PERSPECTIVE
Ecological and Evolutionary Science

Exploring the Evolution of Virulence Factors through Bioinformatic Data Mining

Andrew C. Doxey,[a] Michael J. Mansfield,[a] Briallen Lobb[a]

[a]Department of Biology, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT The molecular evolution of virulence factors is a central theme in our understanding of bacterial pathogenesis and host-microbe interactions. Using bioinformatics and genome data mining, recent studies have shed light on the evolution of important virulence factor families and the mechanisms by which they have adapted and diversified in function. This perspective highlights three complementary approaches useful for studying the molecular evolution of virulence factors: identification and analysis of virulence factor homologs, detection of adaptations or functional shifts, and computational prediction of novel virulence factor families. Each of these research directions is associated with distinct questions, approaches, and challenges for future work. Moving forward, bioinformatics will continue to play a critical role in exploring the evolution of virulence factors, including those that target humans. By reconstructing past processes and events, we will be able to better interpret newly sequenced microbial genomes and detect future pathoadaptations.

KEYWORDS bioinformatics, microbial genomics, molecular evolution, pathogens, virulence factors

Pathogenic microorganisms interact with hosts by producing specialized virulence factor (VF) proteins that are capable of interacting with or disrupting host processes. Due to the immense biomedical relevance of virulence factors in human infectious disease, there is a long history of research into their biology, which has revealed a remarkable diversity and sophistication in terms of structure, specificity, and mode of action. Many virulence factors are evolutionarily fine tuned to interact with and disrupt specific receptors, pathways, cell types, tissue types, and host species. This raises important evolutionary questions such as the following. How do virulence factors originate, diversify, and adapt over time? How can we use this knowledge of virulence factor evolution to discover novel virulence factors within the vast and growing collection of sequenced microbial genomes?

With the growing availability of genomes across the tree of life, it has become increasingly possible to perform comprehensive and detailed analyses of virulence factor diversity and evolution. A critical first step in many such approaches is the construction and curation of VF databases (1). The virulence factor database (VFDB), which is derived largely from a set of 74 bacterial pathogen genomes, currently contains 1,074 virulence factors. However, the VFDB also contains an additional 32,312 related VFs that can be computationally detected in genomes through homology (1). When factoring in additional pathogenic species, as well as the growing repertoire of secreted effectors, peptides, and other virulence-related molecules found in different host-pathogen systems (2), this number likely grows by orders of magnitude. Given such a vast quantity of data and possibilities for analysis, a useful schema is to divide bioinformatic analysis of VF evolution into the three conceptual approaches discussed below.

Citation Doxey AC, Mansfield MJ, Lobb B. 2019. Exploring the evolution of virulence factors through bioinformatic data mining. mSystems 4:e00162-19. https://doi.org/10.1128/mSystems.00162-19.

Address correspondence to Andrew C. Doxey, acdoxey@uwaterloo.ca.

Conflict of Interest Disclosures: A.C.D. has nothing to disclose. M.J.M. has nothing to disclose. B.L. has nothing to disclose.

mSystems® *vol. 4, no. 3, is a special issue sponsored by Illumina.*

Exploring virulence factor evolution and diversity using bioinformatics: A perspective on our work mining genomic databases for homologs, adaptions, and new virulence factor families.

Received 1 March 2019
Accepted 11 March 2019
Published 21 May 2019

Identifying VF homologs: exploring the sequence space near characterized VFs. New virulence factors are typically predicted based on detectable homology to previously characterized VF sequences. Newly detected VF homologs can be of great value, as they can expand the known "sequence space" of a VF family, identify VFs in unexpected taxa, and even identify distant relationships between seemingly unrelated VF families. Detected homologs may vary considerably in their sequence similarity to a reference VF, from near-identical orthologs to remote homologs with low sequence similarity. Therefore, it can be challenging to draw the line between where one family ends and another begins, as well as to understand how sequence divergence relates to conservation of VF function.

Homology-based prediction of VFs in genomes has underpinned a number of recent studies by our lab exploring the molecular evolution and diversity of bacterial exotoxins, a major class of VFs. Well-known examples include the botulinum neurotoxins (BoNT) and tetanus neurotoxins (TeNT) produced by various *Clostridia* as well as the diphtheria toxin (DT) produced by members of the *Corynebacterium* genus. BoNT and DT are also notable as the first bacterial toxins to be discovered, dating back to the late 1800s (3). Primarily through applications of PSI-BLAST, we recently identified the first homologs of these toxin families outside their respective bacterial lineages (4–6). The botulinum neurotoxin-like toxins occur outside the *Clostridium* genus in bacteria such as *Weissella* and *Enterococcus* and phylogenetically cluster nearby but outside the characterized BoNT tree (Fig. 1a). Similarly, DT-like toxins occur outside *Corynebacterium*, primarily in other actinobacterial organisms such as *Austwickia* and *Streptomyces*. These DT-like toxins also phylogenetically cluster nearby but outside the DT clade (Fig. 1b).

The identification of divergent homologs of these human disease-causing toxins has intriguing evolutionary implications. Their occurrence in bacteria not known for causing human infectious disease suggests that they may target different host species "in the wild" (e.g., the genome of *Austwickia chelonae* encodes a DT-like toxin, and this organism is associated with infectious lesions in reptiles rather than humans) (7). Furthermore, the sequence divergence and phylogenetic position of BoNT-like and DT-like toxins suggest that they potentially represent ancient lineages that predate the emergence of the well-known forms associated with human disease.

A major drawback of identifying toxins from sequence information is that it is difficult to predict their specific functional properties, such as host and substrate specificity (3); thus, it remains unclear how host specificity maps onto the evolutionary history of these toxin families. Once established, this information may help elucidate key evolutionary adaptations through which ancestral BoNT-like and DT-like toxins acquired specificity for different hosts, including humans.

Detecting VF adaptations and functional shifts. Mapping the evolution of function and specificity within VF families is an important and highly challenging research objective (Fig. 1c). Here, it is important to consider that host-pathogen arms races may drive accelerated evolution of VFs, as well as diversification of different aspects of VF function, including host, cell type, and substrate specificity. State of the art methods for exploring the evolution of specificity involve an integration of experimental and computational approaches and also rely heavily on structural biology (8). However, the ability to reconstruct such events also depends largely on the evolutionary timescale associated with the sequences being analyzed.

One exquisite example of VF adaptation is the molecular evolution of typhoid toxin. Structural studies of typhoid toxin have revealed the basis for its specificity toward humans. Unique amino acid substitutions in its glycan-binding domain allow the typhoid toxin to selectively bind human glycans terminated in Neu5Ac over glycans terminated in Neu5Gc, which are produced by other mammals (9). Although similar patterns of adaptive evolution have likely occurred in the binding domains of BoNT-like and DT-like toxins throughout their history (Fig. 1), it is difficult to pinpoint the key specificity-determining substitutions in these cases because evolutionary distances are

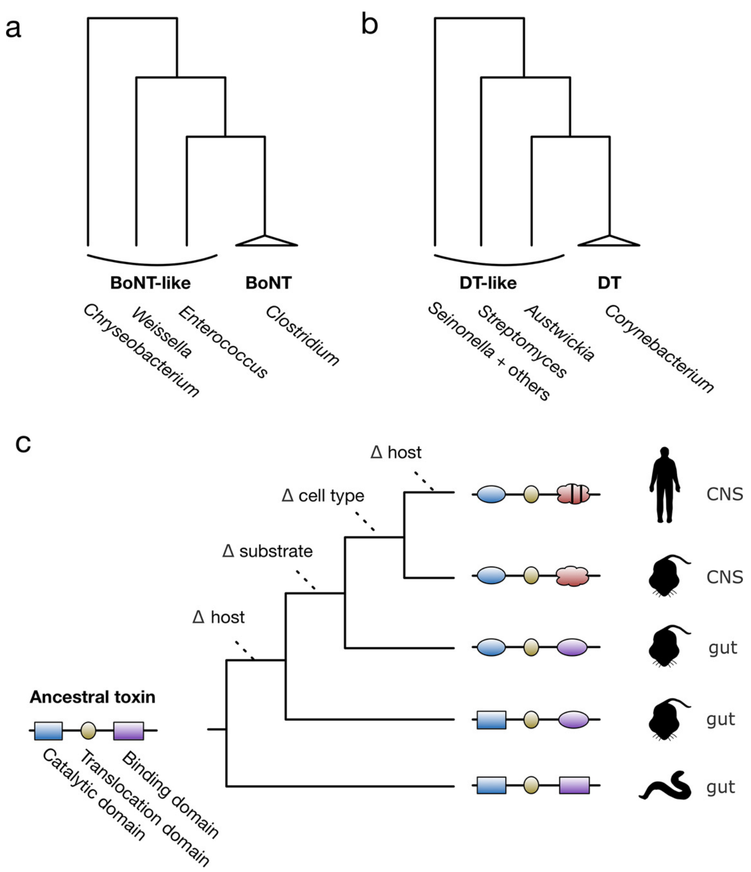

FIG 1 Bioinformatically identified toxins provide insights into the evolutionary origins of major toxin families. (a and b) General overviews of the phylogenies of botulinum neurotoxin (BoNT) and diphtheria toxin (DT), including recently discovered homologs. (c) General model depicting the hypothetical evolution of an ancestral toxin that diversifies over time in terms of host, cell type, and substrate specificity. Changes in the catalytic domain (blue) are associated with alterations in substrate specificity, while changes in the binding domain (purple or red) are associated with alterations in host specificity and cell/tissue type (e.g., from the gut to the central nervous system [CNS]). This model is one potential explanation for observed sequence patterns in these toxin families.

large. In cases like these, methods that detect positive selection involving multiple substitutions may be required. For example, we developed an approach to detect spatially clustered substitutions in protein structure (10). This approach was successful in predicting ancestral adaptations within protein structures, such as the gain of a cellulose-binding site in a pathogenesis-related protein that likely contributes to its antifungal specificity (11).

Even more dramatic functional adaptations are cases in which entire segments, such as protein domains, are acquired or lost from a VF family. Recombination and domain shuffling appear to be very common in the evolution of VFs (2), and in some cases, they may even underlie the origin of new virulence functions. For example, we recently discovered a family of bacterial flagellins called flagellinolysins that have acquired a central metalloprotease domain, allowing them to localize metalloprotease activity within the external flagellar filament (12). This novel function resulting from the evolution of a new domain combination may contribute to virulence in *Clostridium haemolyticum* and other pathogens.

Predicting new VF families and mechanisms. Although the above cases explore VF diversity and evolution by starting with already characterized VF families, the computational discovery of entirely new VF families that are unrelated to existing ones is arguably the most challenging approach. One intriguing way to look for these VF sequences is to explore the thousands of domains of unknown function and "ORFan" families that have been mined from genomic and metagenomic databases (13). Because standard homology search will not work for these cases, one must rely on external information that implicates an uncharacterized protein family in virulence, such as its upregulation in virulence-associated conditions, its association with virulence phenotypes (14), its occurrence in a pathogenicity island (15), or its overrepresentation within genomes of pathogenic or host-associated species (16). Given that many virulence factors exhibit similarities to host proteins, allowing them to interfere with host functions, the identification of host-like or "mimicry" proteins in bacterial genomes is another effective strategy for virulence factor discovery (17). A recent study by Levy et al. (16) applied several of these methods to explore the genomic determinants of plant-associated bacteria. By comparing genomes of plant-associated versus non-plant-associated bacteria, Levy et al. identified thousands of plant-associated gene clusters. These clusters included new protein families that perfectly correlated with pathogenicity lifestyles, and 64 "mimicry" proteins that potentially mimic domains found in plants.

Future directions. With ongoing sequencing, bioinformatics will continue to expand the sequence space of VFs. This will contribute to our understanding of VF evolution and identify key sequence determinants of function. As the inventory of known VFs increases, so will our ability to predict bacterial pathogenicity from genome information, which has important implications for human health and disease. Mapping the evolution of specificity within VF and toxin families will also be critical for identifying and potentially predicting future pathoadaptations toward humans and other species. An exciting direction for future work is the use of sophisticated machine-learning approaches to discover commonalities among different virulence factor families that cannot be recognized using homology and application of these approaches to discover new VFs in genomes. Finally, through large-scale recovery of metagenome-assembled genomes, there has been an explosion of new genomic and microbiome diversity (18). New gene families encoding virulence factors and toxins can be identified in these data sets, without the need for direct culturing and experimentation (3). Through metagenomic exploration of the diversity and roles of virulence factors in host-associated microbiomes and environmental microbial communities, it will become possible to glean insights into the broader ecological roles of VFs in host-microbe interactions and ecosystem function.

ACKNOWLEDGMENTS

We thank the editorial board of *mSystems* for the opportunity to contribute to this special issue. We also thank Josh Neufeld and Laura Hug for providing useful comments on the manuscript.

A.C.D. is supported by an Ontario Early Research Award and by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant.

REFERENCES

1. Liu B, Zheng D, Jin Q, Chen L, Yang J. 2019. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res 47:D687–D692. https://doi.org/10.1093/nar/gky1080.
2. Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L. 2012. Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol Direct 7:18. https://doi.org/10.1186/1745-6150-7-18.
3. Doxey AC, Mansfield MJ, Montecucco C. 2018. Discovery of novel bacterial toxins by genomics and computational biology. Toxicon 147:2–12. https://doi.org/10.1016/j.toxicon.2018.02.002.
4. Mansfield MJ, Adams JB, Doxey AC. 2015. Botulinum neurotoxin homologs in non-Clostridium species. FEBS Lett 589:342–348. https://doi.org/10.1016/j.febslet.2014.12.018.
5. Mansfield MJ, Sugiman-Marangos SN, Melnyk RA, Doxey AC. 2018. Identification of a diphtheria toxin-like gene family beyond the *Corynebacterium* genus. FEBS Lett 592:2693–2705. https://doi.org/10.1002/1873-3468.13208.
6. Zhang S, Lebreton F, Mansfield MJ, Miyashita S-I, Zhang J, Schwartzman JA, Tao L, Masuyer G, Martínez-Carranza M, Stenmark P, Gilmore MS, Doxey AC, Dong M. 2018. Identification of a botulinum neurotoxin-like toxin in a commensal strain of Enterococcus faecium. Cell Host Microbe 23:169–176.e6. https://doi.org/10.1016/j.chom.2017.12.018.
7. Mansfield MJ, Doxey AC. 2018. Genomic insights into the evolution and ecology of botulinum neurotoxins. Pathog Dis 76:fty040. https://doi.org/10.1093/femspd/fty040.
8. Hochberg GKA, Thornton JW. 2017. Reconstructing ancient proteins to understand the causes of structure and function. Annu Rev Biophys 46:247–269. https://doi.org/10.1146/annurev-biophys-070816-033631.
9. Deng L, Song J, Gao X, Wang J, Yu H, Chen X, Varki N, Naito-Matsui Y, Galán JE, Varki A. 2014. Host adaptation of a bacterial toxin from the

human pathogen Salmonella typhi. Cell 159:1290–1299. https://doi.org/10.1016/j.cell.2014.10.057.

10. Adams J, Mansfield MJ, Richard DJ, Doxey AC. 2017. Lineage-specific mutational clustering in protein structures predicts evolutionary shifts in function. Bioinformatics 33:1338–1345. https://doi.org/10.1093/bioinformatics/btw815.
11. Doxey AC, Cheng Z, Moffatt BA, McConkey BJ. 2010. Structural motif screening reveals a novel, conserved carbohydrate-binding surface in the pathogenesis-related protein PR-5d. BMC Struct Biol 10:23. https://doi.org/10.1186/1472-6807-10-23.
12. Eckhard U, Bandukwala H, Mansfield MJ, Marino G, Cheng J, Wallace I, Holyoak T, Charles TC, Austin J, Overall CM, Doxey AC. 2017. Discovery of a proteolytic flagellin family in diverse bacterial phyla that assembles enzymatically active flagella. Nat Commun 8:521. https://doi.org/10.1038/s41467-017-00599-0.
13. Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC. 2015. Remote homology and the functions of metagenomic dark matter. Front Genet 6:234. https://doi.org/10.3389/fgene.2015.00234.
14. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y, Carlson HK, Esquivel Z, Sadeeshkumar H, Chakraborty R, Zane GM, Rubin BE, Wall JD, Visel A, Bristow J, Blow MJ, Arkin AP, Deutschbauer AM. 2018. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature 557:503–509. https://doi.org/10.1038/s41586-018-0124-0.
15. Ho Sui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL. 2009. The association of virulence factors with genomic islands. PLoS One 4:e8094. https://doi.org/10.1371/journal.pone.0008094.
16. Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, Wang K, Devescovi G, Stillman K, Monteiro F, Rangel Alvarez B, Lundberg DS, Lu T-Y, Lebeis S, Jin Z, McDonald M, Klein AP, Feltcher ME, Rio TG, Grant SR, Doty SL, Ley RE, Zhao B, Venturi V, Pelletier DA, Vorholt JA, Tringe SG, Woyke T, Dangl JL. 2018. Genomic features of bacterial adaptation to plants. Nat Genet 50:138–150. https://doi.org/10.1038/s41588-017-0012-9.
17. Doxey AC, McConkey BJ. 2013. Prediction of molecular mimicry candidates in human pathogenic bacteria. Virulence 4:453–466. https://doi.org/10.4161/viru.25180.
18. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 176:649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.