverage originating from metagenomes relevant for one organism but not another, we employed a filtering strategy that, for each *idrD* sequence, considered only metagenomes from which at least half of the nucleotides received coverage. For Figures 4C and 4D, the filtering criterion was applied after subsetting to the C-terminal domain (CTD); that is, the filtration was applied based only on the CTD. The number of metagenomes passing the criterion is displayed in the top right corner of each subpanel in Figure 4. For each metagenome passing the filtration step, each nucleotides' coverage is plotted as a partially-transparent line; thus, metagenomes with similar coverage trends overlap and appear darker.

SNP information from the metagenomes, relative to the seven *idrD* reference sequences, is displayed by vertical bars showing the Shannon Entropy [49]of the frequencies of each nucleotide across the metagenomes (Figure A.6). Higher entropy values correspond to more even SNP diversity, while an entropy of 0 signifies an invariant position. Entropy values were calculated using the observed nucleotide frequencies found in all metagenomic reads mapping to a given nucleotide position.

Table 2.1: Strains used in this study

Strain	Notes	KAG #	DS #	Source
P. mirabilis BB2000 *idrD** + pDS0062	*idrD::Tn-Cm(R)* producing GFPmut2 under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3277	349	This study
P. mirabilis BB2000 *idrD** + pDS0002	*idrD::Tn-Cm(R)* producing IdrD-CTD under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2178	104	This study
P. mirabilis BB2000 *idrD** + pDS0058	*idrD::Tn-Cm(R)* producing IdrD-CTD_{D39A} under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3236	344	This study
P. mirabilis BB2000 *idrD** + pDS0059	*idrD::Tn-Cm(R)* producing IdrD-CTD_{E53A} under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3237	345	This study
P. mirabilis BB2000 *idrD** + pDS0060	*idrD::Tn-Cm(R)* producing IdrD-CTD_{K55A} under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3238	346	This study

Table 2.1 (continued): Strains used in this study

Strain	Notes	KAG #	DS #	Source
P. mirabilis BB2000 *idrD** + pDS0061	*idrD::Tn-Cm(R)* producing IdrD-CTD$_{D39A\ E53A\ K55A}$ under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3239	347	This study
E. coli MG1655 + pBBR1-NheI	MG1655 carrying empty vector (pBBR1 origin, Kan (R))	2076	68	This study
E. coli MG1655 + pDS0002	MG1655 producing IdrD-CTD under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2298	151	This study
E. coli MG1655 + pDS0058	MG1655 producing IdrD-CTD$_{D39A}$ under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3228	336	This study
E. coli MG1655 + pDS0059	MG1655 producing IdrD-CTD$_{E53A}$ under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3229	337	This study
E. coli MG1655 + pDS0060	MG1655 producing IdrD-CTD$_{K55A}$ under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3230	338	This study

Table 2.1 (continued): Strains used in this study

Strain	Notes	KAG #	DS #	Source
E. coli MG1655 + pDS0061	MG1655 producing IdrD-CTD$_{D39A\ E53A\ K55A}$ under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3231	339	This study
OneShot Omnimax 2 T1R Competent Cells	*E. coli* strain for cloning			Thermo Fisher Scientific, Waltham, MA.
S17λpir	*E. coli* mating strain to introduce plasmids into *P. mirabilis*	068		[50]

Table 2.2: Plasmids used in this study

Plasmid	Cloning method (or source)	Primers and gBlocks (5'=> 3')
pBBR1-NheI	[10]	
pDS0002	Anhydrotetracycline promoter (Ptet) with 29bp overhangs (gDS0005) was recombined into amplified pAS1054 by SLiCE	oDS0005: gctagccatttgcccatgg oDS0006: cgtttttgataaaaggatattgttgag gDS0005: tcgcccacccccatgggcaaatggctagcttaagacccactttcacatttaagttgtt tttctaatccgcatatgatcaattcaaggccgaataagaaggctggctctgcaccttg gtgatcaaataattcgatagcttgtcgtaataatggcggcatactatcagtagtaggt gtttccctttcttctttagcgacttgatgctcttgatcttccaatacgcaacctaaagtaa aatgccccacagcgctgagtgcatataatgcattctctagtgaaaaaccttgttggc ataaaaaggctaattgattttcgagagtttcatactgtttttctgtaggccgtgtaccta aatgtacttttgctccatcgcgatgacttagtaaagcacatctaaaacttttagcgttat tacgtaaaaaatcttgccagctttccccttctaaagggcaaaagtgagtatggtgcct atctaacatctcaatggctaaggcgtcgagcaaagcccgcttatttttacatgccaa tacaatgtaggctgctctacacctagcttctgggcgagtttacgggttgttaaaccttc gattccgacctcattaagcagctctaatgcgctgttaatcactttacttttatctaatcta gacatcattaattcctaatttttgttgacactctatcgttgatagagttattttaccactcc ctatcagtgatagagaaagtttttgataaaaggatattgttgagcac
pDS0058	*idrD-CTD*$_{D39A}$ was amplified from pDS0048 (to remove FLAG-tag) and ligated in Ptet vector using restriction digest (SacI and AgeI)	oDS0137: gtcaaggagctctcatgtgc oDS0159: caataaaccggtctaccattcctcaaacgttatattc
pDS0059	*idrD-CTD*$_{E53A}$ was amplified from pDS0049 (to remove FLAG-tag) and ligated in Ptet vector using restriction digest (SacI and AgeI)	Same as above

Table 2.2 (continued): Plasmids used in this study

Plasmid	Cloning method (or source)	Primers and gBlocks (5'=> 3')
pDS0060	*idrD-CTD*$_{K55A}$ was amplified from pDS0050 (to remove FLAG-tag) and ligated in Ptet vector using restriction digest (SacI and AgeI)	Same as above
pDS0061	*idrD-CTD*$_{D39A\ E53A\ K55A}$ was amplified from pDS0051 (to remove FLAG-tag) and ligated in Ptet vector using restriction digest (SacI and AgeI)	Same as above
pDS0062	*gfpmut2* was amplified and ligated in Ptet vector using restriction digest (SacI and AclI)	oDS0161: gtacatgagctctcatgagtaaaggagaagaacttttc oDS0162: caataaaccggtctatttgtatagttcatccatgcc

References

1. Lloyd-Price, J., et al., *Strains, functions and dynamics in the expanded Human Microbiome Project.* Nature, 2017. **550**(7674): p. 61-66.

2. Costea, P.I., et al., *Subspecies in the global human gut microbiome.* Mol Syst Biol, 2017. **13**(12): p. 960.

3. Eren, A.M., et al., *Oligotyping analysis of the human oral microbiome.* Proc Natl Acad Sci U S A, 2014. **111**(28): p. E2875-84.

4. Aas, J.A., et al., *Defining the normal bacterial flora of the oral cavity.* J Clin Microbiol, 2005. **43**(11): p. 5721-32.

5. Verster, A.J., et al., *The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition.* Cell Host Microbe, 2017. **22**(3): p. 411-419.e4.

6. Koskiniemi, S., et al., *Rhs proteins from diverse bacteria mediate intercellular competition.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 7032-7.

7. Aoki, S.K., et al., *A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria.* Nature, 2010. **468**(7322): p. 439-42.

8. Zhang, D., L.M. Iyer, and L. Aravind, *A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems.* Nucleic Acids Res, 2011. **39**(11): p. 4532-52.

9. DIENES, L., *Reproductive processes in Proteus cultures.* Proc Soc Exp Biol Med, 1946. **63**(2): p. 265-70.

10. Gibbs, K.A., M.L. Urbanowski, and E.P. Greenberg, *Genetic determinants of self identity and social recognition in bacteria.* Science, 2008. **321**(5886): p. 256-9.

11. Gibbs, K.A., L.M. Wenren, and E.P. Greenberg, *Identity gene expression in Proteus mirabilis.* J Bacteriol, 2011. **193**(13): p. 3286-92.

12. Wenren, L.M., et al., *Two independent pathways for self-recognition in Proteus mirabilis are linked by type VI-dependent export.* MBio, 2013. **4**(4).

13. Alteri, C.J., et al., *Multicellular bacteria deploy the type VI secretion system to preemptively strike neighboring cells.* PLoS Pathog, 2013. **9**(9): p. e1003608.

14. Cardarelli, L., C. Saak, and K.A. Gibbs, *Two Proteins Form a Heteromeric Bacterial Self-Recognition Complex in Which Variable Subdomains Determine Allele-Restricted Binding.* MBio, 2015. **6**(3): p. e00251.

15. Saak, C.C. and K.A. Gibbs, *The Self-Identity Protein IdsD Is Communicated between Cells in Swarming Proteus mirabilis Colonies.* J Bacteriol, 2016. **198**(24): p. 3278-3286.

16. Alteri, C.J., et al., *Subtle variation within conserved effector operon gene products contributes to T6SS-mediated killing and immunity.* PLoS Pathog, 2017. **13**(11): p. e1006729.

17. Zepeda-Rivera, M.A., C.C. Saak, and K.A. Gibbs, *A Proposed Chaperone of the Bacterial Type VI Secretion System Functions To Constrain a Self-Identity Protein.* J Bacteriol, 2018. **200**(14).

18. Tipping, M.J. and K.A. Gibbs, *Peer pressure from a Proteus mirabilis self-recognition system controls participation in cooperative swarm motility.* bioRxiv 490771, 2018.

19. Jackson, A.P., et al., *Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement.* BMC Genomics, 2009. **10**: p. 584.

20. Alvarez-Martinez, C.E. and P.J. Christie, *Biological diversity of prokaryotic type IV secretion systems.* Microbiol Mol Biol Rev, 2009. **73**(4): p. 775-808.

21. Aoki, S.K., et al., *Contact-dependent inhibition of growth in Escherichia coli.* Science, 2005. **309**(5738): p. 1245-8.

22. Russell, A.B., et al., *Type VI secretion delivers bacteriolytic effectors to target cells.* Nature, 2011. **475**(7356): p. 343-7.

23. Zhang, D., et al., *Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics.* Biol Direct, 2012. **7**: p. 18.

24. Steczkiewicz, K., et al., *Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily.* Nucleic Acids Res, 2012. **40**(15): p. 7016-45.

25. Selent, U., et al., *A site-directed mutagenesis study to identify amino acid residues involved in the catalytic function of the restriction endonuclease EcoRV.* Biochemistry, 1992. **31**(20): p. 4808-15.

26. Venclovas, C., A. Timinskas, and V. Siksnys, *Five-stranded beta-sheet sandwiched with two alpha-helices: a structural link between restriction endonucleases EcoRI and EcoRV.* Proteins, 1994. **20**(3): p. 279-82.

27. Johnson, P.M., et al., *Functional Diversity of Cytotoxic tRNase/Immunity Protein Complexes from Burkholderia pseudomallei.* J Biol Chem, 2016. **291**(37): p. 19387-400.

28. ROTH, G.D. and A.N. THURN, *Continued study of oral nocardia.* J Dent Res, 1962. **41**: p. 1279-92.

29. Georg, L.K. and J.M. Brown, *<i style="">Rothia , gen. nov. an aerobic genus of the family Actinomycetaceae.* International Journal of Systematic and Evolutionary Microbiology, 1967. **17**: p. 79-88.

30. Eddy, S.R., *A new generation of homology search tools based on probabilistic inference.* Genome Inform, 2009. **23**(1): p. 205-11.

31. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

32. Lin, R.J., M. Capage, and C.W. Hill, *A repetitive DNA sequence, rhs, responsible for duplications within the Escherichia coli K-12 chromosome.* J Mol Biol, 1984. **177**(1): p. 1-18.

33. Torres, P.J., R.A. Edwards, and K.A. McNair, *PARTIE: a partition engine to separate metagenomic and amplicon projects in the Sequence Read Archive.* Bioinformatics, 2017. **33**(15): p. 2389-2391.

34. Kosinski, J., M. Feder, and J.M. Bujnicki, *The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function.* BMC Bioinformatics, 2005. **6**: p. 172.

35. Pingoud, A. and A. Jeltsch, *Structure and function of type II restriction endonucleases.* Nucleic Acids Res, 2001. **29**(18): p. 3705-27.

36. Skirgaila, R., et al., *Structure-based redesign of the catalytic/metal binding site of Cfr10I restriction endonuclease reveals importance of spatial rather than sequence conservation of active centre residues.* J Mol Biol, 1998. **279**(2): p. 473-81.

37. Bujnicki, J.M. and L. Rychlewski, *Identification of a PD-(D/E)XK-like domain with a novel configuration of the endonuclease active site in the methyl-directed restriction enzyme Mrr and its homologs.* Gene, 2001. **267**(2): p. 183-91.

38. Pingoud, V., et al., *Evolutionary relationship between different subgroups of restriction endonucleases.* J Biol Chem, 2002. **277**(16): p. 14306-14.

39. Tamulaitis, G., A.S. Solonin, and V. Siksnys, *Alternative arrangements of catalytic residues at the active sites of restriction enzymes.* FEBS Lett, 2002. **518**(1-3): p. 17-22.

40. Feder, M. and J.M. Bujnicki, *Identification of a new family of putative PD-(D/E)XK nucleases with unusual phylogenomic distribution and a new type of the active site.* BMC Genomics, 2005. **6**: p. 21.

41. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

42. Capella-Gutiérrez, S., J.M. Silla-Martínez, and T. Gabaldón, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.* Bioinformatics, 2009. **25**(15): p. 1972-3.

43. Ronquist, F. and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-4.

44. Whelan, S. and N. Goldman, *A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.* Mol Biol Evol, 2001. **18**(5): p. 691-9.

45. S, T., *Some probabilistic and statistical problems in the analysis of DNA sequences , in Some Mathematical Questions in Biology.* DNA Sequence Analysis, 1986: p. 57–86.

46. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.

47. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

48. Eren, A.M., et al., *Anvi'o: an advanced analysis and visualization platform for 'omics data.* PeerJ, 2015. **3**: p. e1319.

49. Shannon, C.E., *A mathematical theory of communication.* The Bell System Technical Journal, 1948. **27**(3): p. 379 - 423.

50. Simon, R., U. Priefer, and A. Puhler, *A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. <i data-test="*journal-title" style="color: rgb(34, 34, 34); font-family: "Source Sans Pro", arial, helvetica, sans-serif; font-size: 14px; letter-spacing: 0.17px; background-color: rgb(255, 255, 255);">Bio/Technology, 1983(1): p. 784 – 791.

Chapter 3

Characterization of IdrE, an immunity protein that counteracts IdrD-CTD-mediated toxicity

All of the work presented in this chapter adapted from a manuscript in progress written by Denise Sirias, Emma Keteku, Dan Utter, and Dr. Karine A. Gibbs. Emma Keteku, Denise Sirias, Sajal Akkipeddi, and Abigail Knecht performed biochemical, genetic and physiological characterization, Dan Utter performed metagenomic analysis, and Dr. Karine A. Gibbs advised. Biochemical data was also presented in Emma Keteku's senior thesis (The Role of IdrE in an Effector/Immunity System of *Proteus mirabilis*).

Abstract

Polymorphic toxin systems are widely distributed among bacteria and are implicated in bacterial competitions. Here, we characterize the immunity protein IdrE of the Rhs polymorphic toxin IdrD found in *Proteus mirabilis* BB2000. Co-expression of IdrE rescues cells from IdrD-CTD-mediated toxicity. We show that IdrD-CTD detection is decreased when co-expressed with IdrE, indicating that IdrE may counteract toxicity by promoting degradation of IdrD-CTD. Predicted secondary structure comparisons of IdrE homologs show a conserved structural domain in the C-terminal region, which is functionally confirmed to be sufficient for loss of IdrD-CTD when co-expressed. Further, we use metagenomic analysis to probe abundance of IdrE, showing that it co-occurs with IdrD-CTD.

Introduction

Bacteria reside in communities where they are involved in competitive interactions. One mechanism of bacterial competition is through polymorphic toxin systems (PTS), such as the Rhs protein family [1, 2]. They contain toxins (or effectors) with a conserved N-terminal region, followed by a divergent region in the C-terminal domain (CTD) [3-5]. The toxic activity of the effectors is found in the CTD[2-6]. These CTDs can be transported by different secretion systems, which is determined by the N-terminal region [3-5]. Cell contact is required to deploy these toxins. Polymorphic toxins are tightly linked to small, downstream open reading frames that encoded immunity proteins that counteract the effects of the toxin [1, 2, 4, 5]. PTS have been implicated in interstrain competition, providing a competitive advantage by inhibiting growth of cells lacking the cognate immunity protein [1, 2, 4, 7].

We reported an *rhs*-encoding locus as necessary for self versus non-self recognition in the Gram-negative bacterium *Proteus mirabilis* BB2000 called *idr* [8]. This is a T6SS-

associated locus, and the IdrA, IdrB, and IdrD proteins require a functional T6SS for export out of the cell [8, 9]. IdrD is an Rhs family protein, which an example of a polymorphic toxin. The corresponding immunity gene is *idrE*, which is encoded downstream of *idrD* [8]. We characterized the activity of IdrD-CTD, and a homolog found in *Rothia aeria* C6B, and found that it has deoxyribonuclease activity and causes death *in vivo* (Chapter 2). Further, we combined our biochemical characterization of IdrD-CTD with metagenomic analyses to look at abundance of this toxin in different microbial communities (Chapter 2). In this study, we address the mechanism by which IdrE counteracts IdrD-CTD-mediated cell death and probe metagenomic datasets for the abundance of *idrE* in different communities, as well as in relation to *idrD-CTD*.

Results

IdrE counteracts IdrD-CTD-mediated toxicity and affects detection of IdrD-CTD-FLAG by an unknown mechanism

We predicted the gene immediately downstream, *idrE*, functions as the immunity gene to *idrD*, similar to other polymorphic toxin systems. Overexpression of the toxin IdrD-CTD causes cell death in swarmer cells of *P. mirabilis* BB2000 with a transposon mutation disrupting *idrD-G* (*idrD**) (Chapter 2). To determine whether IdrE acts as the cognate immunity protein, we introduced anhydrotetracycline-inducible, multi-copy expression vectors for IdrD-CTD, IdrE and the two proteins co-produced (IdrD-CTD-E) into *Escherichia coli* strain MG1655. Induction of IdrD-CTD overexpression inhibited growth as compared to *E. coli* carrying a control vector (Figure 3.1A). Growth inhibition did not occur when IdrD-CTD was co-expressed with IdrE, or when IdrE was expressed alone. Similar results are observed in liquid-grown and swarmer *P. mirabilis* BB2000 *idrD** cells (Figure A.7). These results confirm that IdrE counteracts the DNase activity of IdrD-CTD, providing cells with protection.

We previously found that when visualized via an anti–FLAG western blot, the signal for IdrD-CTD-FLAG (active and inactive forms) decreased in the presence of IdrE-His (Figure A.8). To further test the loss of IdrD-CTD-FLAG signal, we performed a time course where whole cell extracts of *E. coli* BL21(DE3)pLysS expressing $IdrD_{D39A}$-CTD-FLAG (inactivated toxin) with and without IdrE-His were collected over the course of two hours and visualized on an anti-FLAG western blot to detect levels of $IdrD_{D39A}$-CTD-FLAG (Chapter 2). $IdrD_{D39A}$-CTD-FLAG signal when expressed alone started to be detectable at 30 minutes (Figure 3.1B). In the presence of IdrE-His, $IdrD_{D39A}$-CTD-FLAG is not always detectable via western blot. (Figure 3.1B). IdrE-His signal is inconsistent via western blot (Figure A.9). We hypothesize that it is an unstable protein, but this inconsistent signal could also problems with the α–His antibody. These experiments indicate that co-expression of IdrD-CTD and IdrE affects detectable levels of IdrD-CTD; however, the mechanism is unclear.

The recently discovered type VI toxin-antitoxin (TA) system contains a protein antitoxin, which neutralizes the activity of a protein toxin by promoting toxin degradation by acting as an adaptor to the Clp protease system [10]. The only example of this mechanism was shown in the *socAB* TA locus in *Caulobacter crescentus* [10]. We set out to test if IdrE targeted $IdrD_{D39A}$-CTD for degradation by acting as an adaptor to a protease system by observing the stability of the $IdrD_{D39A}$-CTD and IdrE complex in protease-deficient *E. coli* strains. $IdrD_{D39A}$-CTD-FLAG signal in the presence of IdrE-His6x was observed in the following strains of *E. coli*: MG1655 (all proteases), BL21(DE3)pLysS (lacking *lon* and *omp-T* proteases) and *E. coli* W3110 *clpP::cat ΔsmpB-1* (lacking *clpP*) (Figure 3.1C). The latter strain was provided by Dr. Tania Baker's lab at MIT [11]. In all strains, $IdrD_{D39A}$-CTD-FLAG signal is decreased in the presence of IdrE-His (Figure 3.1C). These protease-deficient strains did not restore $IdrD_{D39A}$-CTD-FLAG

signal to the same level as IdrD-CTD$_{D39A}$-FLAG in the absence of IdrE-His6x. This data shows that deletion of *clp*, *lon*, and *omp-T* proteases are not sufficient to restore IdrD-CTD-FLAG signal, indicating that IdrE does not function as an adaptor to one of these protease systems.

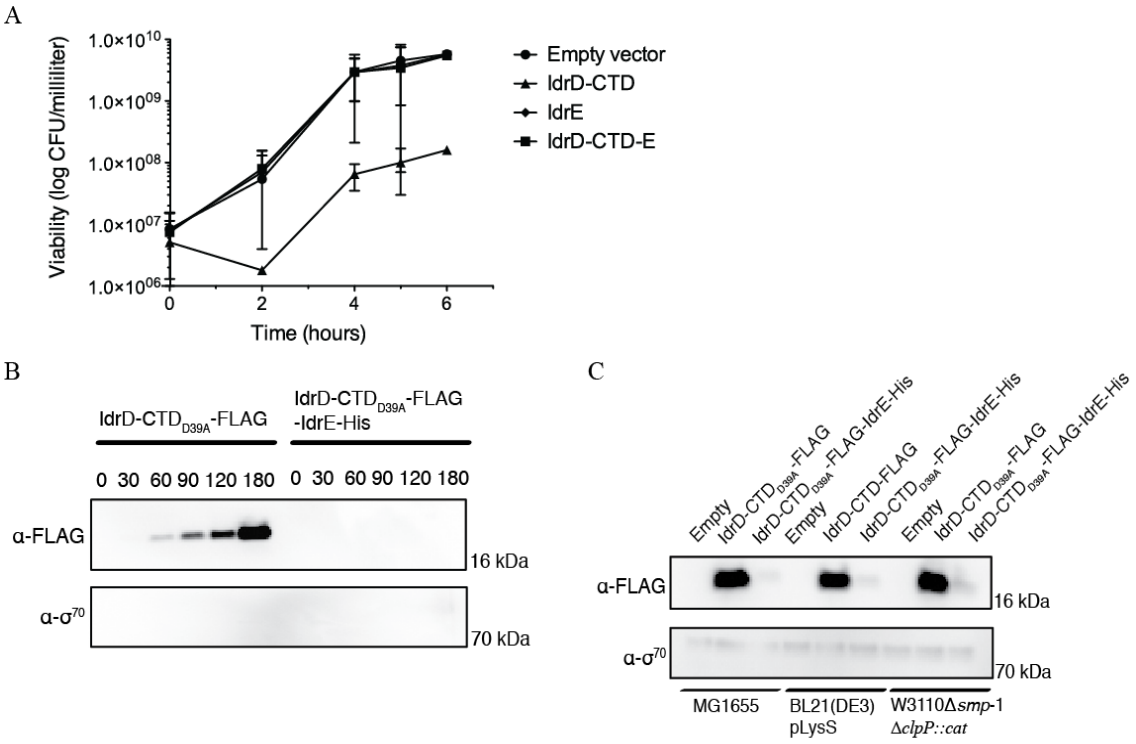


Figure 3.1: Co-expression of IdrE counteracts IdrD-CTD toxicity through an unknown mechanism. (A) Quantification of viable cells after overexpression of IdrD-CTD and IdrE in *E. coli* MG1655. Cells containing an empty vector control or expression vectors of IdrD-CTD, IdrE or both expressed were induced and collected at 0, 2, 4, 5, and 6 hours. Samples were plated for colony-forming units. (B) α-FLAG western blot of $IdrD_{D39A}$-CTD-FLAG-IdrE expression vectors in *E. coli* strain BL21(DE3)pLysS. Whole cell extracts containing expression vectors of $IdrD_{D39A}$-CTD-FLAG and IdrE were collected at the labeled time points and run on an α-FLAG western blot to detect $IdrD_{D39A}$-CTD-FLAG. α-σ^{70} western blot was done as a loading control. Exposure for all blots was 2 minutes. (C) α-FLAG western blot of IdrD-CTD_{D39A}-FLAG-IdrE expression vectors in protease-deficient *E. coli* backgrounds. Whole cell extracts of *E. coli* strains MG1655, BL21(DE3)pLysS, and W3110 Δ*smpB-1* Δ*clpP::cat* containing $IdrD_{D39A}$-CTD-FLAG and IdrE-His expression vectors were collected after 3 hours and run on an α-FLAG western blot to detect $IdrD_{D39A}$-CTD-FLAG. α-σ^{70} western blot was done as loading control.

IdrE homologs reveal potential conserved structural domain at the C-terminal end

Domains of known function have not been found in IdrE based on sequence. The predicted secondary structure of IdrE shows that it is comprised of alpha helices (Figure 3.2A). This is also the case with homologs of IdrE found in other bacteria, which also contain homologs to IdrD-CTD, the cognate toxin (Figure 3.2A; Chapter 2). Intriguingly, alignments of these IdrEs based on predicted secondary structure show a potential conserved set of alpha helices at the 3'-end of the protein, despite low sequence homology (Figure 3.2A).

To determine the necessary domains for IdrE function, we constructed truncations of IdrE. We were able to successfully clone four constructs into a co-expression vector with IdrD-CTD_{D39A}-FLAG (Amino acids 1-85, 1-260, 150-305, and 235-350) (Figure 3.2B). Whole cell extracts were collected from induced BL21(DE3)pLysS cultures and run on a western blot to detect $IdrD_{D39A}$-CTD-FLAG. Three of the four constructs still caused loss of $IdrD_{D39A}$-CTD-FLAG signal: 1-85, 150-305, and 235-305 (Figure 3.2C). Intriguingly, two of these constructs include the structurally conserved 3'-region identified in the IdrE alignments (Figure 3.2A, 3.2B). From these results, truncations of IdrE can cause loss of $IdrD_{D39A}$-CTD-FLAG signal; however further study is required to determine what feature of the IdrE sequence or structure causes loss of IdrD-CTD_{D39A}-FLAG signal.

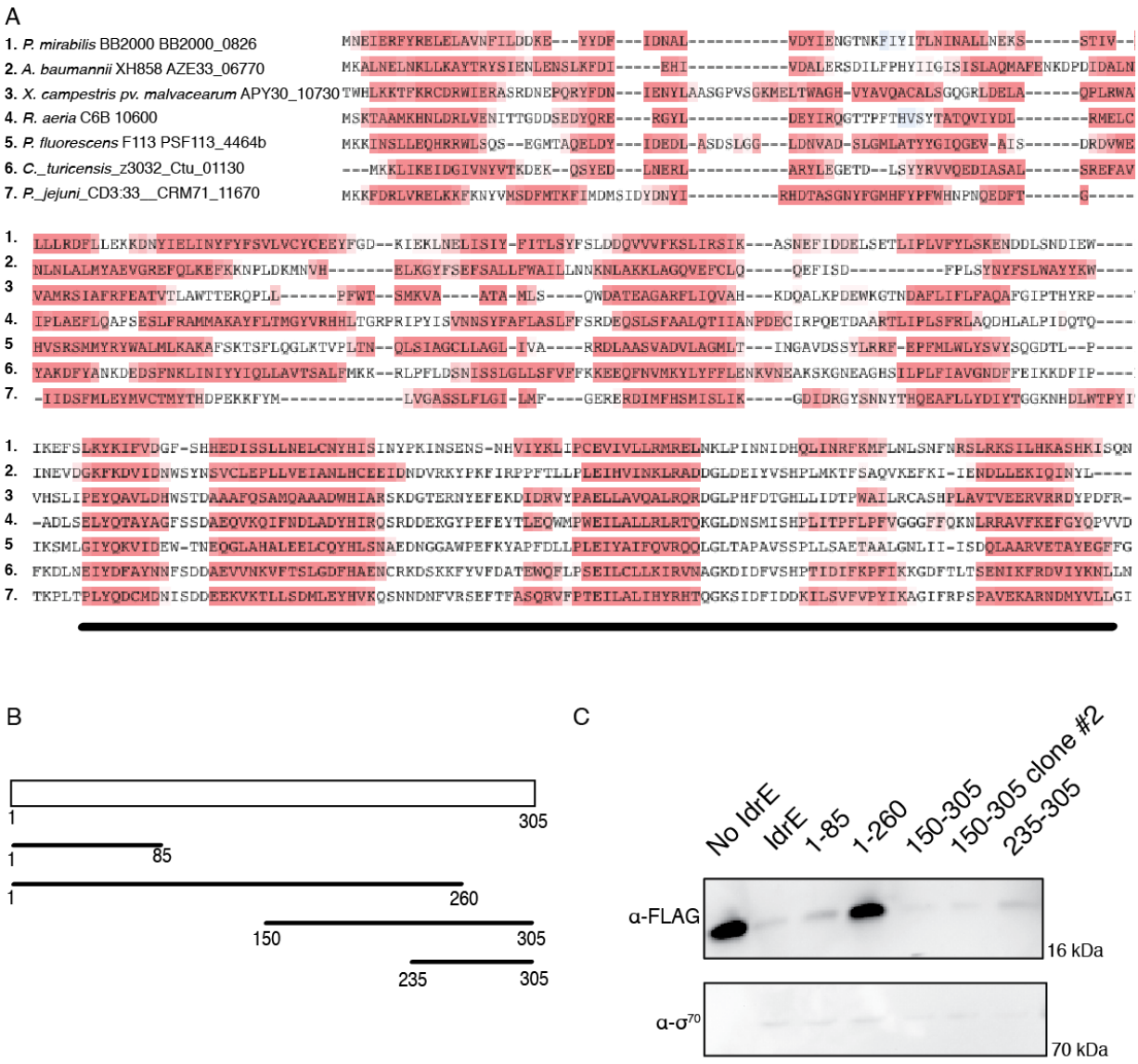

A
1. P. mirabilis BB2000 BB2000_0826
2. A. baumannii XH858 AZE33_06770
3. X. campestris pv. malvacearum APY30_10730
4. R. aeria C6B 10600
5. P. fluorescens F113 PSF113_4464b
6. C._turicensis_z3032_Ctu_01130
7. P._jejuni_CD3:33__CRM71_11670
B
1
305
1
85
1
260
150
305
235
305
C
No IdrE
IdrE
1-85
1-260
150-305
150-305 clone #2
235-305
α-FLAG
16 kDa
α-σ70
70 kDa

Figure 3.2: IdrE homologs reveal similarities in predicted secondary structure (A) Alignment of IdrE-like proteins from representative bacteria containing the cognate toxin. Aligned using MUSCLE on MPI Bioinformatics toolkit. Alpha helices are represented in red; beta sheets are represented in purple. A darker color indicates higher confidence in secondary structure prediction. Black line marks a potential conserved structure. (B) Schematic of IdrE truncations. Various truncations were designed based on predicted secondary structure [12]. Five truncations were successfully cloned: amino acids 1-85, 1-260, 150-305 (two clones), and 235-305. Drawn to scale. (C) α-FLAG western blot of $IdrD_{D39A}$-CTD-FLAG and IdrE truncations co-expression vectors. Whole cell extracts containing expression vectors of $IdrD_{D39A}$-CTD-FLAG and IdrE truncations were collected and run on an α-FLAG western blot to detect $IdrD_{D39A}$-CTD-FLAG. α-σ^{70} western blot was done as loading control. Exposure time for all blots was 5 minutes.

IdrE from R. aeria *C6B has similar mechanism to IdrE from* P. mirabilis *BB2000*

We sought to test if the hypothesized function of IdrE was conserved in another IdrE homolog, based on the conserved secondary structure region that corresponds with the functional truncations in IdrE^{Proteus}. We previously demonstrated that *Rothia aeria* C6B contains a toxic CTD similar to IdrD-CTD that also has DNase activity (Chapter 2). Downstream of this toxin, it also contains a gene similar to *idrE*. IdrE^{Proteus} and its homolog in *R. aeria* share about 25% amino acid identity (Figure 3.3A) [13]. To test if the IdrE^{Rothia} has similar activity to the *P. mirabilis* IdrE, we designed co-expression vectors with IdrD-CTD-FLAG (active and inactive forms) from both *P. mirabilis* and *R. aeria*. Though these co-expression vectors were successfully made, a vector expressing IdrD^{Rothia}-CTD-FLAG (wild type or mutant) could not be cloned. Based on the observation in the *in vitro* DNAse assay that addition of IdrD^{Rothia}-CTD-FLAG results in the complete loss of lambda DNA signal at all concentrations, this inability to clone this sequence is most likely due to IdrD^{Rothia}-CTD-FLAG being more toxic to cells than IdrD^{Proteus}-CTD-FLAG (Chapter 2, Figure 2.2C). Therefore, we moved forward with testing IdrE^{Rothia}-His co-expressed with the IdrD^{Proteus}-CTD-FLAG. Time course assays with these vectors in *E. coli* BL21(DE3)pLysS, as described above, show that co-expression of IdrE^{Rothia}-His also causes a delayed appearance of signal in IdrD$_{D39A}{}^{Proteus}$-CTD-FLAG (Figure 3.3B). This results suggests that IdrE^{Rothia} shares the same activity as IdrE^{Proteus}.

To determine if IdrE^{Rothia} could also counteract *P. mirabilis* IdrD-CTD-mediated cell death, though they are not cognate toxin-immunity pairs, we performed viability assays in *Escherichia coli* strain MG1655 containing expression vectors of IdrD^{Proteus}-CTD-FLAG and IdrE-His (*Proteus* or *Rothia*). IdrD^{Proteus}-CTD-FLAG caused a loss in viability in cells, which was rescued by co-expression with IdrE^{Rothia}-His (Figure 3.3C). Though loss of IdrD^{Proteus}-CTD-

FLAG signal occurs when co-expressed with IdrE^{Rothia}-His, partial rescue from IdrD-CTD-mediated cell death is observed at hours 4 and 5 (Figure 3.3C). These results indicate that IdrE from *P. mirabilis* and *R. aeria* have a similar mechanism, and cross-species protection from a similar non-cognate toxin can occur. However, this protection is partial compared to the full rescue of a cognate toxin, and only occurs at certain time points. This suggests that IdrE homologs could provide temporary protective effects. Further study is required to understand whether this protective effect could influence bacterial interactions in a microbial community.

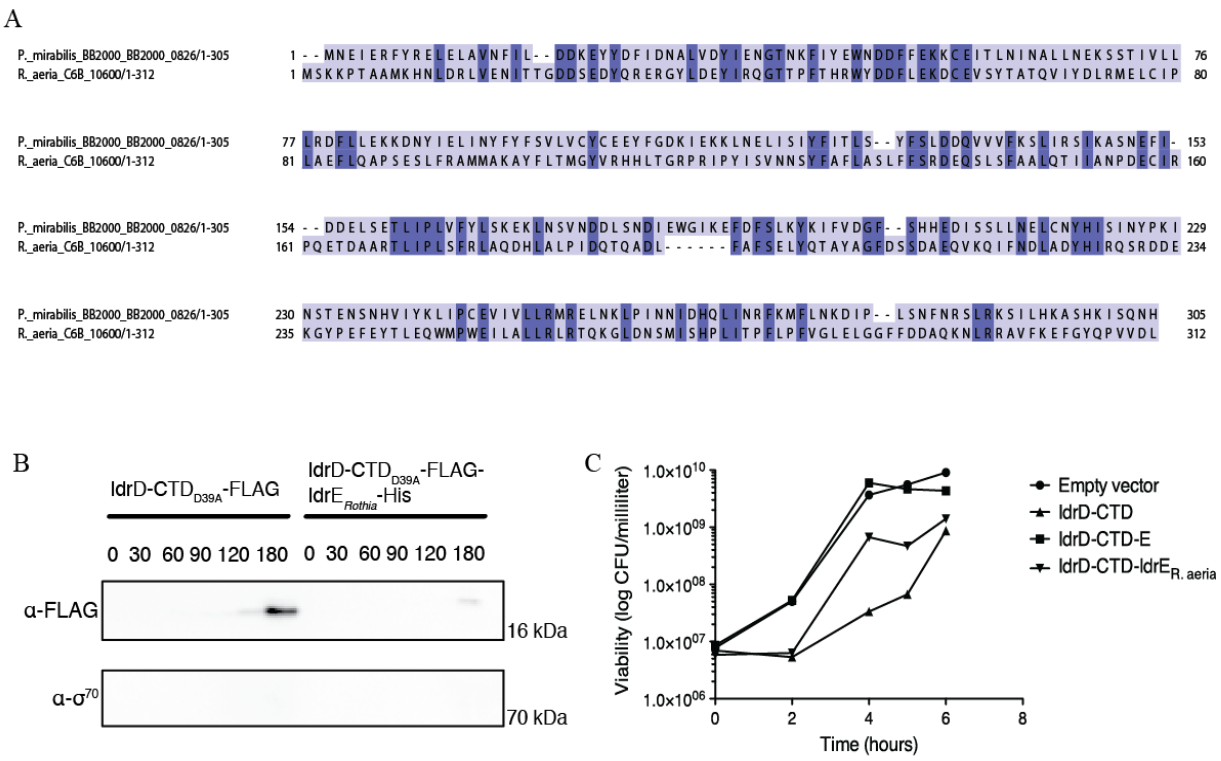


Figure 3.3: IdrE^{Proteus} and IdrE^{Rothia} have similar function. (A) Alignment of IdrE from *P. mirabilis* BB2000 and *R. aeria* C6B. Sequences were aligned using MUSCLE on Jalview. (B) α-FLAG western blot of IdrD$_{D39A}{}^{Proteus}$-CTD-FLAG-IdrE^{Rothia} expression vectors in *E. coli* strain BL21(DE3)pLysS. Whole cell extracts containing expression vectors of IdrD$_{D39A}{}^{Proteus}$-CTD-FLAG and IdrE^{Rothia}-His were collected at the specified time points. IdrD$_{D39A}$-CTD-FLAG was detected by α-FLAG western blot (α-σ^{70} loading control). Exposure for blots was 2 minutes. (C) Quantification of viable cells after overexpression of IdrD^{Proteus}-CTD-FLAG and IdrE^{Rothia}-His in *E. coli* MG1655. Cells containing and empty vector control or expression vectors of IdrD^{Proteus}-CTD-FLAG, IdrE^{Rothia}-His or both expressed were induced and collected at 0, 2, 4, 5, and 6 hours. Samples were plated for colony-forming units.

Metagenomic analysis shows that IdrE co-occurs with IdrD-CTD

In *P. mirabilis* BB2000, *idrE* is found directly downstream of *idrD*. We aimed to look at the abundance of *idrE* in metagenomic datasets, similar to the analysis we did with *idrD* (Chapter 2). We investigated the occurrence of the *idrE* gene in 319 publicly-available human metagenomes (random short-read sequencing of total DNA from a sample) out 1,188 metagenomes identified as likely to contain *idrD* or *idrE* sequences (see Methods, Chapter 2 for details). By mapping each metagenome's short reads onto *idrE* and *idrD* sequences, the abundance of each *idr* sequence in that metagenome's population can be detected.

We found *idrE* sequences from diverse bacteria were abundant in many metagenomes (Figure 3.4A, A.10). However, detection of each *idrE* sequence was generally found in the human microbiome(s) reported to host the parent organism (e.g. *idrE* from *P. mirabilis* was found in human gut but not human oral metagenomes, etc.). Having found that *idrE* is present in the same communities containing *idrD* (Chapter 2), we next wanted to compare the relative abundances of *idrE* and *idrD* in the communities to understand whether one was more abundant than the other.

By comparing the ratio between *idrD* coverage and *idrE* coverage in each metagenome, we discovered that *idrE* is at least as abundant, if not more, than the C-terminal domain (CTD) of *idrD* that contains the active site, but not necessarily more abundant than the entire *idrD* gene (Figure 3.4B, 3.4C). This distinction is likely due to the inclusion of relatively conserved transport domains in other parts of the *idrD* gene, which produced much higher coverage on average than the CTD alone (Chapter 2). The abundance of *idrE* in the metagenomes appears to be directly related to the toxic activity of *idrD-CTD*, consistent with the consensus in the field that an immunity protein protects against a specific toxin [1, 14]. Additionally, some

metagenomes contained *idrE* without *idrD-CTD*, suggesting some selective advantage for having the immunity protein even when the cognate toxin is below our detection limit.

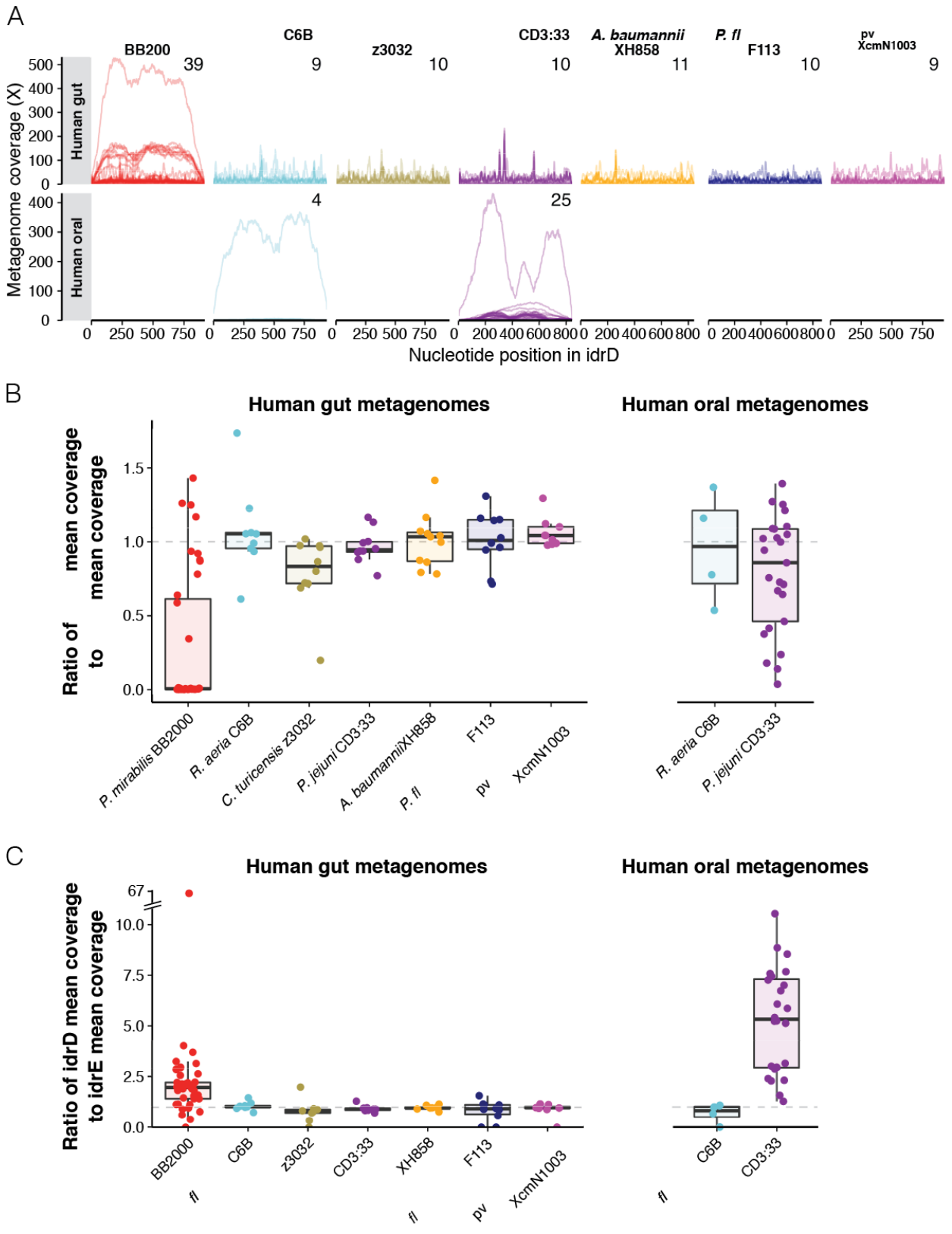

A
BB200
C6B
z3032
CD3:33
A. baumannii XH858
P. fl F113
pv XcmN1003
39
9
10
10
11
10
9
4
25
Human gut
Human oral
Metagenome coverage (X)
Nucleotide position in idrD
B
Human gut metagenomes
Human oral metagenomes
Ratio of mean coverage to mean coverage
P. mirabilis BB2000
R. aeria C6B
C. turicensis z3032
P. jejuni CD3:33
A. baumanniiXH858
P. fl F113
pv XcmN1003
C
Human gut metagenomes
Human oral metagenomes
Ratio of idrD mean coverage to idrE mean coverage
BB2000
C6B
z3032
CD3:33
XH858
F113
XcmN1003

Figure 3.4: *idrE* is abundant in the human microbiome and at least as abundant as *idrD-CTD*. (A) Coverage, the number of metagenomic reads mapping to that position, is plotted for each metagenome (colored, partially transparent lines; each line is a different metagenome). Subpanels show *idrE* sequences from different taxa (subpanel columns) covered by metagenomes obtained from the human gut (top row of subpanels) or human oral cavity (bottom row), for metagenomes covering at least half of that *idrE* sequence. The number of metagenomes plotted is shown in the top right corner of each subpanel. (B) Ratio of mean *idrD* coverage to *idrE* coverage. Each dot represents the ratio from a single metagenome. The boxplots summarise the distribution of the individual ratios. The dotted grey line marks a ratio of 1. Metagenomes are split into same categories (human gut, human oral) as in A. (C) Ratio of mean *idrD-CTD* coverage to mean *idrE* coverage, plotted as in B.

Discussion

We show that co-expression with IdrE causes decreased detection of IdrD-CTD; this activity occurs with both IdrE from *P. mirabilis* and from *R. aeria*. We hypothesize that IdrE promotes the degradation of IdrD-CTD to rescue cells from its DNase activity, though the mechanism is not clear. To further study this mechanism, performing a pulse-chase experiment would provide degradation rates if IdrD-CTD, IdrE, or both are being degraded. Performing this experiment in backgrounds where different proteases are deleted could identify proteases involved based on differences in degradation rates. Loss of IdrD-CTD-FLAG signal still occurs in *E. coli* strains lacking *clp*, *lon*, and *omp-T* proteases indicating that IdrE does not act as an adaptor for these proteases, as we first hypothesized. We have not eliminated the possibility that IdrE could be an adaptor for another protease system not included in this experiment. In the type I toxin-antitoxin systems, the antitoxins are small antisense RNAs that base-pair with the toxin mRNA to inhibit its translation [15]. An alternative hypothesis is that the RNA product of IdrE rescues toxicity from IdrD-CTD. This could be tested by combining lysates containing IdrD-CTD and IdrE subjected to RNase treatment and observing if loss of IdrD-CTD signal still occurs. Alternatively, the formation of a IdrD-CTD and IdrE complex could render the proteins more susceptible to protease degradation in a non-specific manner, though a mechanism such as this has not been previously described.

The metagenomic analysis demonstrates that *idrE* is found in the same communities as its cognate *idrD-CTD* (Fig 3.4). The link between the cognate IdrD^{Proteus}-CTD and IdrE^{Proteus} is supported by the result that IdrE^{Rothia} does not fully rescue IdrD^{Proteus}-CTD. Comparisons of *idrD-CTD* to *idrE* coverage showed that *idrE* can be present in the absence of *idrD-CTD*. This is suggestive of a selective advantage for having the immunity protein, even in the absence of a

functional toxin. The question remains whether an immunity protein that targets its toxin for degradation has an advantage in a certain environment.

Altogether, this study identifies an immunity protein that has a novel mechanism of protecting cells against its toxin. Though the mechanism is not fully elucidated, we have shown that it appears to involve the promotion of toxin degradation and may be dependent on a conserved structural domain in the C-terminal end. Further studies are required to determine the exact mechanism and whether there is an advantage compared to other immunity proteins that only bind the toxin to neutralize their activities.

Materials and Methods

Bacterial strains and media

Strains and plasmids used in this study are described in Table 3.1. Overnight cultures were grown in LB broth (described in Belas et al., 1991) under aerobic conditions at 37°C. All *E. coli* strains were plated on LB agar surfaces (1.5% Bacto agar). When present, antibiotics were used at the following concentration in the media: 35mg/mL kanamycin, 50mg/mL chloramphenicol, and 100mg/mL carbenicillin. For protein production overnights of *E. coli* strains containing anhydrotetracycline-inducible promoters were diluted 1:100 in LB broth with appropriate antibiotic, induced with anhydrotetracycline (10nM) and grown under aerobic conditions for three hours at 37°C.

Plasmid and strain construction

To construct the $idrE^{BB}$ expression plasmid, 897 bp upstream of the *idrA* gene were PCR amplified using primers AS150 and AS151 and cloned into the NheI and SacI sites of plasmid pKG101[16], generating a P_{idrA}-*gfp* expression vector, pAS1034. Next, we amplified the *idrE* gene from BB2000 using primers AS176 and AS177, and replaced the *gfp* gene with *idrE* using the unique SacI and AgeI sites flanking *gfp* in pAS1034. This construction generated the *idrE* expression vector P_{idrA}-$idrE^{BB}$, pAS1042. P_{idrA}-*idrD*-CTD-*idrE* was constructed by PCR amplifying the *idrE* gene from BB2000 using primers AS232 and AS149 and using SLiCE to recombine *idrE* into AgeI-digested pAS1054, resulting in plasmid pAS1059.

The inducible anhydrotetracycline promoter (Ptet) was introduced into the *idrE* expression vectors by generating a gBlock (gDS0006) of the promoter region with 29 bp overhangs for the plasmid, and using SLiCE to recombine into pAS1042. This resulted in the plasmid pDS0001 (*idrE*). A restriction digest on pAS1059 using SacI and AgeI generated an

idrD-CTD-idrE insert. This insert replaced *idrD-CTD* in pDS0002, resulting in pDS0003, an *idrD-CTD-idrE* expression vector.

A C-terminal FLAG tag (GACTACAAGGACGACGATGACAAG) was added to *idrD* by using SLiCE to recombine the gBlock gDS0023 (FLAG tag with 49 bp overhang of *idrD-CTD* and 52 bp overhang of pDS0002) into pDS0002. This vector is pDS0034. The FLAG-tagged *idrD* active site mutants were generated by replacing *idrD-CTD-FLAG* in pDS0034 with the mutants sequences which are encoded in gBlocks gDS0025—resulting in plasmid pDS0048 (D1482A).

Primers oDS0087 and oDS0088 were used to add a C-terminal hexa histidine-tag (CACCACCACCACCACCAC) to *idrE*, which was cloned into pDS0001 using PsiI and AgeI, yielding pDS0030. Primers oDS0089 and oDS0090 were used to add a C-terminal hexa histidine-tag to *idrE* in pDS0003 using SLiCE (oDS0074 and oDS0075 for backbone), yielding pDS0032 (*idrD-CTD-idrE*-His). C-terminal FLAG tag (gDS0023) was added to *idrD-CTD* in pDS0032 using oDS0101 and oDS0102 through SLiCE (pDS0036). This construct was found to be missing two base pairs in the FLAG tag which was corrected using Quikchange (Agilent Technologies, Santa Clara, CA), generating the finally construct of pDS0043.

A restriction digest was performed on gDS0025 (*idrD-CTD-FLAG* D1482A) with SacI and AclI and inserted into pDS0003, replacing wild type *idrD-CTD* with D1482A mutant (pDS0065). Another restriction digest was performed to insert into pDS0043 generating pDS0067.

pAK0004 and pAK0005 were made by amplifying IdrD-CTD-FLAG (oDS0204 and oDS0221) and IdrE^{Rothia} from gDS0040 (oDS0222 and oDS0223); fragments were combined by SOE PCR and inserted into pDS0034 by restriction digest (SacI and Bsu361).

Plasmids were transformed into OneShot Omnimax2 T1R competent cells (Thermo Fisher Scientific, Waltham, MA) and confirmed by Sanger sequencing (Genewiz) using plasmid-specific primers.

W3110Δ*smpB-1 clpP::cat* c. pJF105 was cured of its plasmid before transforming in IdrD-CTD and IdrE expression vectors [11]. Curing protocol was adapted from methods described in Heery et al., 1989. Overnight cultures of the W3110Δ*smpB-1 clpP::cat* c. pJF105 parent strain were grown at 37°C in 10 mL of LB broth + kanamycin, chloramphenicol, and carbenicillin. Cells were then pelleted and washed three times with sterile, deionized water. After a five-minute incubation on ice, cells were pulsed at 2.5kV. One mL of LB broth was added and cells grew for 20 minutes at 37°C before being diluted and plated on selective (LB+ kanamycin, chloramphenicol, and carbenicillin) and nonselective (LB+ kanamycin and chloramphenicol) media and incubated overnight. In a 96-well plate, select colonies from the nonselective media plates were added to LB broth with selective antibiotics and incubated overnight at 37°C. Colonies that did not grow in the broth were then used to prepare competent cells.

Viability Assay

Overnight cultures of *E. coli* MG1655 carrying each expression vector were normalized to OD_{600}=1. 250uL of these normalized cultures were used to inoculate 25mL LB + kanamycin containing 200nM anhydrotetracycline (aTc) and grown for 6 hours at 37°C, shaking. Samples were taken at the following time points and used for a 10-fold dilution series: 0, 2, 4, 5, and 6 hours. Dilutions were plated on LB+kanamycin agar plates.

Gel Electrophoresis and Western Blotting

At the time of collection, sample buffer, consisting of deionized water, 1M Tris, 10% sodium dodecyl sulfate (SDS), 50% glycerol solution, β-mercaptoethanol (BME), and

bromophenol blue, was added to all protein samples. Protein samples were then boiled for 10 minutes after collection and again before gel loading. Samples were loaded onto 12% Tris Tricine gels and gels were run at 120V until samples passed stacking layer at which point voltage was increased to 140V. Samples were then transferred to nitrocellulose membranes with a voltage of 100V at 4°C for one hour. Membranes were then washed three times in TBST, a solution of Tris (pH = 7.4), NaCl, and TWEEN® 20, for five minutes. To prevent nonspecific antibody binding, membranes rocked in a solution of 5% milk in TBST, consisting of dry non-fat milk powder dissolved in TBST, for 30 minutes. Membranes then soaked in primary antibody solutions of concentration 1:4000 (rabbit α-FLAG, mouse α-His6x) or 1:1000 (mouse α-σ^{70}) in 5% milk for either 1 hour at room temperature or overnight at 4°C. After another round of TBST washes, membranes rocked in 1:5000 secondary antibody solutions (goat α-rabbit, goat α-mouse) in 5% milk for 30 minutes at room temperature. The TBST washes were repeated a final time and Immun-Star HRP luminol/enhancer (Bio-Rad) was applied right before chemiluminescence exposure using the ChemiDoc XRS system (Bio-Rad).

IdrD and IdrE Time Course

E. coli BL21(DE3)pLysS carrying the $IdrD_{D39A}$-CTD-FLAG vector and $IdrD_{D39A}$-CTD-FLAG and IdrE-His6x co-expression vector were induced as described above. Culture samples (100ul) were taken every thirty minutes until 2 hours, and one sample taken at 3 hours. Sample buffer (50ul) was added to each sample, which were then run on α-FLAG, α-σ^{70}, and α-His western blots.

Metagenomic Analyses

The 1,188 metagenomes identified as containing idrD sequences from (Chapter 2) were used to investigate the abundance and prevalence of idrE. Briefly, publically-available

metagenomes were screened using the SearchSRA portal (https://www.searchsra.org/) and downloaded. A set of seven idrE sequences, one each from Acinetobacter baumannii XH858, Cronobacter turicensis z3032, Prevotella jejuni CD3:33, Proteus mirabilis BB2000, Pseudomonas fluorescens F113, Rothia aeria C6B, and Xanthomonas citri pv. malvacearum XcmN1003, were used to create a bowtie2 reference database onto which the metagenomic short reads were then mapped with bowtie2 with default parameters [17]. The resulting data was then managed with anvi'o, a platform for analysis and visualization of 'omics data [18]. An anvi'o contigs database was generated with the seven idrE sequences and profiled with the merged results of the bowtie2 mapping (anvi-gen-contigs-database, anvi-profile, and anvi-merge commands, respectively). Per-nucleotide coverage and variability information were then exported with the anvi-get-split-coverages and anvi-gen-variability-profile commands, respectively, for further analyses.

The abundance of idrE in metagenomes was investigated using a custom R script based off of (Chapter 2). We focused only on the 319 metagenomes originating from the human gut or oral microbiomes (Chapter 2). Additionally, for Figure 3.4A only, we dropped outlier metagenomes that contained >1,000x coverage (n = 22). For all analyses, metagenomes were filtered for relevance to each sequence by requiring each metagenome to cover at least half of that sequence's nucleotide positions; the number of metagenomes passing this criterion is shown in the top right of each subpanel in Figure 3.4A. Each individual metagenome's coverage is then plotted as a semi-transparent trace. The ratio plots (Figure 3.4B, 3.4C) were generated by taking mean coverage of each gene by each relevance-filtered metagenome (criterion applied after subsetting to CTD region for Figure 3.4B) and dividing the mean *idrD* coverage by the mean *idrE* coverage. The points are the raw ratios plotted over the boxplots summarising their

distribution.

Table 3.1: Strains used in this study

Strain	Notes	KAG #	DS, EK or AK #	Source
E. coli MG1655 + pBBR1-NheI	MG1655 carrying empty vector (pBBR1 origin, Kan (R))	2076	DS0068	Chapter 2
E. coli MG1655 + pDS0001	MG1655 producing IdrE under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2297	DS0150	This study
E. coli MG1655 + pDS0002	MG1655 producing IdrD-CTD under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2298	DS0151	Chapter 2
E. coli MG1655 + pDS0003	MG1655 co-producing IdrD-CTD and IdrE under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2341	DS0170	This study
E. coli MG1655 + pDS0034	MG1655 producing IdrD-CTD-FLAG under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2428	DS0191	This study
E. coli MG1655 + pDS0043	MG1655 co-producing IdrD-CTD-FLAG and IdrE-His under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2485	DS0228	This study

Table 3.1 (continued): Strains used in this study

Strain	Notes	KAG #	DS, EK or AK #	Source
E. coli MG1655 + pDS0048	MG1655 producing IdrD-CTD_{D39A}-FLAG under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2741	DS0248	This study
E. coli MG1655 + pDS0067	MG1655 co-producing IdrD-CTD_{D39A}-FLAG and IdrE-His under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3374	DS0366	This study
E. coli MG1655 + pAK0004	MG1655 co-producing IdrD-CTD-FLAG and $IdrE_{Rothia}$-His under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	4369	DS0411	This study
E. coli BL21(DE3)pLysS + pBBR1-NheI	BL21(DE3)pLysS carrying empty vector (pBBR1 origin, Kan (R))	3007	DS0297	This study
E. coli BL21(DE3)pLysS + pDS0048	BL21(DE3)pLysS co-producing IdrD-CTD_{D39A}-FLAG under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	2918	DS0284	This study

Table 3.1 (continued): Strains used in this study

Strain	Notes	KAG #	DS, EK or AK #	Source
E. coli BL21(DE3)pLysS + pDS0067	BL21(DE3)pLysS co-producing IdrD-CTD_{D39A}-FLAG and IdrE-His under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	4367	DS0409	This study
E. coli BL21(DE3)pLysS + pAK0005	BL21(DE3)pLysS co-producing IdrD-CTD_{D39A}-FLAG and $IdrE_{Rothia}$-His under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	4323	AK0007	This study
E. coli W3110*ΔsmpB-1 clpP::cat* + pBBR1-NheI	W3110*ΔsmpB-1 clpP::cat* carrying empty vector (pBBR1 origin, Kan (R))	4368	DS0410	This study
E. coli W3110*ΔsmpB-1 clpP::cat* + pDS0048	W3110*ΔsmpB-1 clpP::cat* co-producing IdrD-CTD_{D39A}-FLAG under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3627	EK0029	This study
E. coli W3110*ΔsmpB-1 clpP::cat* + pDS0067	W3110*ΔsmpB-1 clpP::cat* co-producing IdrD-CTD_{D39A}-FLAG and IdrE-His under the control of an anhydrotetracyline-inducible promoter (pBBR1 origin, Kan (R))	3630	EK0031	This study

Table 3.1 (continued): Strains used in this study

Strain	Notes	KAG #	DS, EK or AK #	Source
OneShot Omnimax 2 T1R Competent Cells	*E. coli* strain for cloning			Thermo Fisher Scientific, Waltham, MA.
E. coli str. K12 substr. MG1655	Wild-type laboratory strain of *E. coli*			[19]
OneShot BL21(DE3) pLysS Competent Cells	Strain for protein overexpression.			Thermo Fisher Scientific, Waltham, MA.
E. coli W3110*ΔsmpB-1 clpP::cat*	*E. coli* W3110 with deletions of *smpB* (ssrA tagging) and *clpP* (protease subunit)			[11]

Table 3.2: Plasmids used in this study

Plasmid	Cloning method (or source)	Primers (5'=> 3')
pDS0001	Anhydrotetracycline promoter (Ptet) with 29bp overhangs (gDS0006) was recombined into amplified pAS1042 by SLiCE	oDS0005: gctagccatttgcccatgg oDS0006: cgtttttgataaaaggatattgttgag gDS0006: tcgcccacccccatgggcaaatggctagcttaagacccactttcacatttaagttgtt tttctaatccgcatatgatcaattcaaggccgaataagaaggctggctctgcaccttg gtgatcaaataattcgatagcttgtcgtaataatggcggcatactatcagtagtaggt gtttccctttcttctttagcgacttgatgctcttgatcttccaatacgcaacctaaagtaa aatgccccacagcgctgagtgcatataatgcattctctagtgaaaaaccttgttggc ataaaaaggctaattgattttcgagagtttcatactgtttttctgtaggccgtgtaccta aatgtacttttgctccatcgcgatgacttagtaaagcacatctaaaactttagcgttat tacgtaaaaaatcttgccagctttccccttctaaagggcaaaagtgagtatggtgcct atctaacatctcaatggctaaggcgtcgagcaaagcccgcttatttttacatgccaa tacaatgtaggctgctctacacctagcttctgggcgagtttacgggttgttaaaccttc gattccgacctcattaagcagctctaatgcgctgttaatcactttacttttatctaatcta gacatcattaattcctaatttttgttgacactctatcgttgatagagttatttaccactcc ctatcagtgatagagaaagtttttgataaaaggatattgttgagctc
pDS0003	Restriction digest of *idrD-CTD-idrE* from pAS1059 ligated into pDS0002	
pDS0034	SLiCE of FLAG tag into pDS0002	gDS0023: tggccgaaatcccaaattttatgatcagaatataacgtttgaggaatgggactacaa ggacgacgatgacaagtagaccggtttattgactaccggaagcagtgtgaccgtg tgcttctcaaatg
pDS0043	His6x was amplified and inserted into pDS0003 via SLiCE (oDS0074 and oDS0075 for backbone). C-terminal FLAG tag was amplified and added to *idrD-CTD* in pDS0032 through SLiCE. Quikchange used to correct missing two base pairs in the FLAG tag	oDS0074: gtgattctgactaattttatgtgatgc oDS0075: tagaccggtttattgactaccgg oDS0089: aagcatcacataaaattagtcag oDS0090: ctgcttccggtagtcaataaac oDS0101: ccattcctcaaacgttatattc oDS0102: taggatgaatgagatagaacg oDS0145: aacgtttgaggaatggtagaccggtttattgac oDS0146: gtcaataaaccggtctaccattcctcaaacgtt

Table 3.2 (continued): Plasmids used in this study

Plasmid	Cloning method (or source)	Primers (5'=> 3')
pDS0048	Restriction digest (SacI and AgeI) of gDS0025 into pDS0002	gDS0025: gtcaaggagctctcatgtgctttagtgctcgggtaggtgcttttggtgagaaaagag ttatgaaatacttatctggagcgggctataaaaaagtttttctgtacaaaacaattctg ggcatggtctggctatagttgctttaagaccagatggaaaatttgatatttttgaagtta aaagttcgacaataggacaattttctttatcttcccgccaagctacaggcgatgatttt gcaaaaatagttcttttaaacgatgtgaaaaaaggaggttataatattatcgatataga tggtaatgttaaagcaattacaagtaaacaagctagatacatttataataacatagga acaaccgagtgggttcaggtaaatgttggccgaaatcccaaattttatgatcagaat ataacgtttgaggaatgggactacaaggacgacgatgacaagtagaccggtgtca ag
pDS0067	Restriction digest (SacI and AclI) of pDS0065 inserted into pDS0043	

Table 3.2 (continued): Plasmids used in this study

Plasmid	Cloning method (or source)	Primers (5'=> 3')
pAK0004	Amplified IdrD-CTD-FLAG and gDS0040 and combined inserts by SOE PCR. Restriction digest of insert with SacI and Bsu361 for ligation into anhydrotetracycline-inducible vector	oDS0204: gtacatgagctctcatgtgctttagtgctcg oDS0221: cggtgggtttttgctcatcctacttgtcatcgtcgtccttgtagtc oDS0222: ggacgacgatgacaagtaggatgagcaaaaaacccaccgcc oDS0223: aactggcctcaggttagtgatggtgatgatggtgtaaatcaacaacagg gDS0040: atgagcaaaaaacccaccgccgccatgaagcacaatcttgaccgtctggttgaaa acattactaccggtgacgacagcgaagactatcaacgcgaacgcggctatctgga cgaatacatccgccaaggcaccaccccgttcacccaccgctggtatgacgacttttt tggagaaggactgtgaagtcagctacaccgccacccaagtcatctacgatctgcg catggaactgtgcatcccgcttgccgaatttctgcaagctccaagcgaaagcctgtt ccgcgccatgatggcaaaagcctatttccttaccatgggctatgtacgccaccacct caccggcagaccgcgcattccctatatcagcgtaaacaatagctacttcgccttcct cgcctccctctttttcagccgcgacgagcagagcctcagcttcgccgccctgcaaa ccatcattgccaacccagacgaatgcatccgaccgcaggaaaccgatgccgccc gcaccctcatccccctgtcgttccgcctcgcccaagaccatcttgccctgccgattg accaaacgcaggcagatttgttcgccttctccgaactctaccaaaccgcctacgca ggctttgacagcagcgacgccgaacaggtgaagcagattttcaacgacctcgccg actaccatatccggcaaagccgcgacgatgaaaaaggttaccccgaattcgaata cacgttggaacaatggatgccgtgggaaatcctcgccctgctgcgcctgcgtacg caaaaaggcttggacaacagtatgattagccacccgctgattactccgtttctcccct ttgtcggcttagaactgggcggctttttcgacgacgcgcaaaaaaacctgcgccgc gccgtgttcaaagaatttggttaccaacctgttgttgatttacaccatcatcaccatca ctaa
pAK0005	Same as above, except with $IdrD_{D39A}$-FLAG	Same as above

References

1. Poole, S.J., et al., *Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems.* PLoS Genet, 2011. **7**(8): p. e1002217.

2. Koskiniemi, S., et al., *Rhs proteins from diverse bacteria mediate intercellular competition.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 7032-7.

3. Jackson, A.P., et al., *Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement.* BMC Genomics, 2009. **10**: p. 584.

4. Aoki, S.K., et al., *A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria.* Nature, 2010. **468**(7322): p. 439-42.

5. Zhang, D., L.M. Iyer, and L. Aravind, *A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems.* Nucleic Acids Res, 2011. **39**(11): p. 4532-52.

6. Ma, J., et al., *PAAR-Rhs proteins harbor various C-terminal toxins to diversify the antibacterial pathways of type VI secretion systems.* Environ Microbiol, 2017. **19**(1): p. 345-360.

7. Chen, H., et al., *Identification of a contact-dependent growth inhibition system in the probiotic Escherichia coli Nissle 1917.* FEMS Microbiol Lett, 2018. **365**(11).

8. Wenren, L.M., et al., *Two independent pathways for self-recognition in Proteus mirabilis are linked by type VI-dependent export.* MBio, 2013. **4**(4).

9. Zepeda-Rivera, M.A., C.C. Saak, and K.A. Gibbs, *A Proposed Chaperone of the Bacterial Type VI Secretion System Functions To Constrain a Self-Identity Protein.* J Bacteriol, 2018. **200**(14).

10. Aakre, C.D., et al., *A bacterial toxin inhibits DNA replication elongation through a direct interaction with the β sliding clamp.* Mol Cell, 2013. **52**(5): p. 617-28.

11. Flynn, J.M., et al., *Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals.* Mol Cell, 2003. **11**(3): p. 671-83.

12. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nat Protoc, 2015. **10**(6): p. 845-58.

13. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

14. Zhang, D., et al., *Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics.* Biol Direct, 2012. **7**: p. 18.

15. Thisted, T., et al., *Mechanism of post-segregational killing: Sok antisense RNA interacts with Hok mRNA via its 5'-end single-stranded leader and competes with the 3'-end of Hok mRNA for binding to the mok translational initiation region.* EMBO J, 1994. **13**(8): p. 1960-8.

16. Gibbs, K.A., L.M. Wenren, and E.P. Greenberg, *Identity gene expression in Proteus mirabilis.* J Bacteriol, 2011. **193**(13): p. 3286-92.

17. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

18. Eren, A.M., et al., *Anvi'o: an advanced analysis and visualization platform for 'omics data.* PeerJ, 2015. **3**: p. e1319.

19. Blattner, F.R., et al., *The complete genome sequence of Escherichia coli K-12.* Science, 1997. **277**(5331): p. 1453-62.

Chapter 4

Discussion

Bacteria form complex community structures where cells are in close contact. Contact-dependent interactions, such as polymorphic toxin systems, are often competitive and are hypothesized to play a role in the formation of these complex structures. Polymorphic toxins are defined by a modular structure where the N-terminal region contains components required for secretion and the C-terminal region contains the toxic domain [1-3]. The conserved features of the polymorphic toxin-immunity pairs have helped identify many novel enzymes, which could potentially have biotechnological or therapeutic applications [4]. In this study, we elucidate the molecular function of a polymorphic toxin and its cognate immunity protein from *P. mirabilis* BB2000. We combine these results with metagenomic analyses to address the role of similar toxin-immunity pairs in microbial communities.

In Chapter 2 we show that the mechanism by which IdrD-CTD causes cell death is through DNA-degrading activity of IdrD (Figure 2.1). We have found that IdrD^{Proteus}-CTD, IdrD^{Rothia}-CTD, and related proteins comprise a new protein subfamily of the PD-(D/E)XK superfamily. With the addition of another 23 potential members found through bioinformatics examinations, we will submit these alignments for consideration as a Pfam domain as one does not yet exist for this protein subfamily (Figure A.3). PD-(D/E)XK superfamily members are characterized by a conserved structural fold in the enzymatic core and have sequence and structure variability elsewhere . This variability has made it difficult to identify new subfamilies and to determine the nucleic acid target, because target specificity is predicted to be embodied outside of the catalytic core [5, 6].

We find these IdrD-CTD proteins particularly interesting, because we hypothesize that IdrD-CTD could have a role in establishing and retaining spatial organization in a more complex community such as the gut and oral microbiomes. Polymorphic toxins like IdrD-CTD are

hypothesized to contribute to competition with other bacteria in microbiomes and to differentiation between genetic lineages [1, 3, 7, 8]. We had previously identified *idrD* as a self versus non-self recognition factor that allows one strain of *P. mirabilis* to physically and spatially exclude another [9]. We have also previously shown that the encoded IdrD protein is exported via cell-contact associated machinery and requires physical contact to cause death among neighboring cells [9, 10]. Another polymorphic toxin, a putative nuclease of a different protein family to IdrD, was identified as a fitness factor for polymicrobial catheter-associated urinary tract infections in the clinical isolate *P. mirabilis* HI4320[11, 12]. This indicates that the role for polymorphic nuclease toxins in *P. mirabilis* interactions also apply to a polymicrobial community. Further, we have also identified and characterized a rD-CTD-like protein in the genus *Rothia* (Figure 1.2) which is an abundant, diverse inhabitant of supragingival plaque [13]Costea [14]. The microbiota inhabiting supragingival dental plaque are strikingly organized with stark separation between spatially-adjacent bacterial populations [15]. Consistent with our hypothesis for IdrD-CTD's conserved role in spatial organization, we found that IdrD-CTD^{Proteus} and IdrD-CTD^{Rothia} retain the same enzymatic function though differing in nucleotide sequence and in originating microbiome niche. This hypothesis could be addressed by observing the expression of these proteins within a structured community and the effect of introducing strains lacking these proteins in a community. However, this analysis is prevented by the low abundance of these sequences in isolated microbiomes.

This study exemplifies a way to probe the abundance of polymorphic toxins, as well as their functional subdomains, within different types of communities based on publicly available metagenomes. Molecular analysis of IdrD-CTD was required to understand the differences in abundance of the subdomains in different *Prevotella* species in the oral metagenomes. We show

that by combining biochemical characterization and metagenomic analyses of IdrD-CTD^{Proteus} and similar proteins, we are able to probe for the abundance of not only the gene as a whole, but also of subdomains within the gene. Observing the spread of toxins with similar targets could determine if specific targets provide more of a competitive advantage than other in different microbial communities. One could examine the prevalence and distributions of other polymorphic toxins within metagenomic datasets using this method. Indeed, several metagenomic analyses investigating the abundance of various cell-contact dependent effectors, often originating from a single species, have previously suggested roles in establishing overall dominance in a community and the presence of niche-specific specialization of effectors [16, 17]. Emerging from these studies is the possibility that low-abundance and strain-specific genes could provide a window for understanding differentiation among groups, both in species and strains. Rapid and large-scale metagenomics analysis combined with molecular characterization of factors necessary for cell-cell communication and cell-cell interactions opens the possibilities of better understanding how interactions among resident microbes correlate with, and possibly contribute to, behaviors within a host.

We hypothesize that IdrD-CTD is a binding partner of IdrE based on other bacterial toxin-immunity pairs. Similar to other studies, we attempted to perform co-immunoprecipitations of epitope-tagged IdrD-CTD and IdrE. This experiment was unsuccessful because IdrE was not detectable on western blots. We also observed that IdrD-CTD signal was decreased in the presence of IdrE, which was confirmed in a time-course experiment of IdrD-CTD-FLAG signal (Figure 3.1). We hypothesize that IdrE is promoting degradation of IdrD-CTD to protect cells against toxicity, but have not identified a mechanism.

A future candidate protease for eludicidating IdrE mechanism is the HslVU (ClpQ) protease. HslV, the peptidase subunit, has a similar catalytic mechanism to the eukaryotic and archaebacterial 20S proteasomes [18]. Additionally, homologs of the components of this protease are present in both *E. coli* MG1655 (*hslU* and *hslV*) and *P. mirabilis* BB2000 (BB2000_3234 and BB2000_3235). This protease is a candidate because accessory factors in both bacterial and eukaryotic systems have been found to require a conserved C-terminal end to assemble and activate their respective proteases [19, 20]. In the case of the eukaryotic Rpn12 subunit, a single α-helix results in a conformational change to promote assembly of the proteasome lid, which leads to assembly of full proteasome complex [20]. The predicted secondary structure of IdrE homologs shows that it is comprised of α-helices, and align at the C-terminal end, which has also been functionally shown to decrease IdrD-CTD signal (Figure 3.2). Given these characteristics, perhaps IdrE functions as a chaperone to HslVU, which would be a novel immunity protein mechanism. This could be tested by observing if IdrD-CTD-FLAG signal when co-expressed with IdrE-His is similar to IdrD-CTD-FLAG alone in a strain lacking components of the HslVU protease. Obtaining stable amounts of IdrE and subsequent purification would allow for determination of structure and interactions with binding partners; however, we have found that IdrE is unstable so production and detection would have to be optimized before such studies.

The link between IdrD-CTD and IdrE is supported by the metagenomic analysis, which shows that *idrE* is found in the same communities as its cognate *idrD-CTD* (Figure 3.4). Comparisons of *idrD-CTD* to *idrE* coverage showed that *idrE* can be present in the absence of *idrD-CTD*. This is suggestive of a selective advantage for having the immunity protein, even in the absence of a functional toxin. Many polymorphic toxin systems contain orphan toxin-

immunity pairs, which consist of a displaced C-terminal toxin, not attached to the N-terminal end required for transport, and an intact immunity gene that has the components to be expressed [21]. It has been reported that recombination can occur that restores orphan toxins to the N-terminal domain resulting in expression of the former orphan pair [22]. Because IdrE^{Rothia} provides protection against the non-cognate IdrD^{Proteus}-CTD and certain time points, we hypothesize that instances where an immunity gene is present without a functional toxin could be a mechanism to protect against similar but non-cognate toxins. This hypothesis could be tested in the context of cell-mediated competition by competing a strain with an intact toxin-immunity pair and one with a homologous orphan toxin-immunity pair. If this hypothesis were correct, it would shift the current paradigm that only a cognate immunity protein can protect against a toxin.

Reference

1. Aoki, S.K., et al., *A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria.* Nature, 2010. **468**(7322): p. 439-42.

2. Jackson, A.P., et al., *Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement.* BMC Genomics, 2009. **10**: p. 584.

3. Zhang, D., L.M. Iyer, and L. Aravind, *A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems.* Nucleic Acids Res, 2011. **39**(11): p. 4532-52.

4. Jamet, A. and X. Nassif, *New players in the toxin field: polymorphic toxin systems in bacteria.* MBio, 2015. **6**(3): p. e00285-15.

5. Steczkiewicz, K., et al., *Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily.* Nucleic Acids Res, 2012. **40**(15): p. 7016-45.

6. Kosinski, J., M. Feder, and J.M. Bujnicki, *The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function.* BMC Bioinformatics, 2005. **6**: p. 172.

7. Aoki, S.K., et al., *Contact-dependent inhibition of growth in Escherichia coli.* Science, 2005. **309**(5738): p. 1245-8.

8. Koskiniemi, S., et al., *Rhs proteins from diverse bacteria mediate intercellular competition.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 7032-7.

9. Wenren, L.M., et al., *Two independent pathways for self-recognition in Proteus mirabilis are linked by type VI-dependent export.* MBio, 2013. **4**(4).

10. Zepeda-Rivera, M.A., C.C. Saak, and K.A. Gibbs, *A Proposed Chaperone of the Bacterial Type VI Secretion System Functions To Constrain a Self-Identity Protein.* J Bacteriol, 2018. **200**(14).

11. Alteri, C.J., et al., *Subtle variation within conserved effector operon gene products contributes to T6SS-mediated killing and immunity.* PLoS Pathog, 2017. **13**(11): p. e1006729.

12. Armbruster, C.E., et al., *Genome-wide transposon mutagenesis of Proteus mirabilis: Essential genes, fitness factors for catheter-associated urinary tract infection, and the impact of polymicrobial infection on fitness requirements.* PLoS Pathog, 2017. **13**(6): p. e1006434.

13. Lloyd-Price, J., et al., *Strains, functions and dynamics in the expanded Human Microbiome Project.* Nature, 2017. **550**(7674): p. 61-66.

14. Costea, P.I., et al., *Subspecies in the global human gut microbiome.* Mol Syst Biol, 2017. **13**(12): p. 960.

15. 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. E791-800.

16. Verster, A.J., et al., *The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition.* Cell Host Microbe, 2017. **22**(3): p. 411-419.e4.

17. Egan, F., F.J. Reen, and F. O'Gara, *Tle distribution and diversity in metagenomic datasets reveal niche specialization.* Environ Microbiol Rep, 2015. **7**(2): p. 194-203.

18. Bochtler, M., et al., *Crystal structure of heat shock locus V (HslV) from Escherichia coli.* Proc Natl Acad Sci U S A, 1997. **94**(12): p. 6070-4.

19. Hu, K., et al., *Proteasome substrate capture and gate opening by the accessory factor PafE from.* J Biol Chem, 2018. **293**(13): p. 4713-4723.

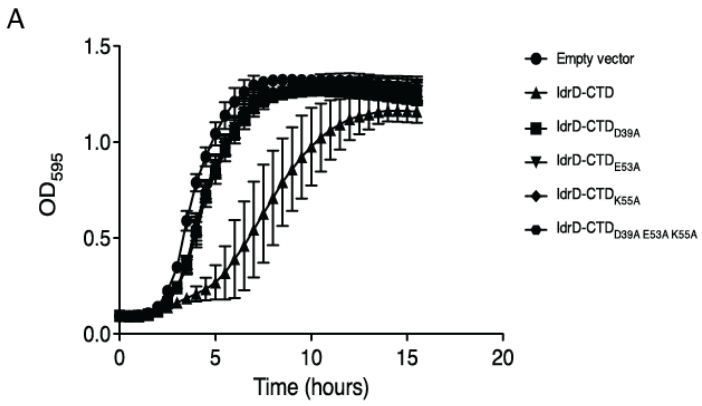
20. Tomko, R.J., et al., *A Single α Helix Drives Extensive Remodeling of the Proteasome Lid and Completion of Regulatory Particle Assembly.* Cell, 2015. **163**(2): p. 432-44.

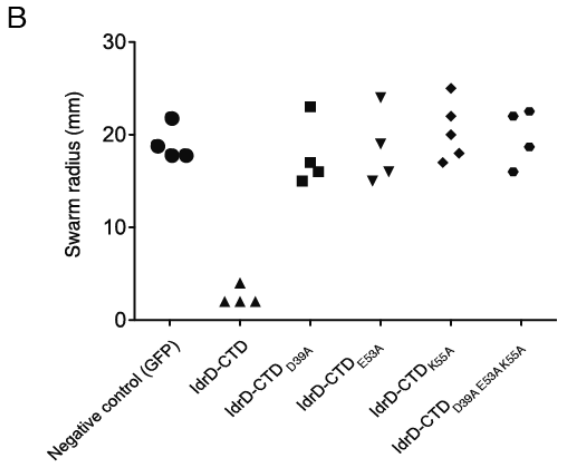
21. Poole, S.J., et al., *Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems.* PLoS Genet, 2011. **7**(8): p. e1002217.

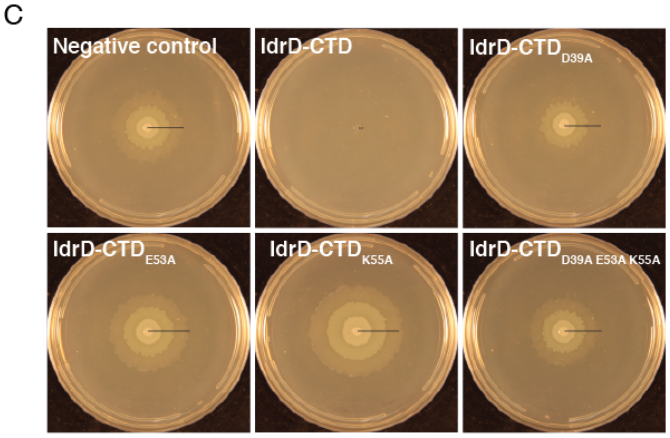
22. Koskiniemi, S., et al., *Selection of orphan Rhs toxin expression in evolved Salmonella enterica serovar Typhimurium.* PLoS Genet, 2014. **10**(3): p. e1004255.

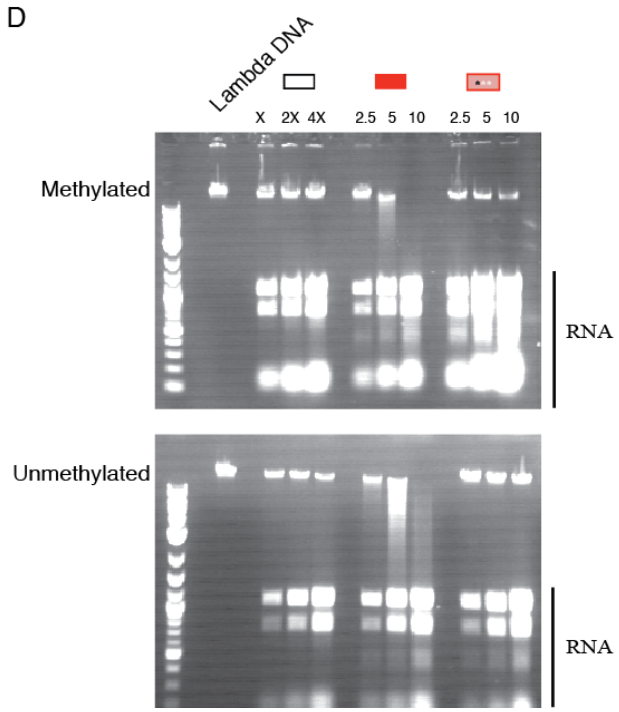
Appendix A

Supplemental Figures and Tables


A
1.5
1.0
0.5
0.0
OD595
0
5
10
15
20
Time (hours)
Empty vector
IdrD-CTD
IdrD-CTD D39A
IdrD-CTD E53A
IdrD-CTD K55A
IdrD-CTD D39A E53A K55A


B
30
20
10
0
Swarm radius (mm)
Negative control (GFP)
IdrD-CTD
IdrD-CTD D39A
IdrD-CTD E53A
IdrD-CTD K55A
IdrD-CTD D39A E53A K55A


C
Negative control
IdrD-CTD
IdrD-CTD D39A
IdrD-CTD E53A
IdrD-CTD K55A
IdrD-CTD D39A E53A K55A


D
Lambda DNA
X 2X 4X 2.5 5 10 2.5 5 10
Methylated
RNA
Unmethylated
RNA

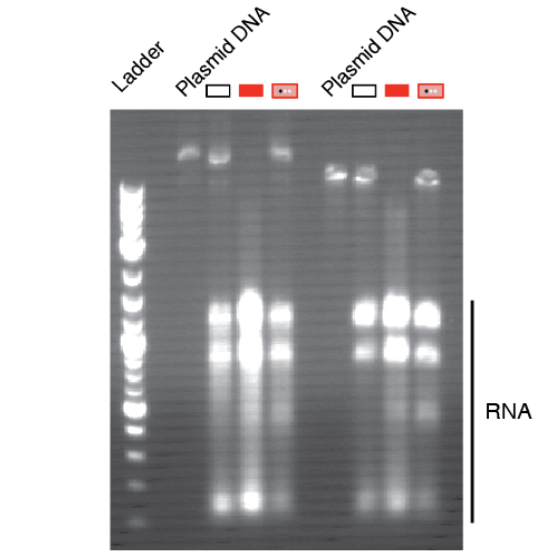

E
Ladder
Plasmid DNA
Plasmid DNA
RNA

Figure A.1: *P. mirabilis* IdrD-CTD is a novel DNAse in the PD-(D/E)XK superfamily. (A) Quantification of viable cells after overexpression of IdrD-CTD and active site mutants in liquid-grown *E. coli* MG1655. Optical density (OD_{600}) was measured over the course of sixteen hours. (B) Swarm distances of *P. mirabilis* BB2000 *idrD** swarmer cells containing expression vectors of IdrD-CTD and mutants (C) Representative images of swarm plates with swarm radius indicated by black line (D and E) Full agarose gels of DNAse assays with (D) plasmid DNA and (E) lambda DNA as substrates, with NEB 2-log DNA ladder. Bands running below 1kB are presumed to be rRNA and tRNA from PURExpress reaction.

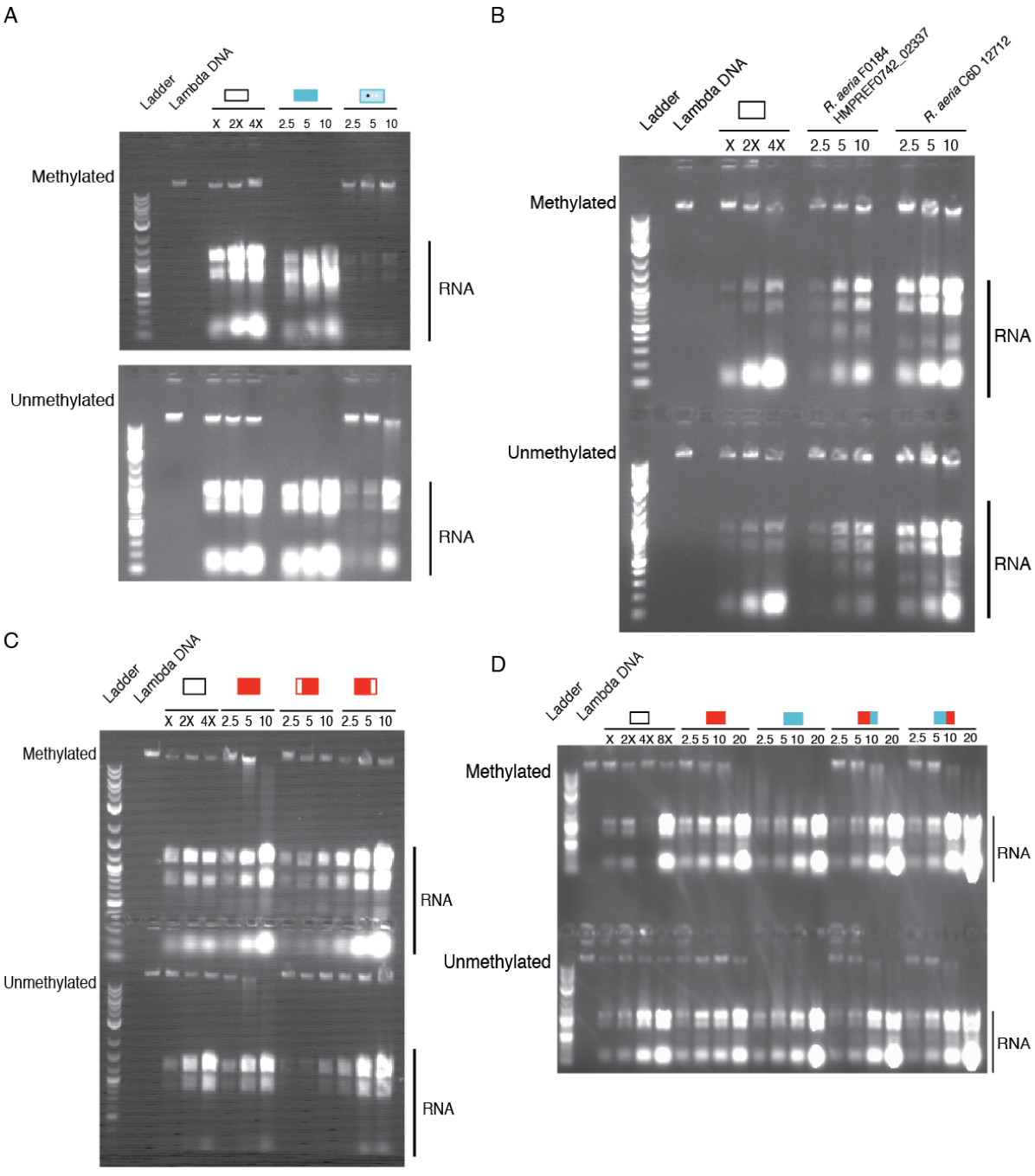

A
Ladder
Lambda DNA
X 2X 4X
2.5 5 10
2.5 5 10
Methylated
RNA
Unmethylated
RNA
B
Ladder
Lambda DNA
R. aeria F0184 HMPREF0742_02337
R. aeria C6D 12712
X 2X 4X
2.5 5 10
2.5 5 10
Methylated
RNA
Unmethylated
RNA
C
Ladder
Lambda DNA
X 2X 4X 2.5 5 10 2.5 5 10 2.5 5 10
Methylated
RNA
Unmethylated
RNA
D
Ladder
Lambda DNA
X 2X 4X 8X 2.5 5 10 20 2.5 5 10 20 2.5 5 10 20 2.5 5 10 20
Methylated
RNA
Unmethylated
RNA

Figure A.2: Similar IdrD-CTDs in *Rothia* show additional subdomain required for DNase function. (A-D) Full agarose gels of DNAse assays with lambda DNA of (A) IdrD^{Rothia}-CTD, (B) Similar IdrD-CTDs from *R. aeria* with N- and C-terminal truncations, (C) N-terminal and C-terminal IdrD-CTD truncations, and (D) *P. mirabilis* BB2000 and *R. aeria* C6B hybrids.

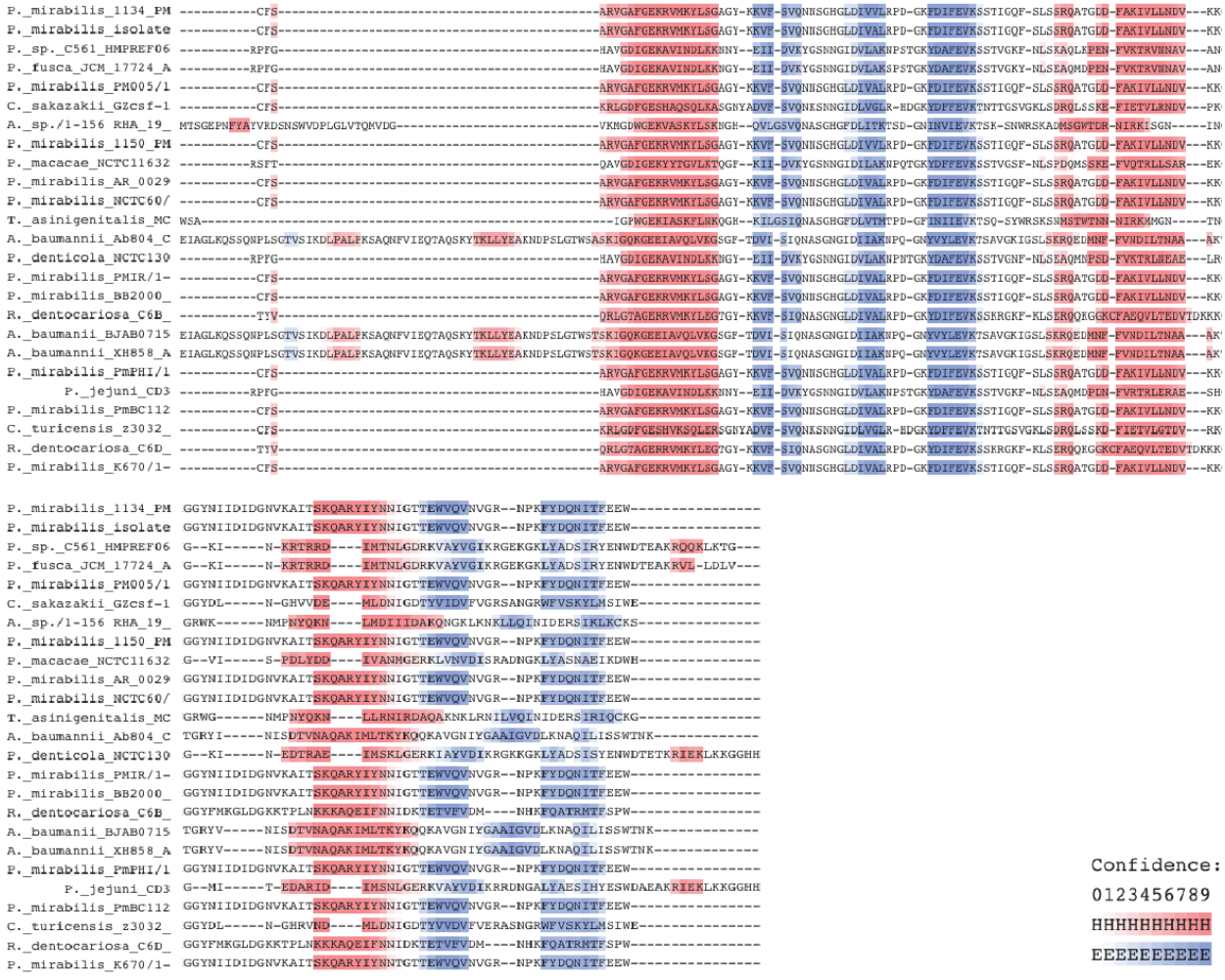


Figure A.3: *idrD* homologs are diverse and phylogenetically widespread. Alignment of rD-CTD and similar CTDs (MUSCLE followed by Ali2D on MPI Bioinformatics toolkit). Predicted alpha helices are red and beta sheets are purple; darker color indicates higher confidence.

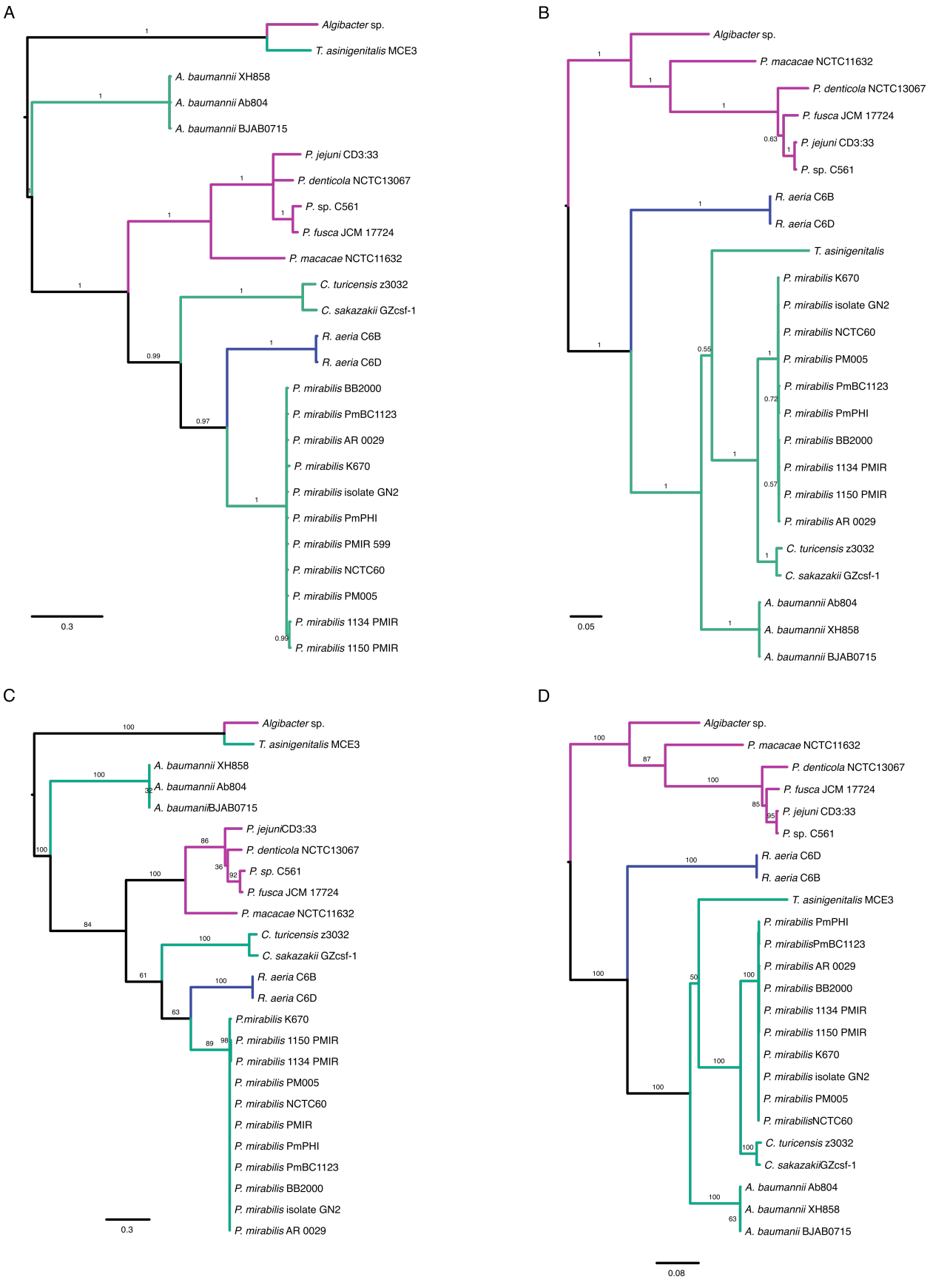

A
Algibacter sp.
T. asinigenitalis MCE3
A. baumannii XH858
A. baumannii Ab804
A. baumannii BJAB0715
P. jejuni CD3:33
P. denticola NCTC13067
P. sp. C561
P. fusca JM 17724
P. macacae NCTC11632
C. turicensis z3032
C. sakazakii GZcsf-1
R. aeria C6B
R. aeria C6D
P. mirabilis BB2000
P. mirabilis PmBC1123
P. mirabilis AR 0029
P. mirabilis K670
P. mirabilis isolate GN2
P. mirabilis PmPHI
P. mirabilis PMIR 599
P. mirabilis NCTC60
P. mirabilis PM005
P. mirabilis 1134 PMIR
P. mirabilis 1150 PMIR
0.3
B
Algibacter sp.
P. macacae NCTC11632
P. denticola NCTC13067
P. fusca JM 17724
P. jejuni CD3:33
P. sp. C561
R. aeria C6B
R. aeria C6D
T. asinigenitalis
P. mirabilis K670
P. mirabilis isolate GN2
P. mirabilis NCTC60
P. mirabilis PM005
P. mirabilis PmBC1123
P. mirabilis PmPHI
P. mirabilis BB2000
P. mirabilis 1134 PMIR
P. mirabilis 1150 PMIR
P. mirabilis AR 0029
C. turicensis z3032
C. sakazakii GZcsf-1
A. baumannii Ab804
A. baumannii XH858
A. baumannii BJAB0715
0.05
C
Algibacter sp.
T. asinigenitalis MCE3
A. baumannii XH858
A. baumannii Ab804
A. baumaniiBJAB0715
P. jejuniCD3:33
P. denticola NCTC13067
P. sp. C561
P. fusca JM 17724
P. macacae NCTC11632
C. turicensis z3032
C. sakazakii GZcsf-1
R. aeria C6B
R. aeria C6D
P.mirabilis K670
P. mirabilis 1150 PMIR
P. mirabilis 1134 PMIR
P. mirabilis PM005
P. mirabilis NCTC60
P. mirabilis PMIR
P. mirabilis PmPHI
P. mirabilis PmBC1123
P. mirabilis BB2000
P. mirabilis isolate GN2
P. mirabilis AR 0029
0.3
D
Algibacter sp.
P. macacae NCTC11632
P. denticola NCTC13067
P. fusca JM 17724
P. jejuni CD3:33
P. sp. C561
R. aeria C6D
R. aeria C6B
T. asinigenitalis MCE3
P. mirabilis PmPHI
P. mirabilisPmBC1123
P. mirabilis AR 0029
P. mirabilis BB2000
P. mirabilis 1134 PMIR
P. mirabilis 1150 PMIR
P. mirabilis K670
P. mirabilis isolate GN2
P. mirabilis PM005
P. mirabilisNCTC60
C. turicensis z3032
C. sakazakiiGZcsf-1
A. baumannii Ab804
A. baumannii XH858
A. baumanii BJAB0715
0.08

Figure A.4: Bayesian (A, B) and maximum likelihood (B, C) phylogenies of rD-CTD (A, C) and the 16S rRNA gene (B, D) representing the species tree. For A-D, Branches are colored by phyla (pink: Bacteriodetes, teal: Proteobacteria, blue: Actinobacteria), and scale bars show expected substitutions per site. Numbers adjacent to each branch report posterior probability (A, B) or bootstrap values (C, D). (A) Bayesian phylogeny of rD-CTD based on amino acid alignments. Translated amino acid sequences of the 23 identified rD-CTD homologs were aligned with muscle, filtered for 70% occupancy, and passed to MrBayes for phylogenetic reconstruction. (B) Bayesian species tree of taxa containing rD-CTD. 16S rRNA sequences were obtained from published genomes of the taxa represented in B, aligned with muscle, filtered for 70% occupancy, and passed to MrBayes for phylogenetic reconstruction. For A and B, nNumbers adjacent to each branch report posterior probability. Scale bars show expected substitutions per site. (C) Maximum-likelihood phylogeny of the rD-CTD alignment used in A but generated with RAxML. (D) Maximum-likelihood phylogeny of the 16S rRNA species tree alignment used in B but generated with RAxML.

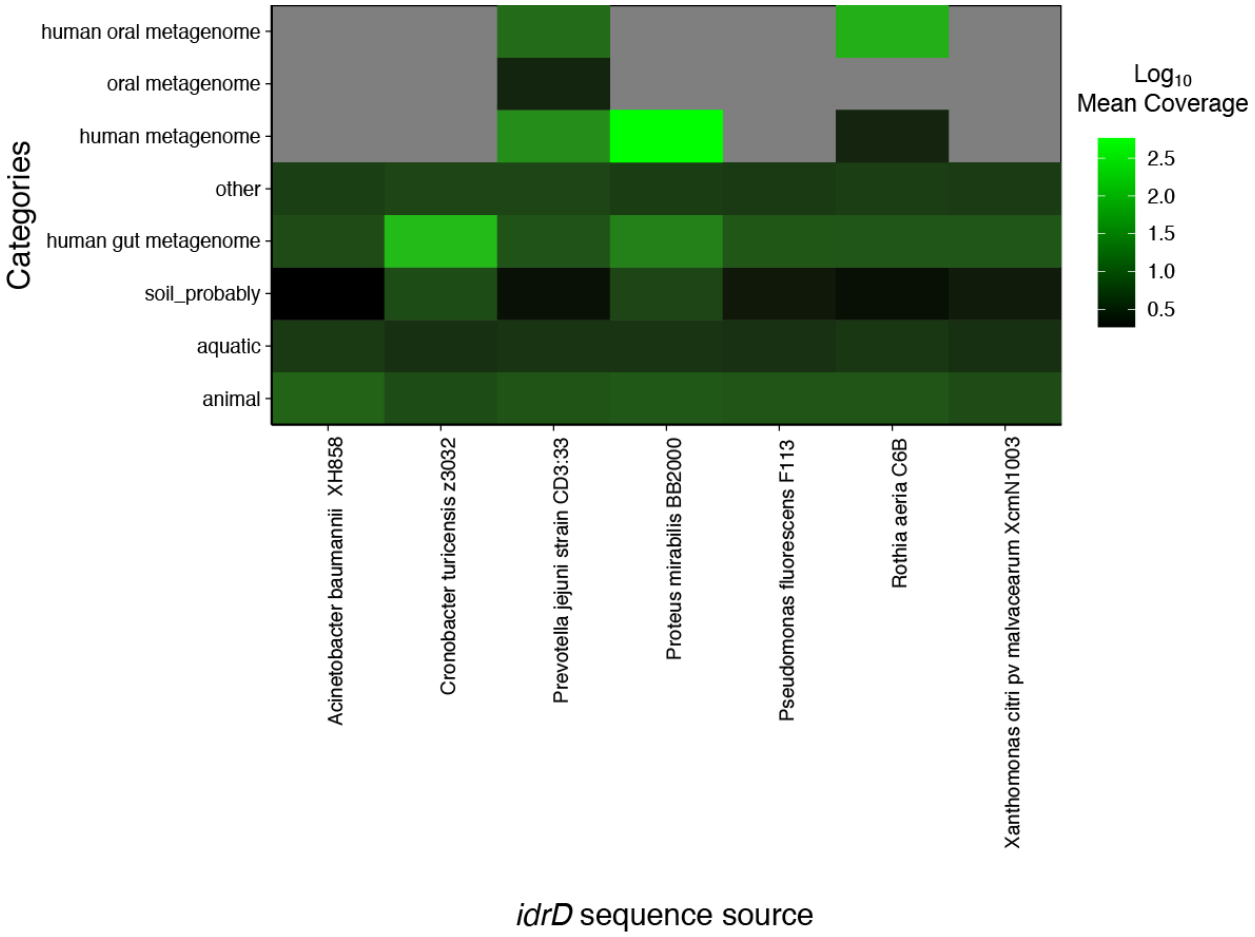


Figure A.5: *idrD*-like sequences are abundant in human metagenomes. This heatmap shows the log10-transformed mean coverage of each *idrD* sequence for different metagenome categories from metagenomes covering at least 50% of that sequence's nucleotides. Brighter green represents more coverage; black represents no coverage; grey cells represent combinations where no metagenomes in that category existed for that sequence due to the 50% coverage requirement.

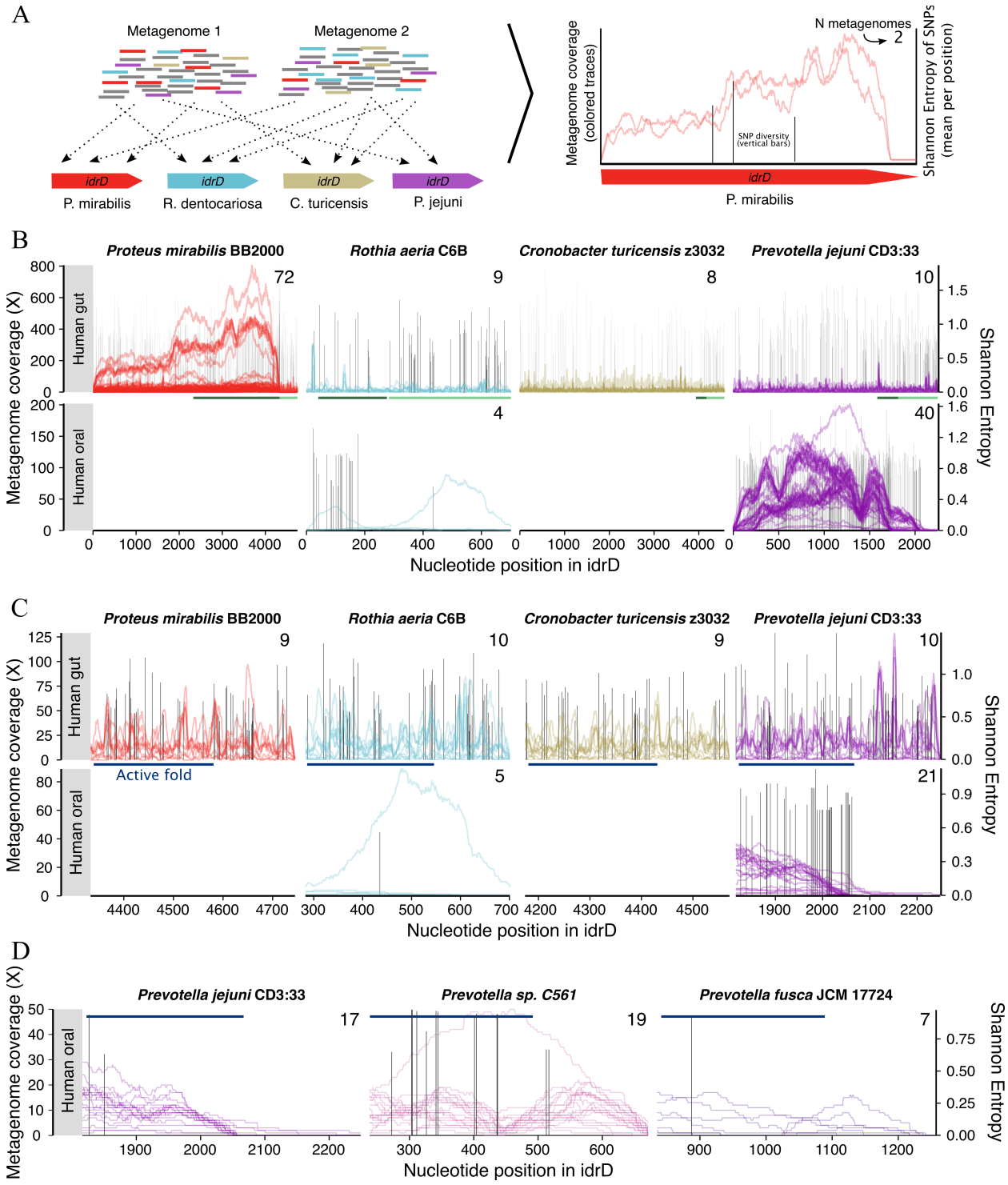
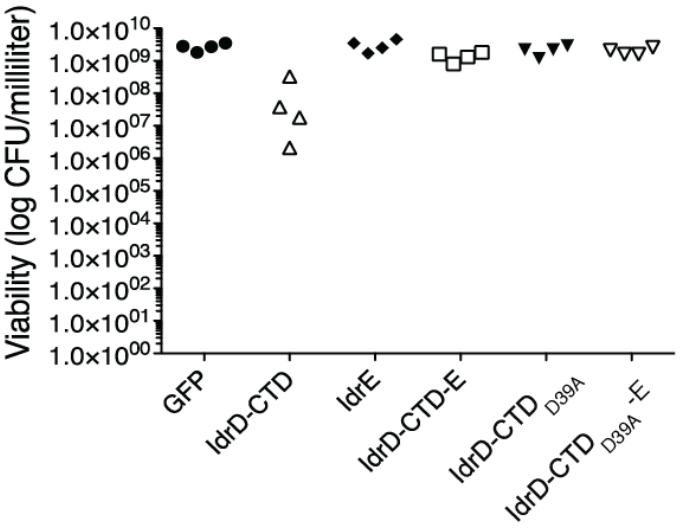
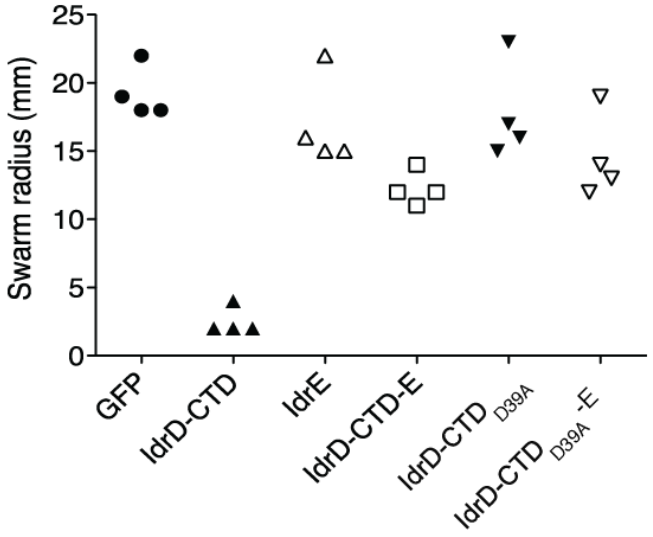

A
Metagenome 1
Metagenome 2
idrD
P. mirabilis
R. dentocariosa
C. turicensis
P. jejuni
Metagenome coverage (colored traces)
N metagenomes
2
SNP diversity (vertical bars)
Shannon Entropy of SNPs (mean per position)
P. mirabilis
B
Proteus mirabilis BB2000
Rothia aeria C6B
Cronobacter turicensis z3032
Prevotella jejuni CD3:33
Human gut
Human oral
Metagenome coverage (X)
Shannon Entropy
Nucleotide position in idrD
C
Active fold
D
Prevotella sp. C561
Prevotella fusca JCM 17724

Figure A.6: Diversity of *idrD*-like sequences in the human microbiome. The coverage data is identical to Figure 4, but the nucleotide variants mapped to the reference *idrD* sequences are reported (black vertical lines). Each line marks a nucleotide position where at least one metagenome mapped a variant nucleotide, and the height of the line corresponds to the Shannon Entropy of the frequencies of each nucleotide mapping to that position.

A


Viability (log CFU/milliliter)
1.0×10^10
1.0×10^09
1.0×10^08
1.0×10^07
1.0×10^06
1.0×10^05
1.0×10^04
1.0×10^03
1.0×10^02
1.0×10^01
1.0×10^00
GFP
IdrD-CTD
IdrE
IdrD-CTD-E
IdrD-CTD D39A
IdrD-CTD D39A -E

B


Swarm radius (mm)
25
20
15
10
5
0
GFP
IdrD-CTD
IdrE
IdrD-CTD-E
IdrD-CTD D39A
IdrD-CTD D39A -E

C

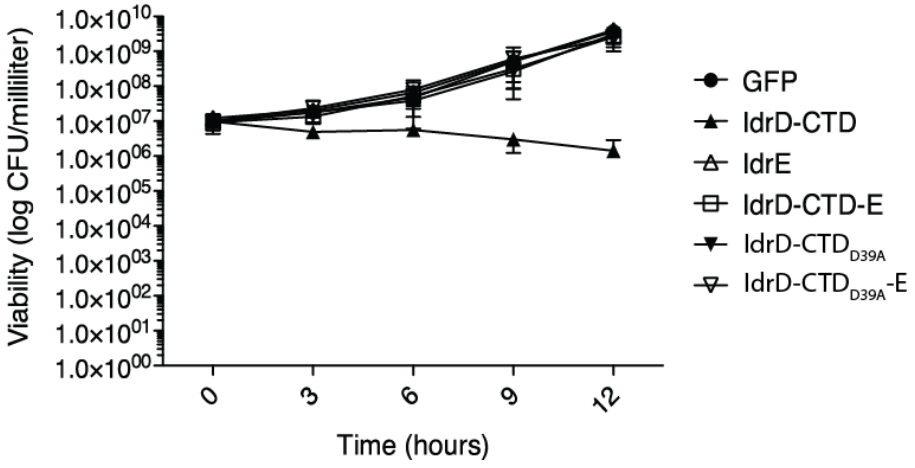

Viability (log CFU/milliliter)
1.0×10^10
1.0×10^09
1.0×10^08
1.0×10^07
1.0×10^06
1.0×10^05
1.0×10^04
1.0×10^03
1.0×10^02
1.0×10^01
1.0×10^00
0
3
6
9
12
Time (hours)
GFP
IdrD-CTD
IdrE
IdrD-CTD-E
IdrD-CTD D39A
IdrD-CTD D39A-E

Figure A.7: Co-expression of IdrE counteracts IdrD-CTD toxicity. (A) Quantification of viable cells after overexpression of IdrD-CTD and IdrE in swarmer cells. *P. mirabilis* BB2000 *idrD** swarmer cells containing expression vectors of IdrD-CTD (active and inactive), IdrE, and the two co-expressed from the same promoter were assayed for colony-forming units per milliliter (plotted on a $\log_{10}$ scale) compared to negative protein production (GFP) control. (B) Swarm distances of *P. mirabilis* BB2000 *idrD** swarmer cells containing expression vectors of IdrD-CTD and IdrE expression vectors. In both (A) and (B) solid data points were previously shown in Figures 1.1B and A.1B, respectively. (C) Quantification of viable liquid-grown *P. mirabilis* BB2000 *idrD** cells containing expression vectors of IdrD-CTD and IdrE expression vectors plotted on a $\log_{10}$ scale.

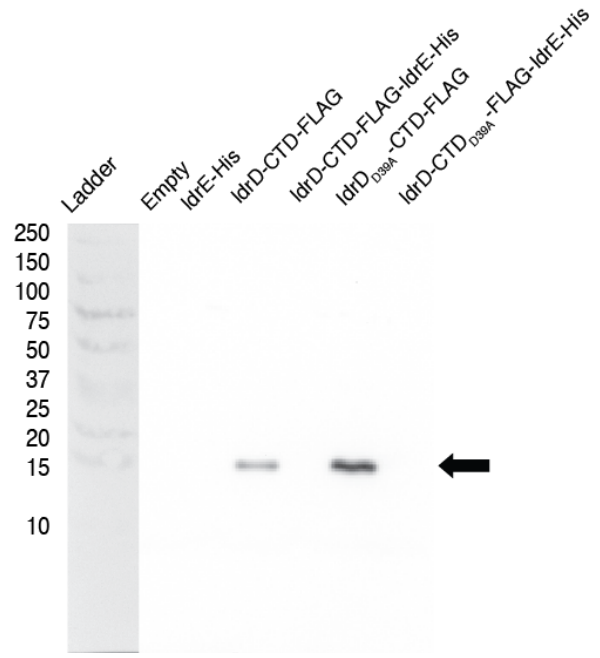


Figure A.8: α-FLAG western blot of IdrD-CTD-FLAG and IdrE-His expression vectors in *E. coli* MG1655. Whole cell extracts of *E. coli* strains MG1655, containing IdrD-CTD-FLAG, $IdrD_{D39A}$-CTD-FLAG and IdrE-His expression vectors were collected after 3 hours and run on an α-FLAG western blot to detect $IdrD_{D39A}$-CTD-FLAG. Arrow indicates size of IdrD-CTD-FLAG.

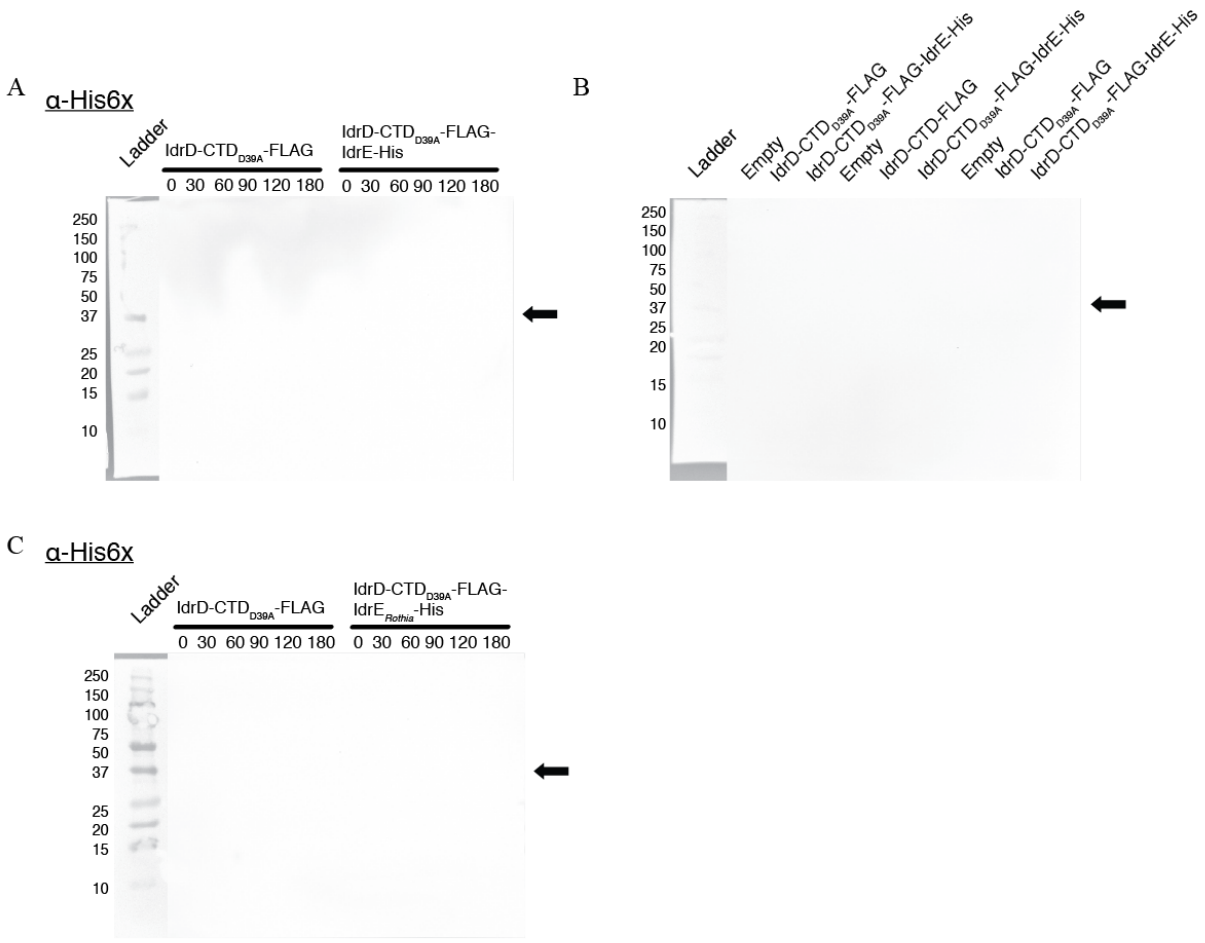


Figure A.9: α-His6x western blots of (A) IdrD$_{D39A}$-CTD-FLAG-IdrE expression vectors in *E. coli* strain BL21(DE3)pLysS, (B) IdrD-CTD$_{D39A}$-FLAG-IdrE expression vectors in protease-deficient *E. coli* backgrounds, and (C) IdrD$_{D39A}^{Proteus}$-CTD-FLAG-IdrE^{Rothia} expression vectors in *E. coli* strain BL21(DE3)pLysS. Whole cell extracts were collected from strains with corresponding expression vectors and run on 12% tris-tricine gel.

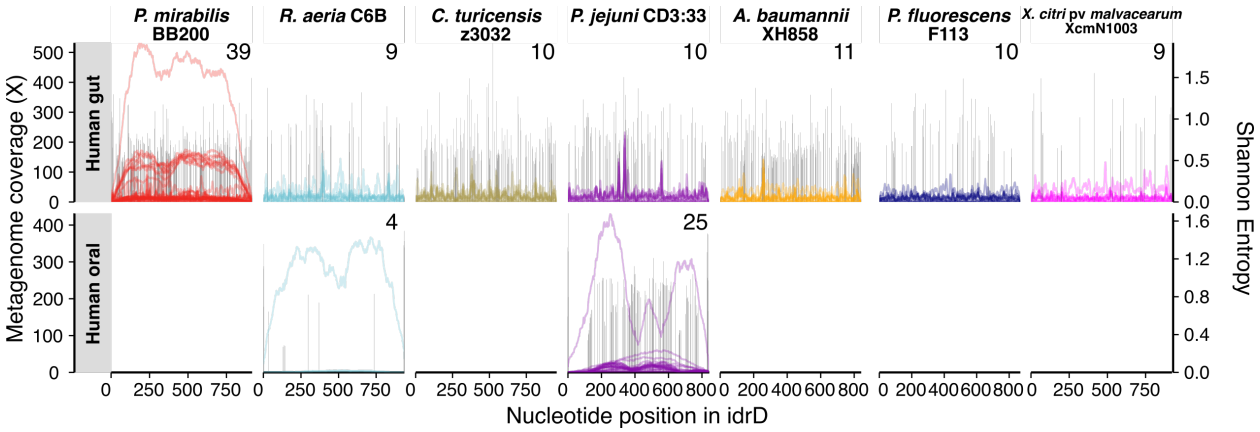


Figure A.10: *idrE*-like sequences in the human microbiome are diverse. The data is identical to Figure 3.4A, but information from nucleotide variation detected in the metagenomes are reported (black vertical lines). Each line marks a nucleotide position where at least one metagenome mapped a variant nucleotide, and the height of the line corresponds to the Shannon Entropy of the frequencies of each nucleotide mapping to that position.

Appendix B

Additional characterization of IdrD-CTD

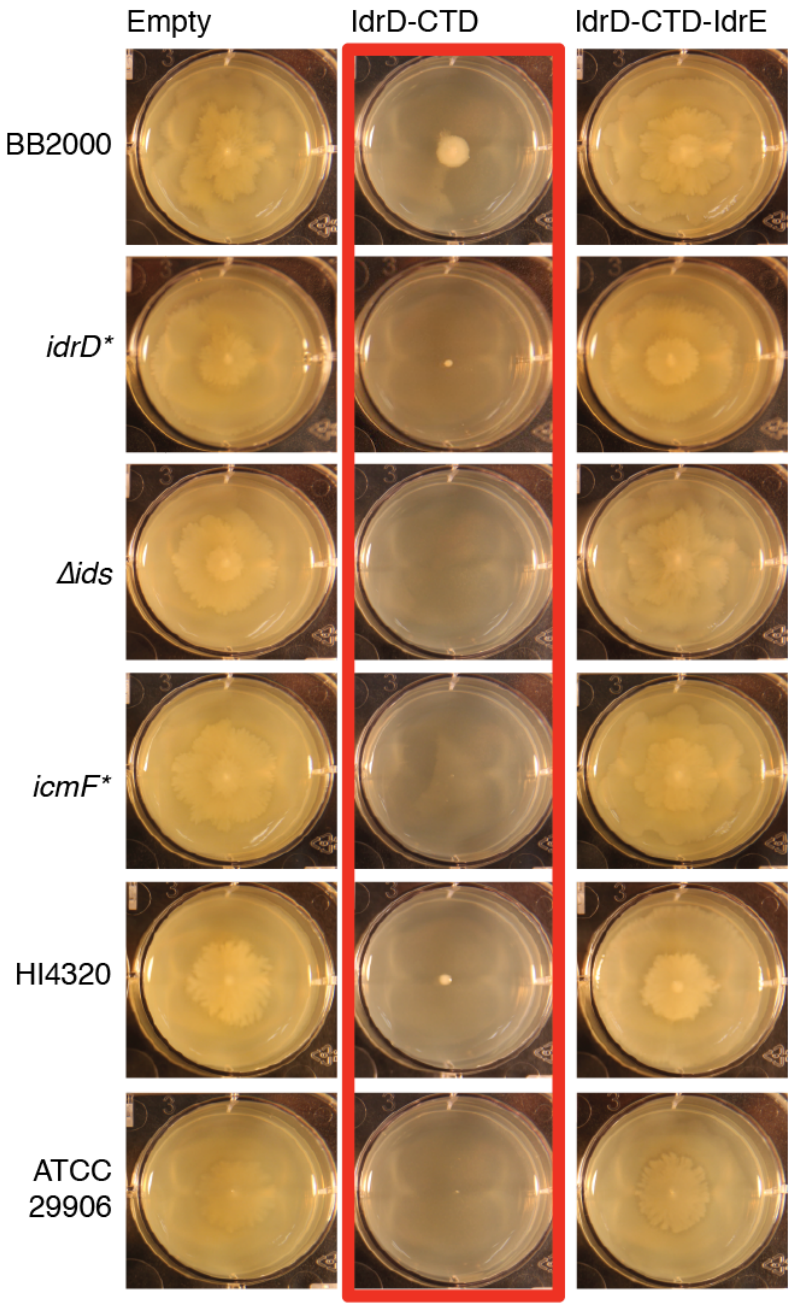


Figure B.1: Swarm assays of *P. mirabilis* containing overexpression vectors for IdrD-CTD and IdrD-CTD-IdrE. The rows correspond to the strains of *P. mirabilis* containing each vector; those in bold are from a BB2000 background. Each column corresponds to the vector present in each strain, all of which confer resistance to kanamycin and are induced by anhydrous tetracycline (10nM anhydrous tetracycline). The plates surrounded by a red box exhibit inhibited swarming.

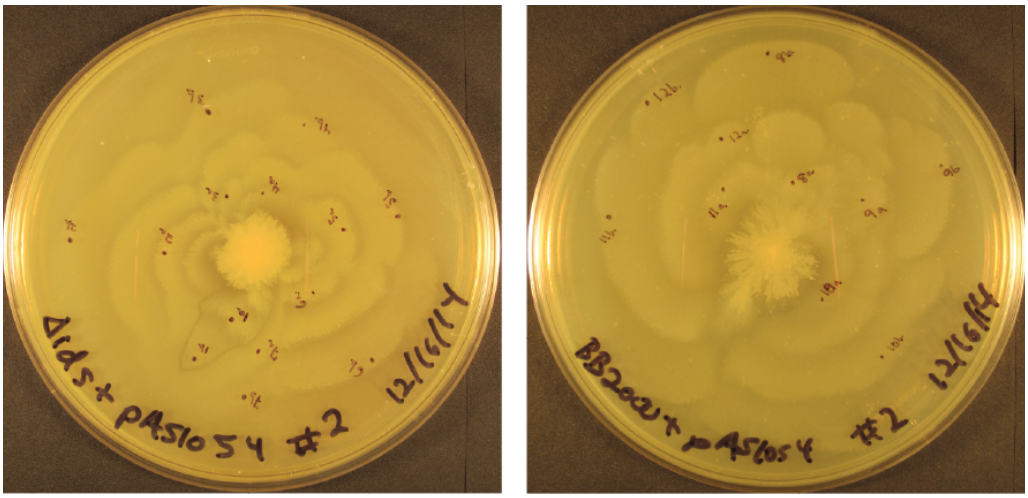


Figure B.2: Suppressor screen to identify target of IdrD-CTD. *P. mirabilis* BB2000 Δ*ids* carrying an IdrD-CTD expression vector under the control of the predicted *idr* promoter (P_{idrA}) was allowed to swarm for ~2 weeks at room temperature. Collected suppressors are marked on the petri dishes. Selected hits are shown in Table A.1; we hypothesize that these hits are involved in regulation of P_{idrA}.

Table B.1: Select mutations from suppressor screen of *P. mirabilis* BB2000 and Δ*ids* carrying an IdrD-CTD expression vector

Background	Locus tag	Gene Product	Mutation
BB2000 and Δ*ids*	BB2000_0822	IdrA	Truncation
BB2000	BB2000_0825	IdrD	Truncation
Δ*ids*	BB2000_816	TssF	Truncation

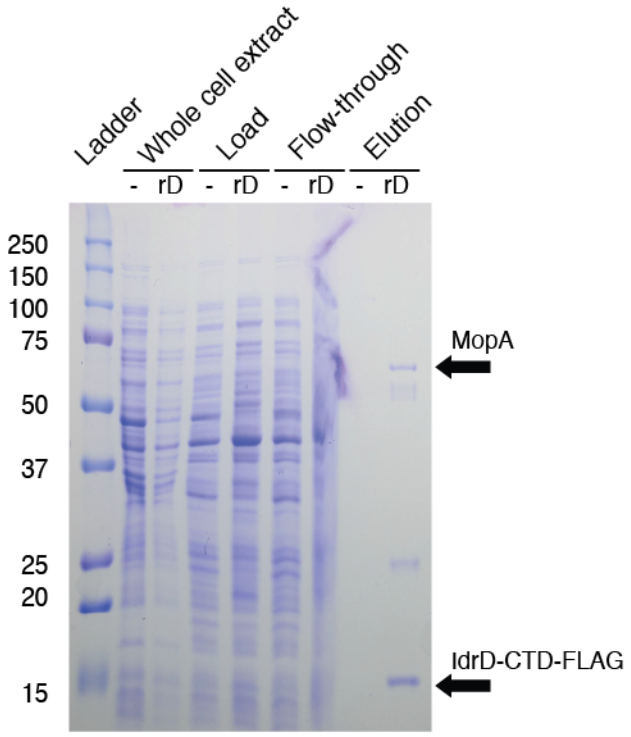


Figure B.3: Immunoprecipitation of IdrD-CTD-FLAG from *E. coli* MG1655. Whole cell extracts were collected from induced *E. coli* MG1655 cells containing empty vector (-) or overexpressing IdrD-CTD-FLAG (rD). Lysates were applied to an α–FLAG resin to pull-down IdrD-CTD-FLAG and any binding partners. The band in the elution fraction of IdrD-CTD-FLAG at ~60kDa was excised and sent off for LC-MS/MS and identified as MopA.

Table B.2: LC-MS/MS results for excised band (~60kDa) from IdrD-CTD-FLAG immunoprecipitation of *E. coli* MG1655 lysates (>2 unique peptides)

Protein	Predicted size (kDa)	Number of unique peptides	Number of total peptides	Coverage (%)
MopA	**57.29**	**52**	**252**	**2.6548**
DnaK	69.07	9	10	2.2442
NuoC	68.21	7	7	2.9191
Tig	48.22	6	6	2.4846
SdhA	64.38	5	5	2.7069
TrxA	15.98	4	5	2.9032
Lpp	8.32	4	4	3.1351
TufB	43.29	4	4	2.5454
RpsA	61.12	4	4	2.4598
RpsB	30.95	3	3	3.2139
RplL	12.29	3	3	2.6832

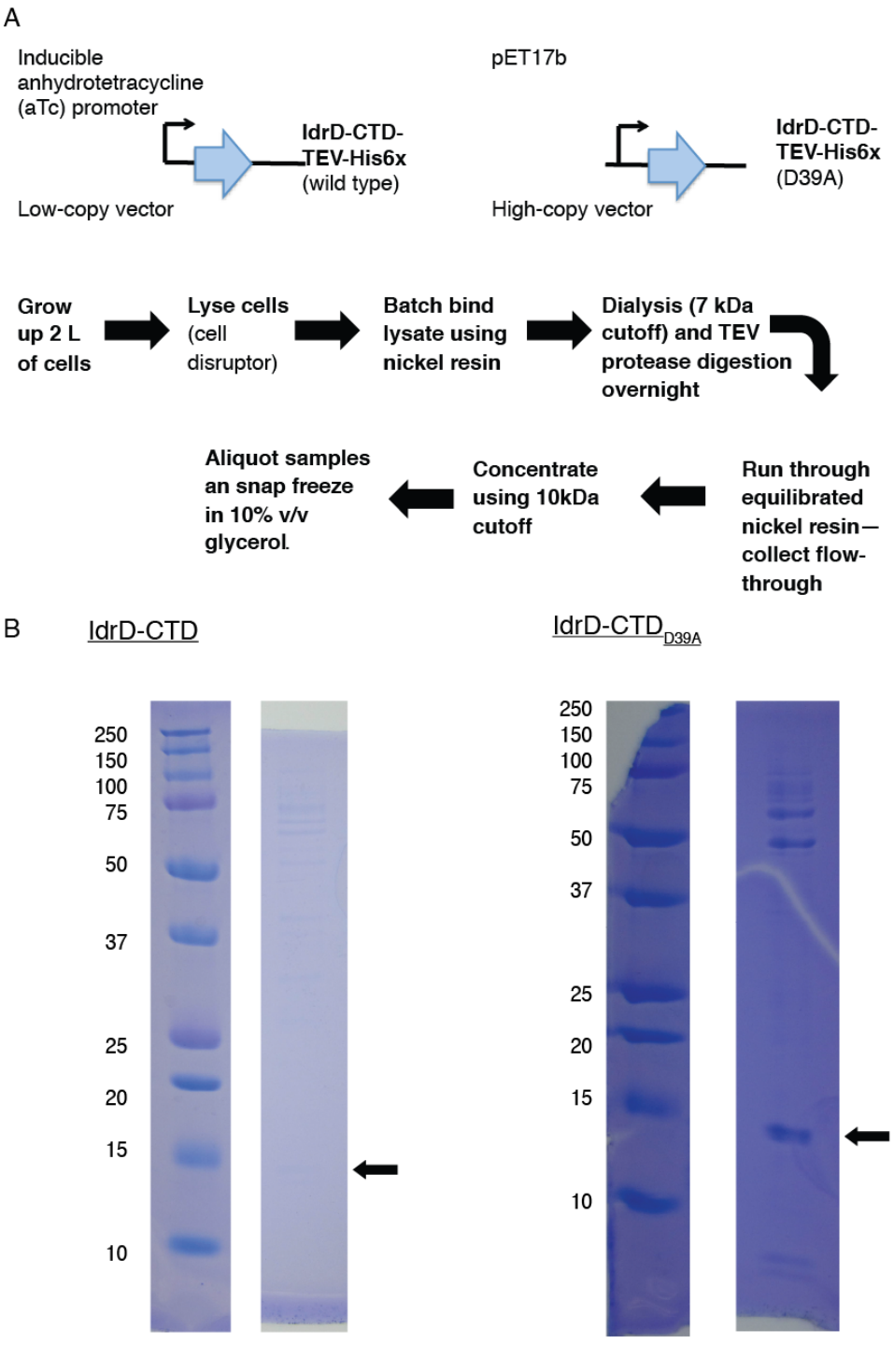

A
Inducible anhydrotetracycline (aTc) promoter
IdrD-CTD-TEV-His6x (wild type)
Low-copy vector
pET17b
IdrD-CTD-TEV-His6x (D39A)
High-copy vector
Grow up 2 L of cells
Lyse cells (cell disruptor)
Batch bind lysate using nickel resin
Dialysis (7 kDa cutoff) and TEV protease digestion overnight
Run through equilibrated nickel resin—collect flow-through
Concentrate using 10kDa cutoff
Aliquot samples an snap freeze in 10% v/v glycerol.
B
IdrD-CTD
IdrD-CTD D39A
250
150
100
75
50
37
25
20
15
10

Figure B.4: IdrD-CTD protein purification. (A) Schematic of purification steps. Two different expression vectors were used to express IdrD-CTD and IdrD-CTDD39A. Wild type IdrD-CTD-TEV-His6x could not be cloned into protein production vector pET17b. Both versions were subsequently purified by the steps described. (B) Final elution fractions of IdrD-CTD. IdrD-CTDD39A is visible on Coomassie-stained gel; however IdrD-CTD is not.

Materials and Methods

P. mirabilis swarm assays

Performed as described in Chapter 2.

Isolation of spontaneous mutant strains of *P. mirabilis* BB2000 and BB2000-derived Δ*ids* carrying an IdrD-CTD expression vector

Spontaneous mutant strains of BB2000 and BB2000-derived Δ*ids* carrying an IdrD-CTD expression vector under the control of the predicted *idr* promoter (P_{idrA}) were isolated. Genomic DNA (gDNA) from these isolates was extracted using phenol-chloroform extractions. gDNA was sheared using a Covaris S 220 (Covaris, Woburn, MA). Libraries prepared using the PrepX ILM DNA Library Kit (WaferGen Biosystems, Fremont, CA) for the Apollo 324 NGS Library Prep System (WaferGen Biosystems, Fremont, CA). Using an Illumina HiSeq 2500 system (Illumina, San Diego, CA), the library was sequenced as 100 base pair (bp), paired-end reads. To identify suppressor-specific polymorphisms, reads were aligned to the *P. mirabilis* BB2000 and Δ*ids* genomes (GenBank accession no. CP004022) using Geneious (Biomatters, Auckland, New Zealand). All genome sequencing was performed by the Bauer Core Facility at Harvard University.

Anti-FLAG immunoprecipitations from *E. coli* cell extracts

E. coli MG1655 cells were grown in 25 mL of LB supplemented with kanamycin and anhydrotetracylcine (inducer) under shaking conditions at 37°C for 3 hours. Cell pellets were re-suspended in 1 mL cell lysis buffer (50mM Tris HCl pH 7.4, 150mM NaCl, 1% triton X-100, 1 mM EDTA) supplemented with 40 µl of either Complete protease inhibitor cocktail (Roche, Basel, Switzerland). Pellets were lysed by vortexing with cell disruptor beads (0.1-diameter,

Electron Microscopy Sciences, Hatfield, PA). Lysates were centrifuged and 900 µL of supernatant was applied to 40 µL pre-equilibrated α-FLAG M2 antibody resin (Sigma-Aldrich, St. Louis, MO; Biotools, Houston, TX). Lysates were incubated with resin for two hours at 4°C after which unbound cell extract was removed. Five wash steps were performed in wash buffer (50mM Tris HCl pH 7.4, 150mM NaCl, 1% triton X- 100). Bound proteins were eluted with 50 µl of elution buffer (50mM Tris HCl pH 7.4, 150mM NaCl, 200 ng/µl 3XFLAG peptide) for 45 minutes at 4°C. The elution was re-centrifuged and the top 40 µl was retained. All samples were run on 12% tris-tricine gel and subjected to a Coomassie blue stain. Band of interest was excised and sent LC-MS/MS analysis (Taplin Mass Spectrometry Core Facility, Harvard Medical School, Boston MA).

Protein purification of IdrD-CTD and IdrD$_{D39A}$-CTD

For IdrD-CTD-TEV-His, *E. coli* BL21 (DE3) pLysS cells (Thermo Fisher Scientific, Waltham, MA) were grown as described for immunoprecipiation. For IdrD$_{D39A}$–CTD-TEV-His, *E. coli* BL21 (DE3) pLysS cells were grown in 2 L of LB supplemented with carbenicillin under shaking conditions at 30°C. When optical density at 600 nm (OD_{600}) was between 0.6 and 1, cultures were cooled on ice, and induced with 1 mM isopropyl-β-D-1 thiogalactopyranoside (IPTG). Cultures were then incubated overnight shaking at 16°C. Cells were harvested by centrifugation and lysed using a cell disruptor. Lysate was applied to nickel resin (2 ml of slurry per liter of culture), and incubated for 1.5 hours at 4°C. Lysate-resin mixture was applied to a column, flowed over twice post-binding, and subjected to three washes. Elution was performed at least five times (one column volume each). Elution fractions were run on a 12% tris tricine gel to determine which fractions contained protein. Fractions with protein were pooled and put in a dialysis cassette along with TEV protease (1:50 mg of TEV to protein); cassette was left at

4°C overnight. Sample was applied to column with nickel resin twice, and flow-through was collected. Samples were concentrated, run on a 12% tris tricine gel, and stained with Coomassie blue to detect protein.