Molecular mimicry of host proteins is a common strategy adopted by bacterial pathogens to interfere with and exploit host processes. Despite the availability of pathogen genomes, few studies have attempted to predict virulence-associated mimicry relationships directly from genomic sequences. Here, we analyzed the proteomes of 62 pathogenic and 66 non-pathogenic bacterial species, and screened for the top pathogen-specific or pathogen-enriched sequence similarities to human proteins. The screen identified approximately 100 potential mimicry relationships including well-characterized examples among the top-scoring hits (e.g., RalF, internalin, yopH, and others), with about 1/3 of predicted relationships supported by existing literature. Examination of homology to virulence factors, statistically enriched functions, and comparison with literature indicated that the detected mimics target key host structures (e.g., extracellular matrix, ECM) and pathways (e.g., cell adhesion, lipid metabolism, and immune signaling). The top-scoring and most widespread mimic pattern detected among pathogens consisted of elevated sequence similarities to ECM proteins including collagens and leucine-rich repeat proteins. Unexpectedly, analysis of the pathogen counterparts of these proteins revealed that they have evolved independently in different species of bacterial pathogens from separate repeat amplifications. Thus, our analysis provides evidence for two classes of mimics: complex proteins such as enzymes that have been acquired by eukaryote-to-pathogen horizontal transfer, and simpler repeat proteins that have independently evolved to mimic the host ECM. Ultimately, computational detection of pathogen-specific and pathogen-enriched similarities to host proteins provides insights into potentially novel mimicry-mediated virulence mechanisms of pathogenic bacteria.

**Introduction**

Molecular mimicry can be broadly defined as sequence or structural resemblance between microbial and host molecules. This has been studied extensively within the context of autoimmunity, whereby similarities between foreign and self molecules can lead to cross-reactive epitopes and ultimately autoimmune disease. However, there is a growing body of evidence that molecular mimicry of host proteins is a broader strategy adopted by bacterial pathogens to exploit and subvert host processes during infection and plays a role in a wide range of virulence pathways, including pathogen recognition and binding to human cells, evasion of the host immune response, and intracellular survival in host immune cells.

Mimics within pathogens are thought to originate through two evolutionary mechanisms. Pathogen genomes can obtain host genes directly through lateral transfer (reviewed in Koonin et al.13). Such cases frequently have detectable homology between pathogen and host proteins, a complex sequence or domain composition, and limited occurrence of the mimic in one or a small number of pathogenic species. For example, Coxiella burnetii, the causative agent of human Q fever, encodes two eukaryote-like sterol reductases. These mimics may play a role in formation of the cholesterol-rich Coxiella parasitophorous vacuole, which serves as a barrier to sequester nutrients and ions and also facilitates pathogen survival inside the host cell. These enzymes are extremely rare in prokaryotes and are thought to have arisen in Coxiella by lateral transfer from a eukaryotic source. A second possible mechanism is convergent or parallel evolution of a pathogenic protein toward resemblance of a host protein. Here, over time, co-evolutionary forces generate pathogen proteins that resemble host proteins structurally, or resemble smaller sequence fragments of host proteins, without homology between the pathogen and host proteins. For example, enterohemorrhagic Escherichia coli (EHEC) secretes a type III effector (EspF) into human host cells, which stimulates actin polymerization by interacting with host WASP proteins. Exploitation of host functions is achieved through subtle structural mimicry of the host WASP autoinhibitory helix, but there is no detectable sequence similarity between the two proteins. By stimulating actin polymerization, EspF mediates attachment of EHEC to host epithelial cells, which is critical to its virulence mechanism. Another example of convergent evolution is that of the Yersinia effector protein, invasin, which has evolved to mimic the integrin-binding surface of fibronectin. This surface mediates high affinity binding to β1 integrins on host M cells, which
mammalian collagens, from which we hypothesized that collagens may be an additional mimicry target of pathogenic bacteria, which could play a role in adhesion of pathogens to the host extracellular matrix. However, detection of sequence similarity between host and pathogenic proteins is by itself not indicative of mimicry or pathogen-specific exploitation of host functions.

Here, motivated by our previous work and the broad goal of detecting host-pathogen mimicry at a genomic scale, we performed an analysis of bacterial pathogen vs. non-pathogen proteomes and compared their similarities to the host (i.e., human) proteome. We screened for cases where the detected similarities to host proteins are pathogen-specific or are enriched in a variety of pathogenic species compared with non-pathogens, thus producing a list of candidate pathogen mimics and their human targets. It is important to note that while the analysis is based on the human proteome, host specificity of the predicted mimics toward human is not certain, and the predictions may reflect mimicry of proteins from alternative eukaryotic hosts. A similar approach has been used to screen for molecular mimicry candidates in protozoan parasites, yet to our knowledge a large-scale computational analysis has not been performed for human pathogenic bacteria.

Ultimately, our results provide additional evidence that collagens and extracellular matrix proteins in general are targets of mimicry by a range of pathogenic bacteria. Moreover, we report the unexpected result that such mimics have evolved independently in a range of bacterial pathogens through separate amplifications of short peptide repeats. In addition to extracellular matrix proteins, the screen predicted numerous known and potentially novel mimicry relationships that are candidates for future experimental investigation.

**Results**

Detecting pathogen mimics of human proteins by comparative proteome analysis. We developed and applied a computational pipeline (outlined in Fig. 1) to identify potential pathogen mimics of human (host) proteins using comparative proteome analysis. First, proteome sequence data was retrieved for human and 128 bacterial species including 62 bacterial pathogens of humans and 66 non-pathogens as annotated by the Comprehensive Microbial Resource (CMR). An all-by-all BLAST analysis was performed, in which all human proteins were searched against all bacterial proteomes, and hits with E-values < 1E−06 were collected and compared between the two groups (Fig. 1).

From the set of 33,340 human proteins, 9149 (27.4%) had BLAST matches with E < 1E−06 in one or more bacterial proteomes (Fig. S1A). For the average human protein, the frequency of pathogen vs. non-pathogen genomes containing a match, and the top pathogen vs. non-pathogen alignment score (bitscore) and E-value, was highly similar (Fig. S1B–E). Moreover, a larger fraction of the human proteome was similar to proteins in the non-pathogen set (Fig. S1A), which may reflect pathogen-associated genome reduction.

While molecular mimics may be encoded by pathogenic and non-pathogenic species, and serve different biological purposes...
Applying these filters resulted in a final list of 355 human proteins predicted as potential targets of pathogen molecular mimicry (Table S1A). These predicted mimicry relationships occurred in a small number of bacterial species (average = 3.2 species, 2.5%) compared with that observed for non-mimics (average = 36.8 species, 28.8%), which as expected appears to contain all ubiquitous proteins in the data set (2333 found in human and a majority of bacterial species, 208 conserved across all species).

The 355 human proteins were predicted targets of 231 total pathogen proteins from 53 pathogen species (Table S1B). We also removed redundancy within this set (see Methods) in order to generate a smaller list of 95 highly unique relationships (Table S1C) for subsequent analysis (see Methods). The top 25 most unique mimicry predictions are listed in Table 1. As explored in later sections, the top predictions include extracellular matrix proteins (collagen and leucine-rich repeat proteins), as well as several virulence factors and known examples of molecular mimicry (Table 1).

(e.g., virulence, mimicry of immune epitopes, and survival of commensal bacteria inside host), we aimed to identify mimics that may play specific roles in pathogen virulence. Here, we use the definition of pathogen mimics as bacterial pathogen-encoded proteins that share significant similarities with host proteins for the purposes of interacting or interfering with host machinery for the pathogen’s benefit.10,22 To identify such a subset, we further processed this list of bacteria–human protein similarities to identify those specific to or enriched in pathogens and diminished or completely absent in the non-pathogen group (controls), and thus indicative pathogen–host specificity and potential molecular mimicry. This involved applying the following criteria: that a hit is specific to or at least 2-fold enriched in pathogens and that a top pathogen hit has a greater alignment score than the corresponding top non-pathogen hit (bitscore difference >10, see Methods) (Fig. 1). These parameters capture a subset of potential mimics within the distributions shown in Figure S1C–E.

| No. | Pathogen protein | Human protein (NCBI gi number, gene name) | Description of human protein | No. human proteins in cluster | No. pathogen species in which mimic is found | Min. BLAST E-value |
|-----|-------------------|------------------------------------------|-----------------------------|-------------------------------|---------------------------------------------|-------------------|
| 1   | spr1403           | 29725624, COL23A1                        | collagen α-1(XXIII) chain   | 42                            | 7                                           | 6.00E–25          |
| 2   | BA_3841           | 11386161, COL9A2                        | collagen α-2(I) chain precursor | 1                             | 6                                           | 1.00E–11          |
| 3   | SpyM3_1561        | 122937309, LRRC4B                       | leucine-rich repeat-containing protein 4B precursor | 33                            | 5                                           | 5.00E–15          |
| 4   | SpyM3_0728        | 115527062, COL6A2                       | collagen α-2(VI) chain isoform 2C2 precursor | 3                             | 5                                           | 2.00E–11          |
| 5   | IpI2569           | 116256356, COL4A4                       | collagen α-4(IV) chain precursor | 5                             | 5                                           | 2.00E–12          |
| 6   | VIV_2676          | 7656971, N4BP2L2                        | NEDD4-binding protein 2-like 2 isoform 2 | 1                             | 5                                           | 3.00E–12          |
| 7   | IpI2411           | 4505983, PPFIA1                         | liprin-α-1 isoform b         | 34                            | 5                                           | 4.00E–10          |
| 8   | ML2177            | 31742508, UPP1                          | uridine phosphorylase 1     | 4                             | 3                                           | 4.00E–22          |
| 9   | CPE0622           | 119395714, FKTN                         | fukutin isoform a            | 2                             | 3                                           | 6.00E–16          |
| 10  | BA_2967           | 296434275, GPRASP2                      | G-protein coupled receptor-associated sorting protein 2 | 5                             | 3                                           | 3.00E–12          |
| 11  | nfa38270          | 22748883, TMEM68                        | transmembrane protein 68    | 1                             | 3                                           | 1.00E–15          |
| 12  | MT_0370           | 14149793, QRICH2                        | glutamine-rich protein 2     | 1                             | 3                                           | 4.00E–14          |
| 13  | CTC_02331         | 85386053, ATP6V0A4                      | V-type proton ATPase 116 kDa subunit a isoform 4 | 7                             | 3                                           | 1.00E–11          |
| 14  | ECH_0498          | 20270337, LEO1                          | RNA polymerase-associated protein LEO1 | 1                             | 3                                           | 3.00E–10          |
| 15  | nfa31870          | 21618331, CRAT                          | carnitine O-acetyltransferase precursor | 22                            | 2                                           | 1.00E–53          |
| 16  | CB1_1158          | 117414150, TM7SF2                       | delta(14)-sterol reductase   | 5                             | 2                                           | 2.00E–50          |
| 17  | IpI1919           | 51479145, ARFGEF1                       | brefeldin A-inhibited guanine nucleotide-exchange protein 1 | 19                            | 2                                           | 3.00E–34          |
| 18  | RP374             | 117606360, PSD3                        | PH and SEC7 domain-containing protein 3 isoform a | 1                             | 2                                           | 3.00E–15          |
| 19  | LMOF2365_0121     | 8922995, PLCXD1                        | PI-PLC X domain-containing protein 1 | 2                             | 2                                           | 8.00E–14          |
| 20  | BPSS0088          | 239747149, LOC100287429                | PREDICTED: zinc finger protein 84-like isoform 2 | 27                            | 2                                           | 4.00E–14          |
| 21  | APH_0455          | 132626688, MDC1                         | mediator of DNA damage checkpoint protein 1 | 1                             | 2                                           | 1.00E–13          |
| 22  | SpyM3_0116        | 4502313, ATP6V0C                       | V-type proton ATPase 16 kDa proteolipid subunit | 1                             | 2                                           | 4.00E–09          |
| 23  | TCE_0021          | 5174415, CEPT1                         | choline/ethanolaminephosphotransferase 1 | 2                             | 2                                           | 1.00E–12          |
| 24  | BCE_5203          | 310115369, LOC100294033                | PREDICTED: protein FAM115A-like isoform 2 | 4                             | 2                                           | 4.00E–11          |
| 25  | TP_0671           | 50726996, CHPT1                         | cholinephosphotransferase 1  | 1                             | 2                                           | 3.00E–11          |

All were specific to human pathogenic bacteria and did not appear in the non-pathogen species data set.
Known mimics. Prediction #3 corresponds to detected similarity between a human leucine-rich repeat protein (gi 122937309) and a leucine-rich repeat protein from Streptococcus pyogenes (SpyM3_1561) (Table 1) as well as Listeria internalin virulence factors (e.g., lmo0801 and LMOh7858_0295 [Table S1A]). This reflects the established example of human LRR mimicry by internalin.7,19,20 Predictions #17 and #18 are detected relationships between two human guanine exchange factor (GEF) containing proteins and the pathogen proteins RalF (Legionella pneumophila) and sec7 (Rickettsia prowazekii), and thus reflect GEF mimicry.18 Prediction #31 corresponds to detected similarity between a human tyrosine phosphatase (gi 108802617) and yopH from Yersinia pseudotuberculosis, another known case of pathogen-specific molecular mimicry.7,8

Homology to virulence factors. Sixty-one of 95 predicted mimics (non-redundant set) were detected as homologous (BLAST E-value < 1E−06) to known virulence factors from the MvirDB,27 suggesting possible roles in mimicry-mediated virulence. These include the Legionella LepB effector (prediction #7, similar to numerous human coiled coil proteins) indicating possible functions in regulation of secretory traffic, Listeria PlcA (prediction #19, similar to human PI-PLC), and Helicobacter and human fucosyltransferase (prediction #68, FucT) (Table 1; Table S1). Helicobacter FucT is another known case of mimicry as its produces the lipopolysaccharide component Lewis X trisaccharide, which is thought to mimic host sugars to escape immune detection.28

Enriched virulence-related functions. Gene enrichment analysis was then performed using DAVID29 to identify statistically enriched functions and protein families among the full set of predicted human mimicry targets (Table 2; Table S2A). The top four
The top two functional enrichments for predicted mimics from phytopathogens were "apoptosis" ($P = 5.54\times10^{-27}$) and "programmed cell death" ($P = 1.42\times10^{-25}$). Indeed, induction of apoptosis is a known virulence mechanism of plant pathogens, 30 suggesting that the approach may be applicable to different host-pathogen relationships.

**Comparison with literature.** We then analyzed the top 95 unique detected mimicry relationships (non-redundant set) for additional literature supporting potential roles in virulence. Thirty-three of these relationships (~35%) were found to be highly suggestive of mimicry-mediated virulence mechanisms with support from existing literature (Table 3). Known mechanisms of pathogen exploitation of host cells that are also reflected throughout the detected relationships include pathogen-specific enriched function terms were: “Extracellular matrix” (Benjamini $P = 3.39\times10^{-42}$), collagen ($P = 3.11\times10^{-29}$), “Extracellular matrix structural protein” ($P = 2.56\times10^{-24}$), and “ARF guanyl-nucleotide exchange factor activity” ($P = 2.11\times10^{-19}$) (Table S2A). Other intriguing top enriched terms included “O-acyltransferase activity” ($P = 2.08\times10^{-06}$), “cell adhesion” ($P = 2.95\times10^{-06}$), and “inflammation mediated by chemokine and cytokine signaling pathway” ($P = 6.84\times10^{-03}$). Terms that were highly ranked and functionally relevant but of weaker statistical significance included “1k kinase/NFkB cascade” ($P = 1.18\times10^{-02}$), “lysosome” ($P = 1.38\times10^{-02}$), “Toll-like receptor pathway” ($P = 5.15\times10^{-02}$), and interaction with host ($P = 9.30\times10^{-02}$). Thus, the top enriched terms appear to be consistent with virulence related mechanisms of pathogenic bacteria. As a comparison, we also applied the same analysis to phytopathogens (578 mimics detected, data not shown), which identified different enrichment categories. The top two functional enrichments for predicted mimics from phytopathogens were “apoptosis” ($P = 5.54\times10^{-27}$) and “programmed cell death” ($P = 1.42\times10^{-25}$). Indeed, induction of apoptosis is a known virulence mechanism of plant pathogens, 30 suggesting that the approach may be applicable to different host-pathogen relationships.

**Comparison with literature.** We then analyzed the top 95 unique detected mimicry relationships (non-redundant set) for additional literature supporting potential roles in virulence. Thirty-three of these relationships (~35%) were found to be highly suggestive of mimicry-mediated virulence mechanisms with support from existing literature (Table 3). Known mechanisms of pathogen exploitation of host cells that are also reflected throughout the detected relationships include pathogen-specific

---

**Figure 3.** Top pathogen vs. non-pathogen protein similarities to a selected set of predicted human mimicry targets. Predicted human mimicry targets were selected from the top 25 detected relationships (Table 1), and the top BLAST matches by bitscore (x-axis) in pathogen vs. non-pathogen proteomes (frequency on y-axis) have been plotted. In each case, it can be seen that a subset of pathogen proteomes encode putative mimics that exhibit much greater similarities to human proteins than similarities found in non-pathogen proteins. A selected portion of the alignment is shown for the top-scoring pathogen mimic in each case. See Data File S1 for additional details regarding pairwise alignments.
modulation of host lipid metabolism, modulation of the host nucleotide pool and induction of apoptosis, and cell adhesion. The remaining 65 cases may represent previously uncharacterized mimicry relationships, and provide numerous targets for investigation of pathogenicity mechanisms.

Numerous detected mimicry relationships and function enrichments (Table S2) are linked to lipid metabolism. Putative mimics affecting host lipid metabolism include detected pathogen counterparts of human cholinephosphotransferase, PCTP like protein, fukutin, acyltransferases, phospholipase A2, carnitine O-palmitoyltransferase, and sterol reductase (Table 3). For example, *Mycobacterium pneumonia* is known to incorporate host lipids (e.g., phosphatidylcholine), and has been shown to modulate host sphingolipid metabolism. A human-like carnitine O-palmitoyltransferase (CPT) was detected in *Mycobacterium pneumonia*, which may play a role in these functions. As another example, part of the virulence mechanism of *Bacillus anthracis* involves escape from host macrophages, but the mechanism is largely unknown. A human-like phospholipase A2 was detected in *Bacillus anthracis* (BA_3805), which, in other pathogens, has been shown to play a role in phagosome escape as well as entry and lysis of host cells.

Cells damaged by invading pathogens release nucleotides into the external environment, which act as “danger signals” and stimulate pathogen-killing immune responses, or induce apoptosis. Detected mimicry relationships indicative of host nucleotide pool modulation include the previously studied *Legionella* lpl1869 with detected similarities to the human NTPDase CD39, and proteins similar to human P-loop NTPases and ATPases from *Vibrio* and other species (Table 3). Some of these putative mimics may play roles in inhibition of apoptotic signaling. *Rickettsia* RC0370 appears to mimic the P-loop NTPase domain of human NACHT proteins involved in immune signaling and apoptosis. Furthermore, adenylate kinase (AK) in *Pseudomonas aeruginosa* has been shown to act as a virulence factor regulating external ATP-dependent macrophage cell death, and the analysis identified several human-like AKs in a range of pathogenic species that represent novel candidate virulence factors (Table 3; Table S1A). Another example of potential disruption of apoptotic signaling involves detected between the *Anaplasma* protein, APH_0455, and the human protein NFBD1/MDC1, with the two proteins sharing a repetitive motif (QPSTXDXQPXT, see Data File S1 for BLAST alignments). Interestingly, *Anaplasma phagocytophilum* is known to prevent apoptosis in neutrophils, and MDC1 is known to have anti-apoptotic activity through inhibition of p53 phosphorylation.

Before pathogens can appropriate host pathways, they must adhere to and invade host cells. This was the strongest detected mimicry function in our data set, as both the top-scoring putative mimics (Table 1) and enriched functions (Table 2) relate to the extracellular matrix and its components (specifically, collagens and leucine-rich repeat proteins). Both collagens and LRRs have been implicated in virulence and play a role in host cell adherence and invasion, and are the focus of the following section. Other detected pathogen proteins with possible roles in host cell adhesion include a *Bacillus* adhesin (BA_0871), which shares a unique repetitive TEKP motif (Data File S1) with human zonadhesin, and *Treponema* proteins similar to human ankyrins that may interact with the host cytoskeleton.

**Extracellular matrix proteins predicted as major targets of molecular mimicry.** Collagens. Prediction #1 (Table 1) and function enrichment #1 and #2 (Table 2) relate to detected similarities between human collagens and collagen-like proteins in pathogenic bacteria. Figures 2 and 3 highlight collagens as clear outliers in the E-value and bitscore distributions, similar to that of the known mimicry relationship, RalF-GEF. Another collagen mimicry function in our data set, as both the top-scoring putative mimic (Table 1) and enriched functions (Table 2) relate to detected similarities between human collagens and collagen-like proteins in pathogenic bacteria. Figures 2 and 3 highlight collagens as clear outliers in the E-value and bitscore distributions, similar to that of the known mimicry relationship, RalF-GEF. Another collagen mimicry relationship was also the most abundant pathogen-specific pattern detected, occurring in 7 pathogens and 0 non-pathogens (Table 1).

The top scoring predicted collagen mimic was *spr1403* (PclA) from *Streptococcus pneumoniae*, a protein that has previously been shown to contribute to host-cell adherence and invasion.
### Table 3. Predicted mimicry candidates in human pathogenic bacteria and potential roles in virulence

| Pred. # | Human protein/s (putative mimic/s) | Pathogen protein/s (internalin) | Pathogen species | Virulence mechanism (known or proposed) |
|---------|-----------------------------------|---------------------------------|-----------------|-----------------------------------------|
| 1       | Collagen                          | spri1403 + 11 others            | Diverse         | Involvement in host-cell adherence/invasion<sup>15</sup> |
| 3, 50, 60 | Leucine-rich repeat proteins      | SpyM3_1561, Lmo0801 (internalin), + 7 others; RC0370; RC0830 | Diverse         | Internalin-related; adhesion and invasion of host epithelia<sup>12</sup> Possible role of RC0370 in modulation of host immune and apoptotic signaling<sup>35,34</sup> |
| 6, 74, 79 | P-loop NTPases and ATPases        | VC_1610, VPA1750, VP1457 + 27 others | H. pylori, L. interrogans, Vibrio spp., and 9 others | Possible modulation of host external ATP pool and macrophage cell death<sup>35,37</sup> |
| 7       | Coiled-coil proteins              | lpl2411 (lepB) + 5 others       | E. faecalis, E. coli, L. pneumophila, P. multocida, P. aeruginosa, S. pneumoniae | Disruption of vesicular protein-trafficking<sup>38,39</sup> |
| 8       | Uridine phosphorylase             | ML2177 + 5 others               | E. coli, H. influenzae, M. leprae, P. multocida, S. flexneri | Use of host uridine for pyrimidine scavenging by M. leprae (cannot directly take up nucleosides and bases)<sup>40</sup> |
| 9       | Fukutin                           | LicD proteins (CPE0622 + 4 others) | C. perfringens, Rickettsia spp. | *Modification of cell-surface glyproteins or glycolipids; transfer of phosphorylcholine residues<sup>41</sup> |
| 11, 54  | Acdytransferases                  | MT_1971, nfa38270 + 2 others    | M. avium, M. tuberculosis, N. farcinica | Possible involvement in lipid biosynthesis from host-derived precursors; possible use as energy store and/or virulence-related, immunomodulatory lipids<sup>42,43</sup> |
| 15      | Carnitine palmitoyl-transferase   | MPN114 nfa31870                 | M. pneumoniae, N. farcinica | Possible biosynthesis of or modification of host phosphatidylcholine<sup>44</sup> or modulation of host sphingolipid metabolism<sup>45</sup> |
| 16      | Delta(14)-sterol reductase        | CBU_1158                        | C. burnetii      | Cholesterol metabolism for structural and/or signaling roles during parasitophorous vacuole formation<sup>21,22</sup> |
| 17, 18  | ADP ribosylation factor (ARF) guanine exchange factor | lpl1919 (RalF), RP374 (sec7) | L. pneumophila, R. prowazekii | RalF recruits ARF GTPases for Legionella phagosome formation<sup>14</sup> |
| 19      | Phosphatidyl-inositol-specific phospholipase C | LMOF2365_0212 + 6 others | B. cereus, L. monocytogenes, S. aureus | Escape from primary vacuole<sup>66</sup> |
| 21      | Mediator of DNA damage checkpoint protein 1 (MDC1) | APH_0455                       | Anaplasma phagocytophilum | Possible disruption of host cell apoptotic signaling in neutrophils<sup>47,60</sup> |
| 23, 25  | Choline/ethanolamine-phosphotransferase | TDE_0021, TP_0671 | Treponema spp. | Possible production of phosphatidylcholine; host-phospholipid mimicry<sup>60,52</sup> |
| 24      | Predicted FAM115A-like proteins   | BCE_5203 + 2 others (enhancins) | B. anthracis, B. cereus | Host–protein (e.g., mucin) proteolysis<sup>51</sup> |
| 30      | Ankyrin                           | TP_0835 + 1 other               | Treponema spp.   | Possible interaction with host cytoskeleton as previously hypothesized<sup>52</sup> Ankyrin-like effectors have been identified in other pathogens with other roles in virulence<sup>53</sup> |
| 31      | Tyrosine-protein phosphatase non-receptor type 20 | YPCD1.67c (yopH) + 3 others | Y. pestis, Y. pseudotuberculosis | Disruption of host macrophage signal transduction pathway<sup>14</sup> |
| 38      | Paraoxonase                       | LA0399                          | Leptospira interrogans | Possible loss of hemostasis through hydrolysis of a platelet-activating factor<sup>27,25</sup> |
| 39      | Periaxin, apoB (leucine-proline rich repeats) | MT_1796, Rv1753c (PPE family proteins) | Mycobacterium tuberculosis | Targeting of DRP2-dystroglycan complex; possible interaction with host lipids;<sup>29,59</sup> may also facilitate survival in host macrophages |
Numerous detected proteins may play similar roles in pathogenesis, including: BA_3841, BCE_1581, BC_3345, CPE0955, CPR_1027, ECs1228, EF_2090, SPy1983, SpyM3_0738, Z1483, lpl2569, BC_2381, CPE_1202, BCE_3739, ECs2941, and Z2340. Another putative collagen mimic detected by the analysis (Spy1983 [SclA] from *Streptococcus pyogenes*) has recently been demonstrated to act as an adhesin during pathogenesis of streptococcal infection.74 These pathogen-specific collagen-like proteins (CLPs) are significantly more human-like than CLPs found in other bacteria, as shown by the bitscore distributions (Fig. 3). To investigate this further, we analyzed motif content for the predicted subset of collagen mimics found in pathogens, vs. all other CLPs including those found in non-pathogens. The human-like, pathogen-specific subset of CLPs (putative collagen mimics) were found to have significantly more tetrapeptides in common with human collagens than non-pathogen CLPs (Fig. S2). Many of these tetrapeptides contain “GP” motifs characteristic of human collagen sequences. For example, the average number of GP motifs per sequence length is 4.95% for the putative collagen mimics, 4.86% for human collagens, and only 0.26% for predicted non-mimics. Thus, the detected sequence similarities are due to similarities in peptide composition rather than sequence or statistical artifacts.

**Leucine-rich repeat proteins.** In mammals, leucine-rich repeat proteins are a second class of proteins abundant in the extracellular matrix, where they function in cell growth, adhesion, migration, and bind with other ECM components including collagens.79 As mentioned above, prediction #3 involves detected similarities between human LRRs and internalin-related factors (e.g., Imo0801, LMOh7858_0295), which are known to function in adherence and invasion of host cells (Table 1; Table S1A). LRR-containing proteins such as NOD-like receptors and Toll-like receptors also serve as important pathogen-detection molecules, recognizing key pathogen-associated molecular patterns (PAMPs).20,76 Both Toll-like receptors (prediction #27) and NOD-like receptors (prediction #50 and #60) were detected as potential targets of LRR mimicry (Table S1C).

**Evolution of ECM mimics via independent repeat events.** As the top scoring pattern overall, we analyzed the detected similarities between human extracellular matrix proteins and putative bacterial pathogen mimics. The two detected relationships are largely distinct according to their taxonomic distribution. Detected collagen mimics were associated predominantly with *Firmicutes* pathogens, and LRR mimics were identified in pathogens from a range of phyla also including *Spirochaetes* and *Bacteroidetes*. One species (*Streptococcus pyogenes*) appears to encode both types of mimicry proteins.

To investigate how the set of putative collagen mimics have evolved in pathogenic bacteria, we analyzed and compared the sequence architecture of predicted ECM mimics from different species. Despite possessing a common repetitive collagen architecture, we found that collagen-like repeats from different bacterial pathogenic species have distinct repeat architectures, indicative of separate evolutionary origins. To demonstrate this, CLPs from different pathogens were divided into their peptide repeats, which were aligned and used to create sequence logos (Fig. 4A) and phylogenetic trees (Fig. 5, left). While the collagen-like repetitive pattern (GXXx) is common to all detected CLPs, the progenitor repeat sequences are different, and the repeat lengths are also variable (Fig. 4A). Phylogenetic trees were constructed based on an alignment of pathogen collagen-like repeats as well as a top-aligning human repeat to the set of repeats within each pathogen protein (see Methods). As revealed by the sequence logos and the tree, the detected collagen mimics from *Streptococcus pneumoniae* (spr1403), *Streptococcus pyogenes* (SpyM3_0738), *Clostridium perfringens* (CPR_1027), *Bacillus anthracis* (BA_3841), and *Legionella pneumophila* (lpl2569) appear to have independently evolved their similarity to human

---

**Table 3. Predicted mimicry candidates in human pathogenic bacteria and potential roles in virulence (continued)**

| Pred. # | Human protein/s | Pathogen protein/s (putative mimic/s) | Pathogen species | Virulence mechanism (known or proposed) |
|--------|-----------------|---------------------------------------|-----------------|----------------------------------------|
| 64     | Serine protease  | VC_1200                               | *Vibrio cholerae* | May process cholera toxin A (CT A) in the host63 |
| 68     | Fucosyltransferase | HP0379 + 2 others                     | *H. hepaticus, H. pylori* | Production of Lewis x trisaccharide; mimicry of host cell-surface sugars; immune evasion79 |
| 69     | Phospholipase A2 | BA_3805                               | *Bacillus anthracis* | Possible role in invasion/entry of host cells, escape from phagosome, and cell lysis64,65 |
| 70     | PCTP like protein | VV2_0046, VVA0553                     | *V. vulnificus* | Possible modulation of host lipid (e.g., cholesterol) metabolism66 |
| 76     | Adenylate kinase | CPE2384 + 7 others                    | Diverse         | An adenylate kinase toxin from *P. aeruginosa* plays a role in macrophage cell death67 |

---

**Table 1**

| Pathogen species | Virulence mechanism (known or proposed) |
|------------------|----------------------------------------|
| *Legionella pneumophila* | Replication in host macrophages; manipulation of host pathways regulated by CD39; Modulation of host NTPs and NDPs64,65 |
| *Bacillus anthracis* | Acts as an adhesin that mediates cell attachment to collagen79 |
| *Rickettsia prowazekii* | Possible involvement in biosynthesis of lipopolysaccharide62 |
| *Vibrio cholerae* | May process cholera toxin A (CT A) in the host63 |
| *H. hepaticus, H. pylori* | Production of Lewis x trisaccharide; mimicry of host cell-surface sugars; immune evasion79 |
| *Bacillus anthracis* | Possible role in invasion/entry of host cells, escape from phagosome, and cell lysis64,65 |
| *V. vulnificus* | Possible modulation of host lipid (e.g., cholesterol) metabolism66 |
| Diverse | An adenylate kinase toxin from *P. aeruginosa* plays a role in macrophage cell death67 |
different repeat architectures (Fig. 4B) and form unique clusters in the phylogenetic tree (Fig. 5, right), implying separate origins via independent repeat amplifications. As described in Ward et al., the first residues LxxLxLxxNx of leucine-rich repeats correspond to a conserved, interior portion of the LRR structure, while the remaining sequence encodes variable residues on the other face of the LRR structure. Interestingly, variable segments from pathogenic LRRs aligned strongly with variable segments from human NOD-like receptors (e.g., NLRC3) (Fig. 4C). Interestingly, NLRC3 was the top BLAST match to lpl1579 in the human proteome, and it has been demonstrated to have inhibitory effects on T-cell function. It is important to note that a BLAST search of lpl1579 across all eukaryotes identified predicted mimics from a range of mammals, and so the human NLRC3 is not necessarily the highest scoring hit, which is a general statement that also applies to other detected mimicry relationships.

Further phylogenetic analysis would be needed to characterize collagen via separate repeat amplifications (Fig. 4A and 5, left). Moreover, different human repeats can be found that cluster specifically with each pathogen repeat class (Fig. 5, left), indicating that different pathogen repeats may not only be mimicking different human proteins, but may be derived from different host peptides.

Consistent with the idea that different pathogen CLPs have evolved independently from each other, CLPs in bacteria appear to exhibit a scattered phylogenetic distribution. For example, while CLPs were predominantly detected in pathogens from the Firmicutes phylum, they also exist in the highly pathogenic O157:H7 strain of Escherichia coli (Table S1A) as well as Legionella pneumophila but not other Gammaproteobacteria. Interestingly, while present in E. coli O157:H7, the collagen-like proteins are absent in the uropathogenic E. coli CFT073 strain, and the non-pathogenic K1 strain.

As with the putative collagen mimics, we then analyzed the repeat architecture of the detected LRR mimics, and aligned the sequence logos in the form of the general LRR sequence pattern (based on Ward et al.) (Fig. 4B), and generated phylogenetic trees as described above (Fig. 5, right). Like the detected collagen mimics, LRR mimics from different pathogens exhibit different repeat architectures (Fig. 4B) and form unique clusters in the phylogenetic tree (Fig. 5, right), implying separate origins via independent repeat amplifications.

![Figure 4](image)

**Figure 4.** Independent evolution of ECM mimics from separate repeat amplifications. High-scoring collagen-like (A) and leucine-rich repeat (B) protein mimics were selected and divided into their constituent protein repeats, which were aligned and used to generate sequence logos. Differences between the sequence logos of each repetitive protein suggest evolution from separate progenitor peptides and repeat amplifications. (C) An example demonstrating similarity of leucine-rich repeat sequence conservation patterns between a human NOD-like receptor (NLRC3) and a predicted mimicry candidate (lpl1579) from Legionella pneumophila. The detected level of sequence similarity between these two proteins is far above that observed in non-pathogens (blue) and other pathogens (red) as indicated by the BLAST bitscore distribution (left panel).
As revealed by gene-enrichment analyses performed on both sets separately (Table S2B and C), the repetitive class consisting of CLPs, LRRs, and other candidates, are responsible for the enrichments related to extracellular matrix mimicry and cell adhesion (Table S2B). These terms were not significantly enriched among the non-repetitive class (Table S2C). Conversely, non-repetitive mimics were significantly enriched in terms related to enzymatic function such as "catalytic activity" and "lipid-metabolism," but these were not enriched among the repetitive-class. These results, combined with sequence (Fig. 4) and phylogenetic analyses (Fig. 5), are consistent with the idea that non-repetitive mimics with complex sequence composition are associated with enzymatic modulation of host functions and have likely been acquired in pathogens by horizontal transfer, while repetitive mimics have evolved independently in pathogens to mimic repetitive host structural proteins involving in adherence and invasion of host cells.

Repetitive and non-repetitive mimics have different functions and modes of evolution. The cases described above (putative collagen and LRR mimics) appear to have evolved independently in pathogens to mimic the repetitive architecture of host proteins.

To investigate the sequence repetition of predicted mimics quantitatively, each region of the putative mimics with detected similarity to human proteins was analyzed and divided into putative sequence repeats using the RADAR repeat prediction algorithm.80 Repeat proteins were found to make up a considerable number of detected mimicry relationships, as 34/95 and 207/306 detected mimics were identified as repetitive (containing three or more predicted repeats) in the non-redundant and full set of predictions, respectively (Table S1).

As revealed by gene-enrichment analyses performed on both sets separately (Table S2B and C), the repetitive class consisting of CLPs, LRRs, and other candidates, are responsible for the enrichments related to extracellular matrix mimicry and cell adhesion (Table S2B). These terms were not significantly enriched among the non-repetitive class (Table S2C). Conversely, non-repetitive mimics were significantly enriched in terms related to enzymatic function such as "catalytic activity" and "lipid-metabolism," but these were not enriched among the repetitive-class. These results, combined with sequence (Fig. 4) and phylogenetic analyses (Fig. 5), are consistent with the idea that non-repetitive mimics with complex sequence composition are associated with enzymatic modulation of host functions and have likely been acquired in pathogens by horizontal transfer, while repetitive mimics have evolved independently in pathogens to mimic repetitive host structural proteins involving in adherence and invasion of host cells.

Figure 5. Phylogenetic trees of bacterial pathogen encoded collagen-like repeats (left) and leucine-rich repeats (right) from Figure 4. The repeats are colored in the tree according to their parent protein. Top-aligning repeats from human proteins have also been included and are colored light green. Repeats cluster predominantly by protein of origin, suggesting that different pathogen repeat proteins have evolved by independent repeat amplifications. Interestingly, the pathogen repeat classes generally cluster with a specific human repeat, suggesting that ancestral progenitor repeats may be host-derived.

the evolutionary origins of each detected mimic to verify possible eukaryotic host-bacteria horizontal transfer79 from human or non-human host species.

As revealed by gene-enrichment analyses performed on both sets separately (Table S2B and C), the repetitive class consisting of CLPs, LRRs, and other candidates, are responsible for the enrichments related to extracellular matrix mimicry and cell adhesion (Table S2B). These terms were not significantly enriched among the non-repetitive class (Table S2C). Conversely, non-repetitive mimics were significantly enriched in terms related to enzymatic function such as "catalytic activity" and "lipid-metabolism," but these were not enriched among the repetitive-class. These results, combined with sequence (Fig. 4) and phylogenetic analyses (Fig. 5), are consistent with the idea that non-repetitive mimics with complex sequence composition are associated with enzymatic modulation of host functions and have likely been acquired in pathogens by horizontal transfer, while repetitive mimics have evolved independently in pathogens to mimic repetitive host structural proteins involving in adherence and invasion of host cells.
Another interesting example of the latter is a detected potential mimicry relationship between human protein periaxin and mycobacterial PPE family virulence factors (Rv1918c) (Table 3). Mycobacteria such as *M. leprae* specifically invade human Schwann cells through an interaction with the dystroglycan complex.26,38 Interestingly, human periaxin is a Schwann cell-specific protein that is critical to formation of the dystroglycan-complex. However, as with LRRs and collagens, the detected similarity is not due to homology but rather, a repetitive proline-rich composition that has independently evolved in both proteins (see Data File S1). It is possible that the repetitive proline-rich architecture of PPE proteins may facilitate an interaction with host dystroglycan-complex as is the case for periaxin, or that they may act as membrane-interacting lipoproteins.

**Discussion**

The results of this work suggest that comparison of host–bacteria proteome similarities is sufficient to detect a subset of pathogen mimics that function in bacterial virulence. Our approach identified known examples of mimicry, virulence factors, and potential novel candidates with roles in modulation and exploitation of host functions.

According to a recent study,81 all human proteins possess motifs also present in bacterial proteins. The same study also found no observable difference in overall bacteria–human peptide similarities between pathogenic and non-pathogenic species. Our results show that while overall sequence similarity to human proteins is not significantly enriched in pathogenic vs. non-pathogenic bacteria, there are detectable pathogen-specific or pathogen-enriched similarities to host proteins in key functional pathways related to virulence. These identified pathways and components, including the extracellular matrix, lipid metabolism, and immune signaling, are known targets of exploitation by bacterial pathogens.

**Evolutionary origins of mimicry proteins in pathogenic bacteria.** As discussed in previous literature,7,14 two evolutionary mechanisms are likely responsible for detected sequence mimicry between pathogen and host proteins: direct homology due to lateral transfer of the eukaryotic proteins to one or few bacteria, or similarities due to independent evolutionary processes (convergent or parallel evolution) (Fig. S3). In this work, we identified two sequence classes of detected mimics (non-repetitive and repetitive), which likely fall into these two evolutionary categories.

A new insight revealed by this work relates to the independent evolutionary processes by which pathogen mimics can originate. One documented mechanism underlying convergent evolution of host mimicry is independent origin of a binding surfaces or motif in a pathogen protein that displays no detectable homology with its host counterpart.7,14 Our work provides strong support for an additional mechanism, mimicry of host repetitive proteins via independently evolved peptide repeats. In this scenario, separate progenitor repeats in the pathogen genome are amplified to result in repeat proteins that share the same repetitive architecture but with different sequences for each repeat unit (Fig. 4; Fig. S3).

This is similar to what has been observed for β-trefoil proteins, which also include virulence-associated subfamilies (i.e., ricin toxins) that have undergone separate repeat amplifications while maintaining the same overall structure.82

These detected similarities do not imply overall homology between the full proteins but rather are due to similarity of repetitive architecture. Repetitive host proteins such as collagens, leucine-rich repeat proteins, and adhesins represent ideal targets for this evolutionary mechanism of pathogen mimicry, while complex proteins such as enzymes are not.

Interestingly, not only do the human and pathogen counterparts of these proteins appear to have evolved independently, but repeat amplifications appear to have occurred independently in different pathogenic species. While this may be indicative of convergent evolution, it is also possible that the pathogen proteins evolved by tandem duplications of an original peptide fragment that itself was acquired from (and thus related to) a host fragment, but then evolved a novel composition through independent evolutionary processes. This is similar in some respects to the results of a recent study that identified a recurring phenomenon whereby host-derived proteins in viruses had subsequently converged toward simpler domain architectures.83

In either case, the enrichment of repetitive ECM mimics in pathogens is likely due to convergent or parallel evolutionary processes that are driven by pathogen-specific selective pressures.

These predictions provide starting points for future experimental work characterizing the biological role of predicted pathogen mimics. We have analyzed only a subset of this spectrum, and future work expanding this analysis, and also evaluating the host-species specificity of this approach will be useful. Thus, future use of alternative classification schemes, improved motif detection techniques and structural bioinformatics may provide added sensitivity.

For instance, the classification scheme that is the basis of our comparative approach separates human pathogenic vs. non-pathogenic bacteria. Although this classification is somewhat arbitrary, the putative mimics detected using this scheme likely play virulence associated roles, and it was an objective of this analysis to find such pathogen-associated mimics. However, it is also important to note that other classes of mimics exist that may not be detected by our pathogen/non-pathogen comparison. These include mimics that are not directly involved in virulence but might still play a role in persistence of commensal bacteria inside the host, or have other effects such as the extensively studied role of peptide mimicry in generation of autoimmune disease.1,2 In the case of immune epitope mimicry, small regions of sequence similarity and not overall homology may be sufficient to elicit molecular mimicry,6,5 which would also not be identified using a standard homology detection approach. Our work thus complements previous work,9 which has focused on such cases of immune epitope mimicry. Finally, pathogen mimics of host proteins may have diverged beyond the point of recognizable sequence homology, but be detectable at the level of overall structural similarity.

Ultimately, to extend computational analysis of host-pathogen molecular mimicry, it may be useful to analyze mimicry with...
Computational screen for mimicry candidates using all-by-all BLAST analysis. Protein sequence data sets for human as well as 163 bacterial genomes (see Table S1 for a complete list) were retrieved from the NCBI (RefSeq human protein database build 36 [37742 proteins]) and the Comprehensive Microbial Resource29 at TIGR/JVCI (http://cmr.jcvi.org). To reduce species redundancy, only one proteome per species was kept and the rest were removed. This step removed 35 species, leaving 62 pathogens and 66 non-pathogens.

An all-by-all BLAST analysis was conducted using BLAST v.2.2.16, in which each human protein was used as an individual query in a separate BLAST search of each individual organism’s protein database. Default BLAST parameters were used with “composition-based statistics” to correct for potential compositional bias. A BLAST E-value cutoff of 1E−06 was used to identify putative matches, from which a presence/absence matrix was constructed. BLAST E-values, bitscores, and top pathogen protein matches were recorded for each cell of the matrix. To remove genome/species redundancy, only one genome per species was kept (randomly assigned) and the remaining genomes were removed.

Each human protein (i) was then scored using the fraction of pathogen species with a hit detected by blast (P) divided by the fraction of non-pathogens (NP) containing a hit. Potential protein mimics were selected based on rarity in non-pathogens, enrichment in pathogens, and greater similarity between pathogen and human proteins. The specific criteria were hits found in less than five non-pathogens, P/NP ratio greater than 2, and top pathogen BLAST hit having a bitscore greater than 10 above that of the top non-pathogen hit. A bitscore difference of 10 was chosen to result in a final list equivalent to roughly 1% of the human proteome (99% percentile).

The full list of detected mimicry candidates is shown in Table S1. A smaller list of the most unique relationships was also generated by including only the top human matches to each unique pathogen protein. This generated a smaller set of 95 unique mimicry relationships (Table S1C), the top 25 of which are listed in Table 1.

Phytopathogen mimicry analysis. As a control, and test for generality, we applied the approach to a different host-pathogen relationship (plant/phytopathogen). The same approach was used as described above with Arabidopsis thaliana used as the host proteome, and the following species were defined as phytopathogens: Phytoplasma asteris, Agrobacterium tumefaciens, Ralstonia solanacearum, Pseudomonas syringae, Xanthomonas axonopodis, and Xylella fastidiosa. Xanthomonas campesiris was removed due to redundancy with X. axonopodis. In total, this data set contained six phytopathogens and 16 non-pathogens.

Homology to known virulence factors. All candidate pathogen mimics were searched against the MvirDB30 database of known virulence factors. Hits were again defined as BLAST matches with E < 1E−06.

Gene enrichment analysis. Gene enrichment analysis for overrepresented functions was performed using DAVID.29 The following eight ontologies were used: GOTERM_BP_ALL, GO_TERM_CC_ALL, GOTERM_MF_ALL, PANTHER_BP_ALL, PANTHER_MF_ALL, BIOCARTA, KEGG_PATHWAY, and PANTHER_PATHWAY. The default parameter value of EASE = 0.1 was used.

Detection and analysis of protein repeat sequences. For each putative mimic, the BLAST result was parsed to extract the sequence region detected as similar to the host protein. Repeats were predicted for these regions using RADAR80 run with default parameters. Repetitive mimics were defined as those containing greater than three predicted repeats, all of which had total scores > 90.

For the detected collagen-like and LRR mimics, the detected repeats were aligned, adjusted where necessary to a common length, and sequence logos were generated using seqlogo with default parameters (http://weblogo.berkeley.edu).

For each set of repeats from a pathogen protein, the consensus sequence was used as a query to identify the best-aligning repeat from a corresponding human protein using SSEARCH.84 Collagen-like and leucine-rich repeats were then aligned separately along with their respective human sequences and a phylogenetic tree was generated using parsimony using Phylip.85 Branch lengths were estimated by maximum likelihood using Fasttree86 with the JTT model and CAT approximation with 20 rate categories.

Sequence composition analysis. The compseq program within the EMBOSS suite87 (version 6.3.1) was used to compute motif frequencies for human collagens, the putative collagen mimics and non-mimics.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
This work was supported by the National Science and Engineering Research Council of Canada (NSERC) through grants to BJM (NSERC Discovery Grant) and ACD (NSERC PDF). We thank Trevor Charles for insightful discussions.

Supplemental Materials
Supplemental materials may be found here: www.landesbioscience.com/journals/virulence/article/25180
null