Szkoła Główna Gospodarstwa Wiejskiego
w Warszawie
Instytut Biologii

Marianna Krysińska

Ewolucja domen ADP-rybozylotransferaz i kinaz oraz bioinformatyczne poszukiwanie nowych rodzin

Evolution of ADP-ribosyltransferase and kinase domains and bioinformatic search for new families

Rozprawa doktorska
Doctoral thesis

Rozprawa doktorska wykonana pod kierunkiem

dr. hab. Krzysztofa Pawłowskiego
University of Texas Southwestern Medical Center, Dallas, TX, USA
Howard Hughes Medical Institute, Dallas, TX, USA

Promotor pomocniczy:

dr hab. Małgorzata Dudkiewicz
Katedra Biochemii i Mikrobiologii, Instytut Biologii,
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

Warszawa 2025

śp. Jerzemu Rębelskiemu
za światło zapalone na początku drogi

Moim promotorom
dr. hab. Krzysztofowi Pawłowskiemu
za przestrzeń do popełniania błędów i podejmowania decyzji,
za asekurację każdego pomysłu i gotowość przyjścia z pomocą,
za nauczenie, że odpowiedzią na większość pytań jest czas,
za uświadomienie, że można czegoś nie wiedzieć i świat się od tego nie kończy,
za nieustanne wsparcie w każdym aspekcie,
za przyszłość.
Nie mogłabym wymarzyć sobie lepszego promotora!

dr hab. Małgorzacie Dudkiewicz
za codzienne wsparcie w gąszczu pułapek naukowych i administracyjnych,
za przygotowanie mnie do bycia wykładowczynią i badaczką,
za cenne, spokojne rady, kiedy świat wokół przyspieszał,
za cierpliwość, uważność i autentyczne zaangażowanie,
za nieustające zaufanie, które dodawało odwagi do działania.

Zespołowi, a w szczególności ***dr. Marcinowi Gradowskiemu****, który z niewzruszonym*
spokojem przyjmował na siebie grad moich pytań
o najdziwniejszych porach dnia i nocy.

Najbliższym: ***Mamie****,* ***Tacie*** *i* ***Bratu*** *– bez Was to nie byłoby możliwe.*

Dziękuję za towarzyszenie mi w tej drodze od początku do końca!

Oświadczenie promotora rozprawy doktorskiej przygotowanej przez doktoranta Szkoły Doktorskiej SGGW

Oświadczam, że **rozprawa doktorska autorstwa mgr Marianny Krysińskiej została przygotowana pod moim kierunkiem na podstawie badań realizowanych w ramach kształcenia w Szkole Doktorskiej SGGW**, i stwierdzam, że spełnia warunki do przedstawienia jej w postępowaniu o nadanie stopnia naukowego doktora.

23-09-2025 Krzysztof Pawłowski

Data Czytelny podpis promotora

Oświadczenie autora rozprawy doktorskiej będącego doktorantem Szkoły Doktorskiej SGGW

Świadoma odpowiedzialności prawnej, w tym odpowiedzialności karnej za złożenie fałszywego oświadczenia, oświadczam, że **niniejsza rozprawa doktorska została przygotowana przeze mnie samodzielnie na podstawie badań realizowanych w ramach kształcenia w Szkole Doktorskiej SGGW** i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami prawa, w szczególności z ustawą z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (tj. z dnia 28 października 2022 r., Dz.U. z 2022 r. poz. 2509 ze zm.)

Oświadczam, że przedstawiona rozprawa nie była wcześniej podstawą żadnej procedury związanej z uzyskaniem stopnia naukowego doktora.

Oświadczam ponadto, że niniejsza wersja rozprawy jest identyczna z załączoną wersją elektroniczną.

Przyjmuję do wiadomości, że rozprawa doktorska poddana zostanie procedurze antyplagiatowej.

Data 29.09.2025 Czytelny podpis autora rozprawy Marianna Krysińska

Streszczenie

W dobie rosnącego zagrożenia chorobami zakaźnymi i pojawiających się coraz liczniej anty biotykoopornych patogenów zrozumienie molekularnych mechanizmów infekcji staje się kluczowe dla opracowania skutecznych terapii. Bakterie z rodzaju *Legionella*, odpowiedzialne m.in. za legionellozę, należą do patogenów zdolnych do zaawansowanej manipulacji komórką gospodarza dzięki wytwarzaniu licznych efektorów – białek ingerujących w szlaki sygnalizacyjne i obronne komórki gospodarza. Wśród nich szczególną rolę odgrywają enzymy z nadrodzin ADP-rybozylotransferaz (ART) oraz kinaz białkowych – kluczowych regulatorów procesów komórkowych u prokariontów i eukariontów. Pomimo znacznych postępów w badaniach wiele takich efektorów pozostaje niezidentyfikowanych lub nieopisanych funkcjonalnie.

Celem niniejszej rozprawy doktorskiej była kompleksowa analiza rodzin enzymatycznych ADP-rybozylotransferaz (ART) oraz kinaz z wykorzystaniem metod bioinformatycznych, ze szczególnym uwzględnieniem ich ewolucji, struktury domen białkowych oraz potencjalnych funkcji biologicznych. Przeprowadzono systematyczne badania mające na celu identyfikację zarówno znanych, jak i nowych rodzin tych enzymów, a także przewidywanie funkcji nieopisanych wcześniej domen białek, wykazujących podobieństwo do ART lub kinaz. Analizy objęły zarówno panproteom bakterii *Legionella*, jak i proteom człowieka.

W pierwszej części pracy dokonano przeglądu rodzin ADP-rybozylotransferaz w rodzaju *Legionella*, obejmującego 41 gatunków. Zidentyfikowano 63 białka o przekonującym podobieństwie do znanych ART, zorganizowane w 39 rodzin, z czego aż 26 stanowiły nowe, dotąd nieopisane rodziny podobne do ART. Przykładem jest rodzina DUF2971, o przewidzianej aktywności wobec DNA, oraz największa z nowych rodzin – DUF4291 – o nietypowym miejscu aktywnym i przewidzianej aktywności wobec białek. Większość zidentyfikowanych białek to przewidywane efektory, mogące odgrywać istotną rolę w patogenezie. Wyniki tego badania udostępniono w publicznie dostępnej bazie danych astARTe.

Druga część pracy dotyczyła identyfikacji nowej, wysoce zachowanej w ewolucji eukariotycznej rodziny pseudo-ADP-rybozylotransferaz, której przedstawicielem jest m.in. ludzkie białko LRRC9. Rodzina ta charakteryzuje się obecnością domeny ART flankowanej przez domeny LRR, co sugeruje potencjalną rolę w odpowiedzi immunologicznej. Nietypowa budowa miejsca aktywnego wskazuje na możliwość odmiennej aktywności enzymatycznej lub funkcji sygnalizacyjnej. Wyniki sugerują, że białka te mogą integrować detekcję sygnałów z odpowiedzią komórkową, pełniąc funkcję sensorów wewnątrzkomórkowych.

Trzecia część rozprawy skupiła się na analizie pankinomu *Legionella*. Wykorzystując niestandardowe podejścia bioinformatyczne, odkryto 13 nowych rodzin kinaz i pseudokinaz, potencjalnie pełniących funkcję efektorów. Prace te znacząco rozszerzają znany zestaw

kinaz *Legionella*, wskazując również na ich obecność w innych taksonach bakteryjnych. Zidentyfikowane rodziny stanowią cenny materiał do dalszych badań nad mechanizmami manipulacji procesami komórkowymi gospodarza przez patogeny. Dane udostępniono publicznie w bazie Kintaro.

Zebrane wyniki wnoszą istotny wkład w poznanie różnorodności i ewolucji domen ART oraz kinaz, wskazując na istnienie wielu nieopisanych wcześniej rodzin tych enzymów oraz ich potencjalną rolę w patogenezie i regulacji procesów komórkowych. Opracowane w pracy podejścia i bazy danych stanowią wartościowe narzędzie do dalszych badań biologii molekularnej i bioinformatyki funkcjonalnej.

Słowa kluczowe: *Legionella*, ADP-rybozylotransferazy, kinazy białkowe, bioinformatyka

Summary

In an era of growing threats of infectious diseases and increasingly numerous antibiotic-resistant pathogens, understanding the molecular mechanisms of infection is crucial for the development of effective therapies. *Legionella* bacteria, responsible for Legionnaires' disease, among other things, are pathogens capable of advanced manipulation of the host cell through the production of numerous effectors – proteins that interfere with the host cell's signalling and defence pathways. Among them, enzymes from the ADP-ribosyltransferase (ART) superfamily and protein kinases, key regulators of cellular processes in prokaryotes and eukaryotes, play a special role. Despite significant advances in research, many such effectors remain unidentified or functionally undescribed.

The aim of this doctoral dissertation was to comprehensively analyse the ADP-ribosyltransferase (ART) and kinase enzyme families using bioinformatics methods, with particular emphasis on their evolution, protein domain structure, and potential biological functions. Systematic studies were conducted to identify both known and new families of these enzymes, as well as to predict the functions of previously undescribed protein domains showing similarity to ART or kinases. The analyses covered both the panproteome of *Legionella* bacteria and the human proteome.

In the first part of the study, a review of ADP-ribosyltransferase families in the genus *Legionella* was performed, covering 41 species. Sixty-three proteins with convincing similarity to known ARTs were identified, organised into 39 families, of which 26 were new, previously undescribed ART-like families. An example is the DUF2971 family with predicted activity against DNA, and the largest of the new families, DUF4291, with an atypical active site and predicted activity against proteins. Most of the identified proteins are predicted effectors that may play an essential role in pathogenesis. The results of this study are available in the publicly accessible astARTe database.

The second part of the work concerned the identification of a new family of pseudo-ADP-ribosyltransferases, highly conserved in eukaryotic evolution, represented by, among others, human LRRC9. This family is characterised by the presence of an ART domain flanked by LRR domains, suggesting a potential role in the immune response. The unusual structure of the active site indicates the possibility of different enzymatic activity or signalling functions. The results suggest that these proteins may integrate signal detection with cellular response, acting as intracellular sensors.

The third part of the dissertation focused on the analysis of the *Legionella* pankinome. Using non-standard bioinformatic approaches, 13 new families of kinases and pseudokinases were discovered, potentially serving as effectors. This work significantly expands the known set of *Legionella* kinases, also indicating their presence in other bacterial taxa. The identified families provide valuable material for further research into the mechanisms

by which pathogens manipulate host cellular processes. The data are publicly available in the Kintaro database.

The collected results contribute significantly to the understanding of the diversity and evolution of ART domains and kinases, indicating the existence of many previously undescribed families of these enzymes and their potential role in pathogenesis and regulation of cellular processes. The approaches and databases developed in this work are valuable tools for further research in molecular biology and functional bioinformatics.

Keywords: *Legionella*, ADP-ribosyltransferases, protein kinases, bioinformatics

Spis treści

Badania stanowiące podstawę niniejszej rozprawy doktorskiej były współfinansowane w ramach projektów Narodowego Centrum Nauki:

- projekt nr 2019/33/B/NZ2/01409
- projekt nr 2023/49/N/NZ2/03440

Wykaz publikacji stanowiących rozprawę doktorską

- **Krysińska, M.**, Gradowski, M., Baranowski, B., Pawłowski, K., Dudkiewicz, M. A survey of ADP-ribosyltransferase families in the pathogenic *Legionella*. BMC Genom. (2025). (**IF = 3,7; MNiSW = 140**)*
- Wyżewski, Z., Gradowski, M., **Krysińska, M.**, Dudkiewicz, M., Pawłowski, K. A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution. *Peer*J 9, e11051 (2021). (**IF = 3,061; MNiSW = 100**)*
- **Krysińska, M.**, Baranowski, B., Deszcz, B., Pawłowski, K., Gradowski, M. Pan-kinome of *Legionella* expanded by a bioinformatics survey. Sci. Rep. 12, 21782 (2022). (**IF = 4,6; MNiSW = 140**)*

Łączna punktacja Ministerstwa Edukacji i Nauki według wykazu czasopism punktowanych wynosi **380 punktów**.

Łączny *impact factor* (IF) według Journal Citation Reports wynosi **11,361**.

* Wartości *impact factor* poszczególnych publikacji podano w oparciu o dane udostępnione na In Cites TM Journal Citation Reports, a punktację czasopism podano w oparciu o wykazy czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych wraz z przypisaną liczbą punktów, stanowiące załączniki do komunikatu Ministra Nauki i Szkolnictwa Wyższego według roku publikacji.

1. Wstęp

1.1. Bioinformatyka

Bioinformatyka zajmuje się analizą i interpretacją danych biologicznych przy użyciu metod informatycznych i statystycznych. Obejmuje metody obliczeniowe służące do badania i symulacji struktury, funkcji i ewolucji genów, genomów oraz białek. Tworzy metody wykorzystywane do analizy i zarządzania informacjami biologicznymi gromadzonymi w toku badań genomicznych oraz badań prowadzonych z zastosowaniem wysoko przepustowych technik eksperymentalnych. Bioinformatyka pozwala na analizowanie ogromnych zbiorów danych genetycznych, proteomicznych i metabolomicznych, co prowadzi do nowych odkryć i zrozumienia procesów biologicznych na niespotykaną dotąd skalę[1,2].

Bioinformatyka jest silnie powiązana z wieloma dziedzinami nauki, m.in. biologią molekularną, informatyką, matematyką i statystyką, chemią, fizyką i medycyną, co czyni ją interdyscyplinarną z natury[1,2].

1.1.1. Historia i rozwój bioinformatyki

Historia bioinformatyki sięga lat 60. XX wieku, a termin „bioinformatyka" po raz pierwszy został użyty przez Paulien Hogeweg i Bena Hespera w 1970 roku w odniesieniu do badania procesów informatycznych w systemach biotycznych. W ten sposób naukowcy wyróżnili bioinformatykę jako osobną dziedzinę nauki obok biochemii i biofizyki[3].

Z bioinformatyką związana jest ściśle biologia molekularna, na której rozwój miały wpływ przede wszystkim dwa odkrycia dokonane w latach 50. XX wieku. Pierwszym z nich było opisanie w 1953 roku przez Jamesa Watsona, Maurice'a Wilkinsa i Rosalind Franklin struktury podwójnej helisy DNA, przechowującej informację genetyczną we wszystkich organizmach żywych[4]. Drugim odkryciem było ustalenie przez Fredericka Sangera w 1951 roku pełnej sekwencji aminokwasów insuliny[5]. Eksperymenty biologiczne zaczęły generować ogromną ilość danych, która okazała się trudna do przechowywania i „ręcznego" przetwarzania. Z tego powodu w pracy biologów niezbędne stało się wykorzystanie komputerów i informatyki.

Pionierką w dziedzinie bioinformatyki była Margaret Oakley Dayhoff – autorka pierwszej bioinformatycznej bazy danych *Atlas of Protein Sequence and Structure*, stworzonej w 1965 roku, w której zgromadziła wszystkie dostępne sekwencje białkowe[6]. Kolejny przełom nastąpił w latach 70., wraz z wynalezieniem przez Fredericka Sangera technologii sekwencjonowania DNA[7]. Zapotrzebowanie na analizy bioinformatyczne znacznie się zwiększyło, a rozwój dziedziny przyspieszył. W latach 80. XX wieku powstał

GenBank, amerykańska bioinformatyczna baza danych zbierająca sekwencje nukleotydowe i udostępniająca je naukowcom na całym świecie[8], a niedługo potem założono National Center for Biotechnology Information (NCBI) – instytucję, która zajęła się zarządzaniem GenBankiem. W kolejnej dekadzie rozpoczęto Human Genome Project (HGP), którego celem było zsekwencjonowanie całego genomu człowieka. Ten międzynarodowy projekt zakończono w 2003 roku, dostarczając pełny katalog ludzkich genów i ich sekwencji. Było to jedno z najważniejszych osiągnięć naukowych XX wieku, które nie byłoby możliwe bez zaawansowanych narzędzi bioinformatycznych.

W Europie powstała baza danych na Uniwersytecie w Genewie (SWISS-PROT) i European Molecular Biology Laboratory (EMBL), a w Japonii – baza DDBJ (DNA Data Bank of Japan). Celem SWISS-PROT jest dostarczanie wysokiej jakości danych o sekwencjach białek. Wyróżnia się dokładnymi i ręcznie opracowywanymi informacjami na temat funkcji białek, ich struktury, modyfikacji, lokalizacji komórkowej oraz interakcji. Baza ta stała się istotnym narzędziem w badaniach bioinformatycznych i biochemicznych, a obecnie wchodzi w skład większego zasobu znanego jako UniProt[9].

Z kolei EMBL to jeden z wiodących ośrodków badawczych w dziedzinie biologii molekularnej w Europie. Jest kuratorem EMBL Nucleotide Sequence Database[10] (obecnie część EMBL-EBI), bazy, która gromadzi, przechowuje i udostępnia sekwencje nukleotydowe DNA i RNA. Stanowi ona część międzynarodowego projektu współpracującego z GenBankiem i DDBJ, tworząc globalny system wymiany danych genetycznych.

Kolejnym przełomowym osiągnięciem w rozwoju bioinformatyki było powstanie AlphaFold[11] i RoseTTAFold[12], które zrewolucjonizowały przewidywanie struktur białkowych. Ich twórcy zostali docenieni przez Królewską Szwedzką Akademię Nauk Nagrodą Nobla w dziedzinie chemii w 2024 roku. AlphaFold, opracowany przez DeepMind, i RoseTTAFold, stworzony na Uniwersytecie Waszyngtońskim, pozwalają z dużą precyzją przewidywać trójwymiarowe struktury białek na podstawie sekwencji aminokwasów. Dzięki tym narzędziom możliwe stało się znaczne przyspieszenie badań nad funkcją białek i opracowywaniem nowych leków, co wcześniej było ograniczone przez czasochłonne i kosztowne metody eksperymentalne.

1.1.2. Główne dziedziny bioinformatyki

Bioinformatyka obejmuje wiele różnych obszarów badawczych, z których każdy ma unikalne zastosowania i metody. Zagadnienia, którymi zajmuje się ta dziedzina, można podzielić ze względu na rodzaj badanego obiektu bądź charakter prowadzonych analiz. Do kluczowych działów bioinformatyki wyróżnionych ze względu na badane obiekty należą:

- Genomika – zajmuje się badaniem kompletnego zestawu genów w organizmie. W skład jej zainteresowań wchodzą: sekwencjonowanie genomów, pozwalające na

ustalenie kompletnej sekwencji DNA danego organizmu; genomika porównawcza, pozwalająca na analizę genomów różnych gatunków w celu zrozumienia ewolucji i funkcji genów; oraz anotacja genomów, prowadząca do identyfikacji genów, elementów regulacyjnych i niekodujących fragmentów w sekwencji genomowej (promotorów, enhancerów, transpozonów itd.)[1,3,13].

- Proteomika – bada pełny zestaw białek w komórce lub organizmie. Pozwala na: wykrywanie i identyfikację białek, głównie z wykorzystaniem technik spektrometrii; analizę funkcji białek, uwzględniającą interakcje międzybiałkowe i ich rolę w procesach komórkowych; oraz profilowanie ekspresji białek, umożliwiające określanie poziomów białek w różnych warunkach biologicznych (np. stresu)[13,14].
- Transkryptomika – koncentruje się na badaniu RNA i ekspresji genów. Do jej zainteresowań należą: sekwencjonowanie RNA (RNA-Seq), czyli analiza poziomu ekspresji genów na podstawie ilości RNA; analiza różnicowej ekspresji, pozwalająca na identyfikację genów, których ekspresja zmienia się w odpowiedzi na różne warunki; oraz badanie różnych wariantów mRNA powstających z jednego genu (splicing alternatywny)[1,15].
- Metabolomika – analizuje metabolity i ich rolę w procesach biochemicznych. Uczestniczy w profilowaniu metabolitów – pozwala na identyfikację i analizę ilościową małych cząsteczek chemicznych w komórkach, a także na analizę szlaków metabolicznych poprzez badanie procesów biochemicznych i ich regulacji[1,2].
- Filogenetyka – zajmuje się rekonstrukcją historii ewolucyjnej obiektów (genów, białek i organizmów) na podstawie danych sekwencyjnych. Narzędzia filogenetyczne pozwalają na tworzenie drzew filogenetycznych, analizę pokrewieństw między obiektami oraz badanie zmian genetycznych na przestrzeni czasu[2,15,16].
- Biologia systemów – integruje dane z różnych dziedzin, takich jak genomika, proteomika i metabolomika, w celu modelowania i zrozumienia złożonych sieci biologicznych i procesów. Pozwala również na badanie zmian w systemach biologicznych w czasie oraz tworzy modele matematyczne opisujące interakcje między obiektami biologicznymi[1,2].

Ze względu na charakter prowadzonych analiz w bioinformatyce można wyróżnić trzy główne działy[1,2]:

- Bioinformatyka sekwencyjna – jej obiektem zainteresowań są sekwencje nukleotydowe i aminokwasowe: ich pozyskiwanie (sekwencjonowanie); identyfikacja podobieństw i różnic oraz zależności międzysekwencyjnych; odkrywanie charakterystycznych wzorców w sekwencjach; a także przypisywanie funkcji do określonych sekwencji genowych na podstawie znanych danych.

- Bioinformatyka strukturalna – koncentruje się na analizie trójwymiarowej struktury biomolekuł, przewidywaniu struktury białek, symulacji interakcji między molekułami i symulacji ruchów atomów w cząsteczkach biologicznych.
- Bioinformatyka funkcjonalna – zajmuje się modelowaniem ścieżek metabolicznych, profilowaniem ekspresji genu, przewidywaniem oddziaływań białka, przewidywaniem lokalizacji komórkowej białka.

1.1.3. Technologie i narzędzia bioinformatyczne

Rozwój bioinformatyki był możliwy dzięki postępowi w technologiach komputerowych i obliczeniowych. Do kluczowych zadań współczesnej bioinformatyki zaliczyć można następujące zagadnienia:

- Analiza danych biologicznych w różnych gałęziach nauki – dużą jej część stanowi opracowywanie danych z sekwencjonowania DNA. Technologie takie jak sekwencjonowanie metodą Sangera, sekwencjonowanie nowej generacji (NGS) oraz sekwencjonowanie jednocząsteczkowe (SMRT) pozwalają na szybkie i tanie sekwencjonowanie całych genomów, co prowadzi do powstania ogromnej ilości surowych danych. Narzędzia bioinformatyczne umożliwiają ich przekształcenie w użyteczne informacje biologiczne oraz zintegrowanie ich z dostępnymi informacjami w bazach danych. Bioinformatyka zajmuje się również danymi transkryptomicznymi (obróbka danych z RNA-Seq i mikroRNA), proteomicznymi (identyfikacja i klasyfikacja białek oraz badanie interaktomu) i epigenomicznymi (analiza wzorców metylacji DNA i modyfikacji histonów). Oprócz tego przedmiotem zainteresowania bioinformatyki są: genomika porównawcza (w tym filogenetyka); analiza wariantów genetycznych, m.in. analiza SNP (badanie jednonukleotydowych polimorfizmów, będących wariantami genetycznymi, które mogą wpływać na funkcję genów i wiązać się z predyspozycjami do chorób), analiza mutacji somatycznych (wykrywanie i analiza mutacji, które pojawiają się w komórkach somatycznych, np. w nowotworach), analiza wariantów strukturalnych (identyfikacja dużych zmian w genomie, takich jak delecje, insercje, duplikacje i translokacje, które mogą być związane z różnymi chorobami genetycznymi); analiza szlaków metabolicznych i sygnalizacyjnych oraz modelowanie sieci genetycznych i regulacyjnych. Bioinformatyka ma również swój udział w integracji danych genetycznych z danymi klinicznymi, takimi jak fenotyp pacjentów, historia medyczna i reakcje na leczenie, w celu opracowania spersonalizowanych terapii. Do kolejnych zadań bioinformatyki zaliczyć można coraz powszechniejsze analizy mikrobiomu, a także opracowywanie danych związanych z chorobami: analizę genetycznych podstaw odpowiedzi na leki oraz modelowanie molekularne

leków (w tym projektowanie struktur leków i biomolekuł, symulacje dokowania molekularnego, modelowanie dynamiki molekularnej itd.). Do ostatniej dużej grupy zadań stawianych przed bioinformatyką należą symulacje dynamiki molekularnej – analiza ruchów i dynamiki białek, lipidów i kwasów nukleinowych w czasie, co pozwala na badanie ich właściwości fizycznych oraz funkcji w środowisku komórkowym, a także modelowanie struktur białek, ligandów i kompleksów białko–ligand, co jest kluczowe dla zrozumienia mechanizmów ich działania[1,2,13,15,16].

- Tworzenie i zarządzanie bazami danych biologicznych – bazy danych, takie jak GenBank, EMBL, DDBJ, UniProt i PDB, gromadzą informacje dotyczące sekwencji DNA, RNA, białek oraz struktur molekularnych i udostępniają je naukowcom na całym świecie. Efektywne zarządzanie tymi ogromnymi zbiorami jest kluczowe dla rozwoju bioinformatyki, dlatego do zadań stawianych przed tą dziedziną należy przechowywanie sekwencji, struktur i innych danych biologicznych, ustalanie formatów danych cyfrowych i protokołów wymiany informacji oraz zapewnienie jakości i spójności danych w bazach[1,2].
- Tworzenie i udostępnianie oprogramowania bioinformatycznego – nowe narzędzia i wdrożenia metod są głównym elementem napędzającym rozwój całej dziedziny. Postęp w tym obszarze jest niezbędny do analiz wciąż powiększających się zasobów danych biologicznych[1,16].
- Obliczenia wysokowydajne (HPC) – superkomputery i klastry komputerowe umożliwiają analizę ogromnych zbiorów danych w krótkim czasie, co jest kluczowe dla badań genomowych i proteomicznych[1,2,16].
- Sztuczna inteligencja (AI) i uczenie maszynowe (ML) – algorytmy AI i ML są od kilku lat powszechnie wykorzystywane do analizy danych biologicznych, przewidywania struktur białek, odkrywania leków oraz analizy obrazów medycznych. Wdrożenie i wykorzystanie zaawansowanych metod obliczeniowych do analizy danych ma swoje zastosowanie w przewidywaniu struktury białek czy funkcji genów, analizie obrazów biologicznych (np. rozpoznawanie wzorców w obrazach mikroskopowych) i modelowaniu złożonych relacji między danymi biologicznymi[1].

1.1.4. Zastosowania bioinformatyki

Bioinformatyka znajduje szerokie zastosowanie w wielu dziedzinach nauki i przemysłu. Do kluczowych obszarów, w których wdrażane są wyniki badań bioinformatycznych, należą farmakogenomika, medycyna spersonalizowana i projektowanie nowych leków. Bioinformatyka umożliwia analizę genomów pacjentów oraz analizę funkcji, zmienności i struktur celów leków, co pozwala na dostosowanie leczenia do indywidualnych potrzeb.

Analiza danych genetycznych pomaga w identyfikacji predyspozycji do chorób, doborze leków oraz monitorowaniu odpowiedzi na terapię. Ponadto bioinformatyka wykorzystywana jest w badaniach organizmów w ich naturalnym środowisku – m.in. w analizie materiału genetycznego bezpośrednio z próbek środowiskowych, w badaniach różnorodności genetycznej ekosystemów[1,2,16].

1.2. Białka

Białka, inaczej proteiny, to wielkocząsteczkowe biopolimery o masie cząsteczkowej wahającej się od około 10 tysięcy do kilku milionów daltonów (Da). Z chemicznego punktu widzenia są to biologiczne polikondensaty, zbudowane z reszt aminokwasowych połączonych ze sobą wiązaniami peptydowymi. Występują powszechnie we wszystkich organizmach żywych oraz wirusach. Są syntetyzowane przez wyspecjalizowane struktury komórkowe – kompleksy białek i RNA – zwane rybosomami. Głównymi pierwiastkami wchodzącymi w skład białek są węgiel, tlen, wodór, siarka i azot, a także fosfor i selen oraz niekiedy kationy metali, m.in. Mn^{2+}, Zn^{2+}, Mg^{2+}, Fe^{2+}, Cu^{2+}, Co^{2+}.

Skład pierwiastkowy białek nie zawsze odpowiada składowi samych aminokwasów, ponieważ większość białek stanowią białka złożone (proteidy). Oprócz łańcucha polipeptydowego zawierają one również inne komponenty związane kowalencyjnie lub poprzez słabsze oddziaływania, takie jak cukry, różne związki pełniące funkcje koenzymów oraz jony metali[13,14,17].

1.2.1. Budowa białek

Łańcuch białkowy zsyntetyzowany w komórce przez rybosom przypomina swobodnie unoszącą się w roztworze nitkę, która ulega procesowi tzw. zwijania białka (ang. *protein folding*), w wyniku czego tworzy mniej lub bardziej sztywną bryłę, zwaną strukturą natywną lub konformacją białka. Zwykle tylko cząsteczki, które uległy zwinięciu do takiej formy, mogą pełnić właściwą dla danego białka funkcję biochemiczną. Istnieją jednak białka pozbawione struktury trzeciorzędowej, stanowiące wyjątek od tej reguły[13,14,17].

Pełną strukturę białka można opisać na czterech stopniach skomplikowania:

- struktura pierwszorzędowa białka (sekwencja aminokwasów) – kolejność aminokwasów w łańcuchu polipeptydowym; białko niezwinięte, w formie „nitki”;
- struktura drugorzędowa białka – lokalne przestrzenne ułożenie fragmentów łańcuchów polipeptydowych; do struktur drugorzędowych zaliczana jest:
  - α-helisa (ang. *α-helix*),
  - β-kartka (ang. *β-sheet*),
  - β-zakręt (lub β-zwrot) (ang. *β-hairpin*);

- struktura trzeciorzędowa białka – wzajemne położenie elementów struktury drugorzędowej;
- struktura czwartorzędowa białka – wzajemne położenie łańcuchów polipeptydowych oraz ewentualnie struktur niebiałkowych (grupa prostetyczna).

1.3. Białka enzymatyczne

Białka enzymatyczne są największą grupą wśród cząsteczek katalizujących reakcje chemiczne (obok rybozymów – tzn. enzymów kwasu rybonukleinowego[18] i deoksyrybozymów – enzymów DNA[19,20]) i dużą grupą wśród białek. Pełnią funkcję katalizatorów reakcji biochemicznych poprzez obniżenie energii aktywacji reakcji. Większość reakcji chemicznych zachodzących w organizmach żywych wymaga enzymów do osiągnięcia odpowiedniej wydajności reakcji. Z kolei enzymy są wysoce specyficzne względem substratów i dany enzym katalizuje jedynie niewielką część reakcji, w które może być zaangażowany konkretny substrat. Dzięki temu enzymy regulują procesy metaboliczne niezbędne dla funkcjonowania organizmów żywych.

Białka enzymatyczne to najczęściej białka globularne, które działają samodzielnie lub w większych kompleksach. Wielkość białek enzymatycznych waha się w szerokim zakresie – zawierają od kilkudziesięciu do kilku tysięcy aminokwasów. Tylko niewielka część ich struktury (zwykle od 2 do 4 aminokwasów) jest bezpośrednio zaangażowana w katalizę – tworzy miejsce katalityczne (ang. *catalytic site*). Znajduje się ono obok jednego lub więcej miejsc wiążących (ang. *binding site*), w których reszty aminokwasowe ustawiają substraty. Miejsce katalityczne i miejsce wiążące razem tworzą miejsce aktywne enzymu. Pozostała część (większość) struktury enzymu służy do utrzymania precyzyjnej orientacji i dynamiki miejsca aktywnego i/lub pełni funkcję utrzymywania całego białka w odpowiednim miejscu w komórce względem organelli (tzw. *anchoring*)[13,14].

1.3.1. Mechanizm działania enzymu

Wiązanie substratu z miejscem aktywnym enzymu jest bardzo specyficzną interakcją. Miejsca aktywne to szczeliny lub rowki na powierzchni enzymu, zwykle składające się z aminokwasów z różnych części łańcucha polipeptydowego, które są połączone w trzeciorzędową strukturę zwiniętego białka. Substraty początkowo wiążą się z miejscem aktywnym poprzez oddziaływania niekowalencyjne, w tym wiązania wodorowe, jonowe i hydrofobowe. Po związaniu substratu z miejscem aktywnym enzymu wiele mechanizmów może przyspieszyć jego konwersję do produktu reakcji. Najprostszym modelem interakcji enzym–substrat jest model *lock-and-key*, w którym substrat dokładnie pasuje do miejsca aktywnego. Jednak w wielu przypadkach konfiguracje zarówno enzymu, jak i substratu są modyfikowane przez wiązanie substratu – proces ten nazywany jest indukowanym dopasowaniem. W takich przypadkach

konformacja substratu jest zmieniana tak, aby bardziej przypominała stan przejściowy. Naprężenia powstałe w wyniku takiego zniekształcenia substratu mogą dodatkowo ułatwić jego konwersję do stanu przejściowego poprzez osłabienie krytycznych wiązań. Co więcej, stan przejściowy jest stabilizowany przez ścisłe wiązanie substratu z enzymem, w wyniku czego obniżona zostaje wymagana energia aktywacji. Oprócz łączenia wielu substratów i zniekształcania ich konformacji w celu zbliżenia się do stanu przejściowego wiele enzymów uczestniczy bezpośrednio w procesie katalitycznym. W takich przypadkach specyficzne łańcuchy boczne aminokwasów w miejscu aktywnym mogą reagować z substratem i tworzyć wiązania z półproduktami reakcji. Większość reszt katalitycznych zaangażowanych w te mechanizmy to aminokwasy polarne[14,21,22].

1.3.2. Kofaktory: grupy prostetyczne i koenzymy

Niektóre enzymy nie potrzebują dodatkowych składników, aby wykazać pełną aktywność. Inne, by mogły być aktywne, wymagają wiązania się niebiałkowych cząsteczek zwanych kofaktorami[14,21]. Kofaktory mogą być nieorganiczne (np. jony metali i klastry żelazowo-siarkowe) lub organiczne (np. flawina i hem). Organiczne kofaktory mogą być albo koenzymami, które są uwalniane z miejsca aktywnego enzymu podczas reakcji, albo grupami prostetycznymi, które są trwale związane z enzymem[14].

Koenzymy to małe cząsteczki organiczne, które mogą być luźno lub ściśle związane z enzymem. Niektóre z nich transportują grupy chemiczne z jednego enzymu do drugiego.

Ponieważ koenzymy ulegają zmianom chemicznym w wyniku działania enzymów, rozpatruje się je jako specjalną klasę substratów lub kofaktorów, które są wspólne dla wielu różnych enzymów – np. wiadomo, że około 1000 enzymów wykorzystuje koenzym NADH. Populacja koenzymów jest zazwyczaj regenerowana, a ich stężenia utrzymywane są na stałym poziomie wewnątrz komórki. Ta ciągła regeneracja pozwala na bardzo intensywne wykorzystywanie niewielkiej ilości koenzymów.

Przykładem koenzymu jest dinukleotyd nikotynoamidoadeninowy (NAD^+), który pełni funkcję nośnika elektronów w reakcjach utleniania-redukcji. NAD^+ może przyjąć jon wodorowy (H^+) i dwa elektrony (e^-) z jednego substratu, tworząc NADH. NADH może następnie przekazać te elektrony drugiemu substratowi, ponownie tworząc NAD^+. W ten sposób NAD^+ przenosi elektrony z pierwszego substratu (który ulega utlenieniu) do drugiego (który ulega redukcji)[14,21].

1.3.3. Inhibicja reakcji enzymatycznych

Szybkość reakcji enzymatycznych można zmniejszyć za pomocą inhibitorów enzymów. Wyróżnia się kilka typów mechanizmów hamowania reakcji enzymatycznej i właściwych

im inhibitorów. Inhibitor konkurencyjny (kompetycyjny) i substrat nie mogą wiązać się z enzymem w tym samym czasie[23]. Często inhibitory konkurencyjne są bardzo podobne do rzeczywistego substratu enzymu. Zahamowanie przez nie reakcji enzymatycznej można przezwyciężyć przy wysokim stężeniu substratu. W niektórych przypadkach inhibitor może wiązać się z miejscem innym niż miejsce wiązania substratu i wywierać efekt allosteryczny, zmieniając kształt miejsca wiązania substratu[24]. Ten typ hamowania nazywa się niekompetycyjnym, ponieważ inhibitor nie konkuruje bezpośrednio z substratem, ale działa pośrednio, zmieniając właściwości enzymu. Substrat nadal wiąże się ze swoim zwykłym powinowactwem, jednak inhibitor zmniejsza wydajność katalityczną enzymu. Inhibicji niekompetycyjnej, inaczej niż w przypadku inhibicji kompetycyjnej, nie można przezwyciężyć przy wysokim stężeniu substratu. Inhibitor akompetycyjny nie może wiązać się z wolnym enzymem, tylko z kompleksem enzym–substrat, dlatego najskuteczniej działa przy wysokim stężeniu substratu. W obecności takiego inhibitora kompleks enzym–substrat jest nieaktywny. Natomiast inhibitor nieodwracalny, w przeciwieństwie do wyżej wymienionych inhibitorów odwracalnych, trwale inaktywuje enzym, zwykle poprzez utworzenie wiązania kowalencyjnego z białkiem[14,25].

1.3.4. Klasyfikacja białek enzymatycznych

Wśród białek enzymatycznych wyróżnia się siedem głównych klas (EC, ang. *enzyme commission*) ze względu na rodzaj katalizowanej reakcji[26]. Są to:

- oksydoreduktazy – katalizują reakcje utleniania i redukcji;
- transferazy – przenoszą grupy funkcyjne;
- hydrolazy – katalizują hydrolizę różnych wiązań;
- liazy – rozcinają różne wiązania na drodze innej niż hydroliza czy utlenianie;
- izomerazy – katalizują zmiany izomeryzacyjne cząsteczek;
- ligazy – łączą cząsteczki wiązaniami kowalencyjnymi;
- translokazy – katalizują ruch jonów i cząsteczek przez błony lub ich rozdział wewnątrz błon.

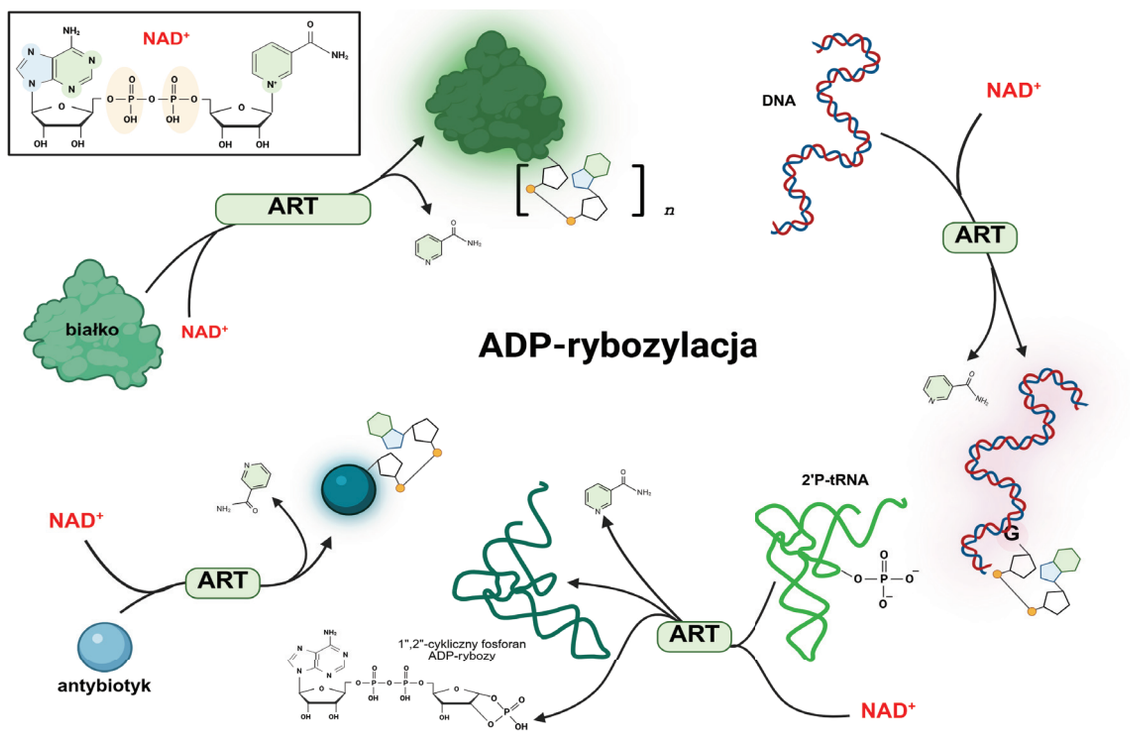
Każda z klas dzieli się dalej na podklasy i podpodklasy w zależności od wykorzystywanego substratu, uzyskiwanego produktu oraz wykorzystywanego mechanizmu chemicznego. Klasyfikacja EC nie odzwierciedla podobieństw ewolucyjnych w poszczególnych grupach enzymów.

1.4. ADP-rybozylotransferazy i ADP-rybozylacja

ADP-rybozylotransferazy (ang. *ADP-ribosyltransferases*, ART) to enzymy z klasy transferaz, modyfikujące chemicznie białka, kwasy nukleinowe i inne cząsteczki poprzez

przyłączenie reszty ADP-rybozy pochodzącej z utlenionego dinukleotydu nikotynoamidoadeninowego (NAD^+) za pośrednictwem wiązań N-, O- lub S-glikozydowych (w zależności od ADP-rybozylowanego łańcucha bocznego aminokwasu)[27–29]. ADP-rybozylacja jest reakcją odwracalną – odwrotny proces przeprowadzają ADP-rybozylohydrolazy oraz glikohydrolaza poli-ADP-rybozy (PARG). Przyłączenie ADP-rybozy przez ADP-rybozylotransferazy występuje w dwóch formach: jako poli-ADP-rybozylacja, w której tworzone i przyłączane do docelowych białek są liniowe lub rozgałęzione polimery, oraz jako mono-ADP-rybozylacja, która wymaga przyłączenia pojedynczej cząsteczki (rys. 1)[28,30–32].

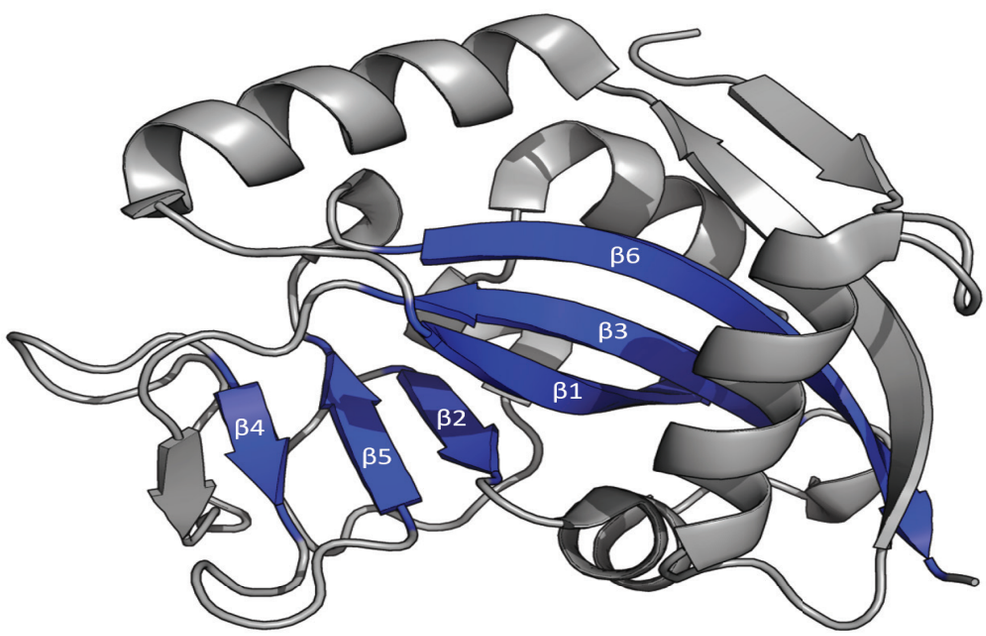
Prawdopodobnie mono-ADP-rybozylacja pierwotnie pojawiła się w bakteriach jako mechanizm obronny przeciwko wirusom, innym gatunkom bakteryjnym i cząsteczkom przeciwdrobnoustrojowym[28,33]. W odróżnieniu od rodzin enzymów katalizujących większość innych modyfikacji białek ART modyfikują różnorodne łańcuchy boczne w białkach. Dotychczas wykazano, że modyfikują: grupę –NH2 z grupy guanidynowej w argininie oraz amidowej bocznego łańcucha glutaminy i asparaginy, grupę fosforanową fosfoseryny, wewnątrzcykliczną grupę –NH– w dyftamidzie (ang. *diphthamide*), grupę –SH w łańcuchu bocznym cysteiny i grupę –COOH glutaminianu i asparaginianu[27,28,30,34–37]. Ponadto ostatnie doniesienia mówią o możliwej ADP-rybozylacji grupy –OH seryny[38], ubikwitynowanej treoniny[39], grupy imidazolowej histydyny, a także grupy hydroksylowej tyrozyny[40]. W kwasach nukleinowych przyłączają ADP-rybozę do zasad azotowych (guaniny i tyminy), fosforylowanych końców DNA i RNA[41].



Rys. 1. Schemat reakcji ADP-rybozylacji

Wiadomo również, że aktywność ADP-rybozylotransferaz (PARP-1 i PARP-2) jest niezbędna do naprawy uszkodzeń DNA: deaminacji, hydroksylacji i metylacji zasad azotowych[42–44]. Białka ART uczestniczą w usuwaniu jedno- i dwuniciowych pęknięć DNA (naprawa poprzez wycięcie zasad, wycięcie nukleotydów, rekombinację homologiczną i łączenie niehomologicznych zakończeń)[44–46]. ADP-rybozylotransferazy odgrywają też istotną rolę w regulacji metabolizmu RNA oraz w obróbce i metabolizmie mRNA[47–51].

Cechą charakterystyczną ADP-rybozylotransferaz jest wysoce specyficzny dla wiązania NAD^+ typ struktury ART, który znajduje się w domenach katalitycznych wszystkich ART, od wirusowych po ludzkie (rys. 2). Niemal wszystkie ADP-rybozylotransferazy zawierają dodatkowe domeny białkowe, oprócz domeny katalitycznej ART, które ukierunkowują enzymy na ich substraty i określone lokalizacje komórkowe[29,52,53]. Kluczowym kryterium stosowanym przy klasyfikacji domen nadrodziny ADP-rybozylotransferaz jest obecność niektórych katalitycznych aminokwasów w strukturze ART[28,29,54].



Rys. 2. Struktura domeny ART na podstawie struktury 2*x*5*y* (PDB)

Kolorem niebieskim oznaczono rdzeń domeny ART, miejsce aktywne tworzone przez trzy motywy znajdujące się odpowiednio na niciach β1, β2 i β5.

ADP-rybozylotransferazy z różnych rodzin charakteryzują się słabym zachowaniem sekwencji aminokwasowej, co niejednokrotnie jest przyczyną trudności w identyfikacji nowych członków nadrodziny jedynie na podstawie analizy sekwencji. Mimo to wszyscy członkowie nadrodziny wykazują dobrze zachowany rdzeń struktury trzeciorzędowej, składający się z podzielonego arkusza β, którego każda połowa zbudowana jest z trzech nici β (rys. 2). Poszczególne odcinki β łączą się ze sobą α-helisami. W efekcie cząsteczka NAD^+ jest blokowana między powierzchniami dwóch jednostek β-kartki[28,29,55,56].

1.4.1. Klasyfikacja ADP-rybozylotransferaz

Aravind z zespołem (2015)[28] podzielili ART na trzy główne klady: klad H-H-h, klad H-Y-[EDQ] i klad R-S-E. Charakteryzują się one różnymi konfiguracjami reszt aminokwasowych w miejscu aktywnym. W domenach kladu H-H-h centrum katalityczne zawiera dwie histydyny i jedną resztę hydrofobową, a struktura trzeciorzędowa ujawnia sześć *β*-kartek i jest pozbawiona wydłużonych wstawek między nimi. Domeny należące do kladu H-Y-[EDQ] mają miejsce aktywne złożone z histydyny, tyrozyny i glutaminianu, asparaginianu lub glutaminy i charakteryzują się, w zależności od rodziny, obecnością wstawek C-końcowych, N-końcowych lub wewnętrznych, które mogą odgrywać rolę w rozpoznawaniu substratu. Natomiast w domenach kladu R-S-E aktywna triada składa się z argininy, seryny (lub treoniny – polarna reszta aminokwasowa) i glutaminianu, a struktura ujawnia obecność siódmej, C-końcowej, *β*-nici, która tworzy strukturę spinki do włosów z szóstą nicią rdzeniową[28].

1.4.2. Świat ADP-rybozylotransferaz

ADP-rybozylotransferazy są dobrze zachowane i szeroko rozpowszechnione wśród wszystkich trzech domen życia: Archaea, Bacteria i Eukaryota oraz u wirusów[28,29]. Najnowsze badania ujawniły ogromną różnorodność funkcji tych enzymów, m.in. uczestnictwo w regulacji podstawowych procesów fizjologicznych na poziomie komórkowym i całego organizmu[53].

Najlepiej poznaną rolą ADP-rybozylotransferaz jest pełnienie przez nie funkcji toksyn bakteryjnych, takich jak toksyna cholery (*Vibrio cholerae*), toksyna błonicy (*Corynebacterium diphtheriae*) i toksyna krztuśca (*Bordetella pertussis*). ADP-rybozylacja białek gospodarza może wywoływać zdarzenia sprzyjające rozwojowi infekcji. Na poziomie komórkowym może zmieniać potencjał apoptotyczny komórki lub zaburzać organizację cytoszkieletu aktynowego i organizację błon komórkowych[57,58]; natomiast na poziomie organizmu – poprzez zaburzenia ciągłości oraz przepuszczalności nabłonków i śródbłonków – może zaburzać komórkową odpowiedź immunologiczną[28,29,37,54].

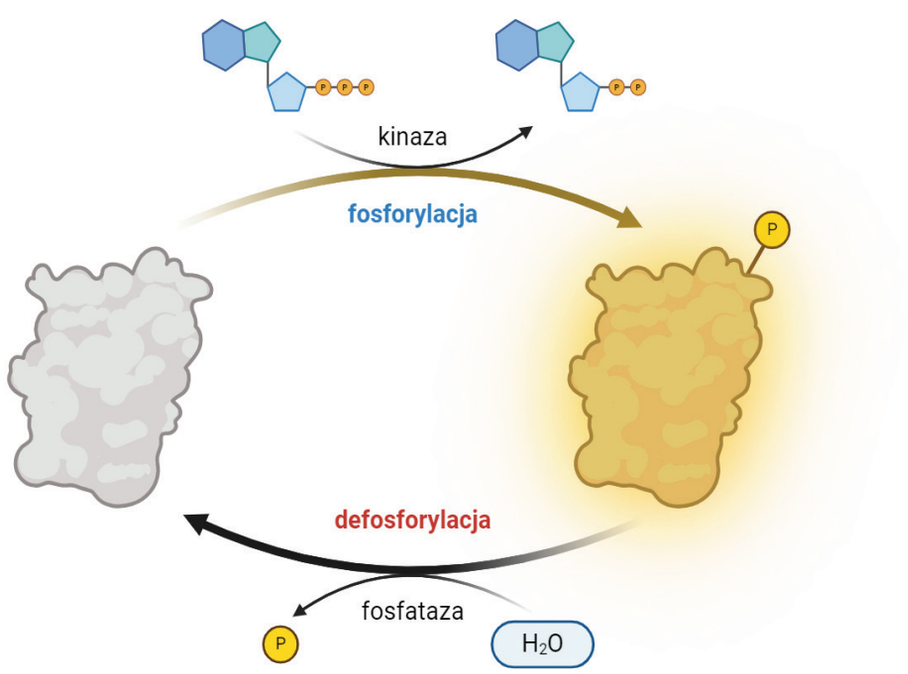
U eukariontów ADP-rybozylacja jest zaangażowana m.in. w regulację transkrypcji, translację, stabilność RNA, sygnalizację komórkową, montaż wrzeciona i podział komórek oraz transport jądrowo-cytoplazmatyczny. ART odgrywają też istotną rolę w odpowiedzi komórkowej na sygnały pozakomórkowe, np. czynniki wzrostu. W warunkach stresu są wiązane zarówno z odpornością wrodzoną, jak i adaptacyjną (szczególnie przeciwko wirusom), kontrolą jakości białek i odpowiedzią na uszkodzenia DNA. Pełnią również ważną funkcję w sygnalizacji i aktywacji komórek odpornościowych oraz regulują odpowiedź komórkową na cytokiny, a także ich ekspresję i wydzielanie[29,32,59,60].

1.4.3. ART u człowieka

U człowieka obecne są zarówno poli-, jak i mono-ADP-rybozylotransferazy. Poli-ADP-rybozylotransferazy są zaangażowane głównie w procesy naprawy DNA, a także regulację długości telomerów oraz ścieżki sygnalizacyjnej Wnt, kluczowej dla rozwoju i różnicowania komórek. Mono-ADP-rybozylotransferazy uczestniczą m.in. w regulacji odpowiedzi immunologicznej i biologii nowotworów.

1.5. Kinazy białkowe i fosforylacja

Kinazy (fosfotransferazy) należą do klasy transferaz i przeprowadzają transfer – reakcję fosforylacji – trzeciej grupy fosforanowej PO_4 (grupy gamma – γ) cząsteczki nukleotydu purynowego, np. adenozynotrójfosforanu (ATP), na cząsteczkę docelową (akceptor). Akceptorem może być białko, lipid lub mała cząsteczka, np. antybiotyk (rys. 3). Fosforylacja jest odwracalną modyfikacją; odwrotny proces przeprowadzają fosfatazy[61].



Rys. 3. Reakcja fosforylacji

1.5.1. Klasyfikacja kinaz

Klasyfikację kinaz białkowych opracowali: Kannan i wsp. (2007), Scheeff i wsp. (2005) oraz Manning i wsp. (2002), bazując na ich funkcji, podobieństwie strukturalnym i sekwencyjnym oraz filogenezie. Wyodrębnione klasy kinaz na różnym poziomie pokrewieństwa – grupy, rodziny – zbudowane są odpowiednio na podstawie miejsca fosforylacji substratu i rodzaju substratu oraz podobieństwa sekwencyjnego, strukturalnego i funkcji. Podrodziny są wyodrębnione w rodzinach kinaz i charakteryzują się wyraźnie zaznaczonym podobieństwem strukturalnym i sekwencyjnym[62–65].

Nadrodzina kinaz według tej klasyfikacji dzieli się na 11 grup. Dziewięć grup zawiera typowe kinazy eukariotyczne. Grupy te zostały wydzielone na podstawie podobieństwa sekwencji domen katalitycznych, obecności domen pomocniczych oraz według znanych modeli regulacji aktywności kinaz. Zidentyfikowane przez Manninga typowe kinazy eukariotyczne, które nie wykazują wystarczającego podobieństwa sekwencyjnego do żadnej z ośmiu grup, umieszczono w grupie „inne”[63]. Kinazy niepasujące do żadnej z dziewięciu grup podzielono na atypowe (ang. *atypical*) i PKL (ang. *protein kinase-like*).

1.6. Klasyfikacja ART i kinaz w oparciu o sekwencje i struktury według bazy rodzin Pfam

Możliwe do zidentyfikowania domeny funkcjonalne i strukturalne są sklasyfikowane w bazie domen rodzin białkowych Pfam. Poszczególne domeny są opatrzone adnotacjami, najczęściej pochodzącymi z literatury, a kolekcja wzbogacona jest użytkowymi zestawami danych dla każdego wpisu, m.in. zestawami sekwencji, ukrytymi modelami Markova (HMM), a także odsyłaczami do innych baz danych, np. bazy struktur białkowych PDB, oraz modelami wygenerowanymi przy użyciu AlphaFold. Dostępne są również dane dotyczące zakresu taksonomicznego, w którym domena została zidentyfikowana, schematy przedstawiające architekturę domen oraz logo sekwencyjne. Wyodrębnione rodziny domen są pogrupowane w klany – zestawy powiązanych wpisów, których relacja może być zdefiniowana poprzez podobieństwo sekwencji, struktury lub profili HMM. Aktualna wersja bazy Pfam – Pfam 37.0 (wrzesień 2024) – zawiera 21 979 rodzin zgrupowanych w 709 klanów.

1.7. *Legionella*

Legionella to rodzaj wolno żyjących, związanych z biofilmem lub z gospodarzem Gram-ujemnych pałeczek z klasy gammaproteobakterii, mających polarnie rozmieszczone wici (od 1 do 3)[66]. Do wzrostu wymaga wysokiego stężenia dwutlenku węgla i dużej wilgotności.

Naturalnym rezerwuarem tych bakterii są wody śródlądowe i morskie, głównie w strefie przybrzeżnej, gdzie łatwo dochodzi do eutrofizacji. Występują one również w glebie i w okolicach gorących źródeł. Gatunkiem, z którego powodu wyodrębniono rodzinę *Legionelliaceae*, jest *Legionella pneumophila*. Została opisana w latach 80. po zidentyfikowaniu jej jako czynnika etiologicznego nietypowego zapalenia płuc, którego epidemia wybuchła wśród uczestników konwentu Legionu Amerykańskiego. Zakażenie *Legionella* u ludzi może powodować ciężkie atypowe zapalenie płuc zwane legionellozą lub łagodniejszą formę zwaną gorączką Pontiac[67–72]. Obecnie rodzaj *Legionella* obejmuje ponad 70 różnych gatunków bakterii, z których blisko połowa została uznana za patogenne dla ludzi.

Po wniknięciu do komórki gospodarza *Legionella* tworzy specjalną wakuolę, w której może przetrwać i się rozmnażać, zwaną wakuolą zawierającą legionellę (LCV)[73–75]. Białka efektorowe przenoszone przez bakterie do komórki gospodarza mogą głęboko zmienić jej zachowanie, co znacznie ułatwia replikację i patogenezę legionelli. Zmiany te obejmują ochronę LCV przed degradacją, manipulowanie układem endomembranowym, tłumienie odpowiedzi immunologicznej gospodarza i wykorzystywanie zasobów gospodarza[76–82].

Efektory bakteryjne to białka eksportowane przez patogenne bakterie do komórek gospodarza. Modyfikują ich szlaki sygnałowe i umożliwiają bakteriom przetrwanie i replikację. Duża część białek efektorowych wciąż nie jest scharakteryzowana, ale wiadomo, że funkcję efektorów, obok proteaz, ligaz ubikwitynowych i glikozylotransferaz, pełnią zarówno ADP-rybozylotransferazy, jak i kinazy[83,84]. ADP-rybozylotransferazy katalizują przenoszenie grup ADP-rybozylowych na białka gospodarza, co często prowadzi do inaktywacji kluczowych białek i destabilizacji cytoszkieletu lub modyfikacji procesów transkrypcyjnych. Z kolei kinazy efektorowe działają przez fosforylację białek w komórce gospodarza, co może zmieniać ich aktywność i prowadzić do zaburzeń w sygnalizacji komórkowej, immunologicznej czy metabolizmie. Obie klasy efektorów (ART i kinazy białkowe) są kluczowe dla skutecznej kolonizacji i unikania odpowiedzi immunologicznej przez bakterie patogenne. Gatunki bakterii *Legionella* wykorzystują do 330 białek efektorowych[84].

2. Cel i zakres pracy

Celem pracy była bioinformatyczna analiza rodzin ADP-rybozylotransferaz i kinaz, w tym badania nad ewolucją domen białkowych, zarówno znanych, jak i nowo zidentyfikowanych; poszukiwanie nowych rodzin białek zawierających domeny wykazujące podobieństwo do ADP-rybozylotransferaz (białka ART-like) i kinaz w publicznie dostępnych bazach danych sekwencji oraz przewidywanie funkcji nieopisanych domen białkowych wykazujących podobieństwo do ADP-rybozylotransferaz oraz kinaz.

Badania obejmowały kompleksową analizę panproteomu bakterii *Legionella* pod kątem obecności nieopisanych wcześniej ADP-rybozylotransferaz i kinaz białkowych, enzymów kluczowych dla infekcji, a także analizę proteomu człowieka.

3. Hipoteza badawcza

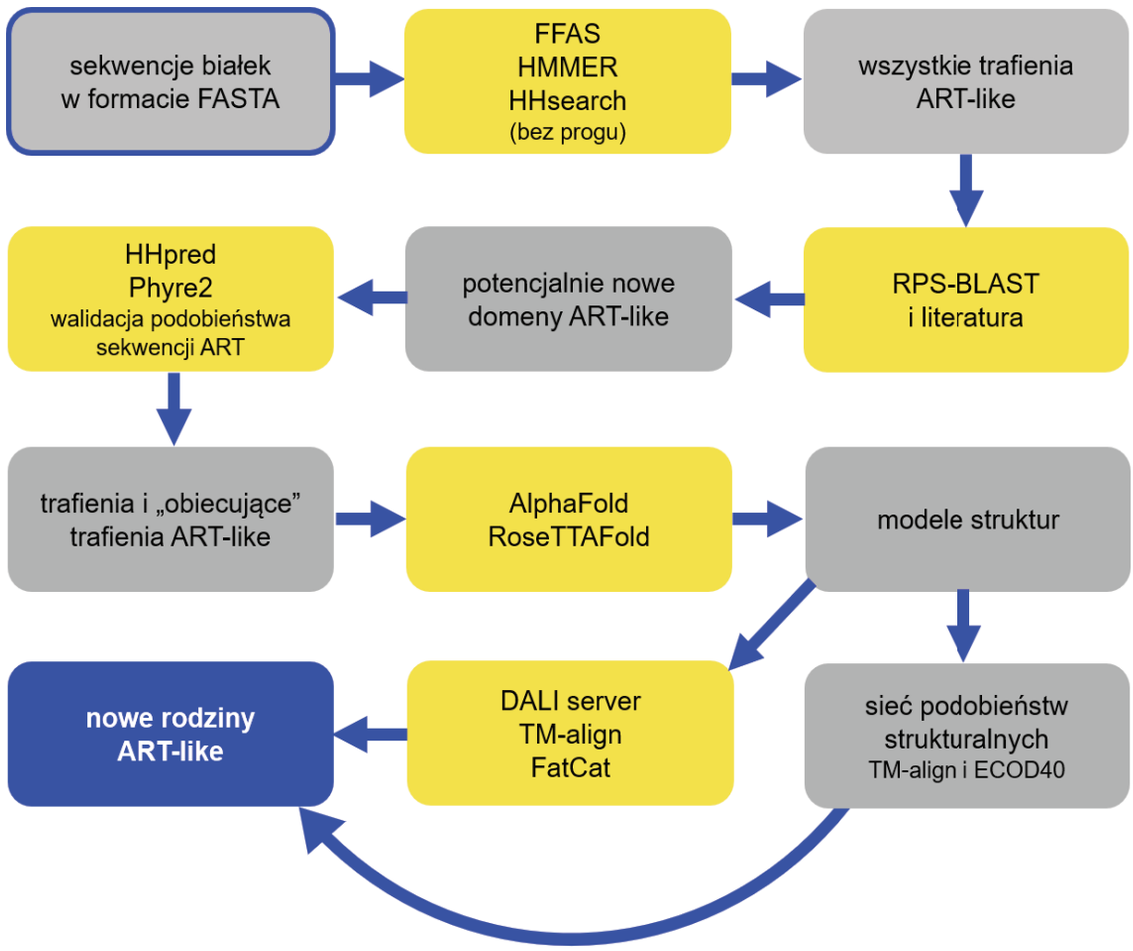
Nowe, nieopisane dotąd domeny białkowe wykazujące podobieństwo do ADP-rybozylotransferaz (ang. *ART-like*) oraz kinaz białkowych są obecne w panproteomie bakterii *Legionella* oraz człowieka i mogą pełnić nieznane dotąd funkcje biologiczne.

4. Metodyka

Do badań zastosowano dwie różne ścieżki: tzw. klasyczną, bazującą na podobieństwie sekwencyjnym i jedynie wspomaganą modelowaniem wybranych struktur białkowych, oraz tzw. alternatywną, opierającą się na modelowaniu struktur białek bądź wykorzystaniu modeli białek udostępnionych w bazach danych.

4.1. Klasyczna ścieżka poszukiwania odległych homologów, niewykrywalnych standardowymi metodami typu BLAST, bazująca na podobieństwie sekwencyjnym

Na potrzeby badań została opracowana wieloetapowa metoda wykrywania i weryfikowania odległych podobieństw sekwencyjnych (również tych z pogranicza wykrywalności). Było to konieczne, ponieważ standardowo wykorzystywany w analizach sekwencyjnych algorytm BLAST pozwala na wykrycie podobieństwa sekwencji na poziomie około 30 %. Dzięki zastosowaniu szeregu czulszych narzędzi i kilkustopniowej walidacji wyników udało się wykryć i zweryfikować podobieństwa do znanych ADP-rybozylotransferaz i kinaz na poziomie około 8–9 % identyczności sekwencji. Analizy obejmowały trzy główne etapy (rys. 4).



Rys. 4. Poglądowy schemat przyjętej metodyki w wersji „klasycznej”

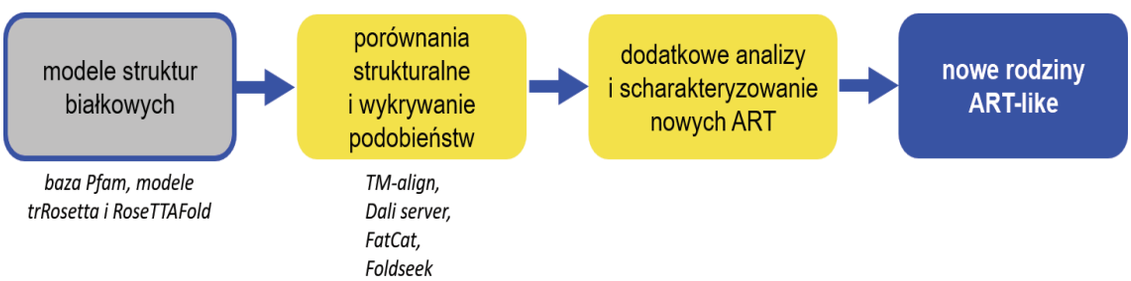
Kolorem szarym oznaczono dane/wyniki, kolorem żółtym – wykorzystywane metody i narzędzia.

Etap pierwszy polegał na przygotowaniu danych (sekwencji aminokwasowych) do analiz. Następnie przeprowadzono wstępną analizę podobieństw (z użyciem czułych algorytmów opierających się na wykorzystaniu profilu częstotliwości aminokwasów lub profilu Ukrytego Modelu Markowa – HMM) i przed kolejnymi etapami odrzucono łatwo wykrywalne wyniki. W końcu zwalidowano otrzymane wyniki metodami sekwencyjnymi i strukturalnymi, a także uzyskano wyniki dodatkowych analiz (analizy klastrów, badania otoczeń genomicznych, przewidywania obecności peptydów sygnałowych i lokalizacji subkomórkowej i in.).

Potwierdzone podobieństwa do znanych ART i kinaz opracowano i zdeponowano w formie publicznie dostępnych baz danych: astARTe (ADP-rybozylotransferazy) i Kintaro (kinazy).

4.2. Alternatywna ścieżka poszukiwań odległych homologów, bazująca na podobieństwie strukturalnym, umożliwiająca identyfikację rodzin niewykazujących podobieństwa sekwencji do znanych ART i kinaz

Podczas badań udostępnione zostały narzędzia do modelowania struktur białkowych opierające się m.in. na głębokich sieciach neuronowych (AlphaFold, RoseTTAFold i ich następcy). Możliwe stało się wielkoskalowe modelowanie struktur białek z wysoką dokładnością. Dzięki udostępnieniu modeli białek (najpierw z trRosetta, a potem z RoseTTAFold) dla rodzin zdefiniowanych w bazie Pfam została opracowana alternatywna metoda wykrywania podobieństw do ART (rys. 5). Obejmowała ona porównania strukturalne wykonane za pomocą algorytmu TM-align pomiędzy modelami białek udostępnianymi w bazie Pfam a zestawem struktur referencyjnych z bazy ECOD dla ART (ECOD40). Po wykryciu podobieństwa strukturalnego do ART przeprowadzano wszystkie analizy analogicznie do rodzin wykrytych podstawową metodą (metodami sekwencyjnymi).



Rys. 5. Alternatywna metoda wykrywania odległej homologii w oparciu o podobieństwa strukturalne

5. Syntetyczne omówienie publikacji

5.1. Publikacja 1. *A survey of ADP-ribosyltransferase families in the pathogenic* Legionella

ADP-rybozylotransferazy (ART) to nadrodzina enzymów zaangażowanych w różne procesy komórkowe, w tym mechanizmy patogenne. Rodzaj *Legionella*, znany z wywoływania choroby legionistów, posiada różnorodne efektory podobne do ART. W niniejszej pracy opisano kompleksowe badanie bioinformatyczne proteomów 41 gatunków *Legionella* w celu zidentyfikowania i scharakteryzowania białek o znacznym podobieństwie sekwencyjnym lub strukturalnym do znanych ART (rodzin ART-podobnych), z uwzględnieniem ich potencjalnej roli w patogenezie i interakcjach z gospodarzem.

Przeprowadzono czułe wyszukiwanie sekwencji w celu wykrycia kandydujących rodzin podobnych do ART. Późniejsza walidacja, w tym przewidywanie struktury takich rodzin, została osiągnięta przy użyciu narzędzi opartych na sztucznej inteligencji, takich jak AlphaFold. Przeprowadzono analizy porównawcze w celu oceny podobieństw sekwencyjnych i strukturalnych między nowymi rodzinami ART-podobnymi a znanymi ART.

Zidentyfikowano 63 białka o przekonującym podobieństwie do ART, zorganizowane w 39 rodzin ART-podobnych, w tym 26 nowych rodzin. Kluczowe odkrycia obejmują następujące rodziny:

- rodzina DUF2971 – wykazuje podobieństwo sekwencji do toksyn DarT i innych ART działających na DNA;
- rodzina DUF4291 – największa nowo zidentyfikowana rodzina; wykazuje strukturalne i sekwencyjne podobieństwo do toksyny błoniczej, co sugeruje zdolność do modyfikacji białek.

Większość członków nowych rodzin ART to przewidywane efektory. Chociaż konieczna jest eksperymentalna walidacja przewidywanych funkcji efektorowych ART, zidentyfikowane nowe rodziny ART-podobne stanowią obiecujące cele dla zrozumienia patogenności bakterii *Legionella* i opracowania strategii terapeutycznych.

Pełny katalog wyników badania został opublikowany w bazie danych astARTe: http://bioinfo.sggw.edu.pl/astarte/.

5.2. Publikacja 2. *A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution*

Obecność wielu całkowicie niescharakteryzowanych białek, nawet w dobrze zbadanych organizmach, takich jak człowiek bądź organizmy modelowe, poważnie utrudnia pełne zrozumienie funkcjonowania żywych komórek. ADP-rybozylacja jest powszechną potranslacyjną modyfikacją białek, kwasów nukleinowych i małych cząsteczek (np. antybiotyków). Pełni istotną funkcję w sygnalizacji komórkowej i procesach infekcji. Zwykle jest przeprowadzana przez enzymy z dużej nadrodziny ADP-rybozylotransferaz (ART).

Wykorzystując metody bioinformatyczne, w niniejszym badaniu zidentyfikowano nową rodzinę domniemanych ADP-rybozylotransferaz, dobrze zachowaną w ewolucji eukariotycznej, z nietypowym miejscem aktywnym. Cechą charakterystyczną tych białek jest domena ART umieszczona pomiędzy flankującymi domenami powtórzeń bogatych w leucynę (LRR). Powtórzenia LRR są zazwyczaj zaangażowane w nieswoistą odpowiedź immunologiczną. Dywergencja sekwencji i brak wyraźnie wykrywalnego „klasycznego" miejsca aktywnego ART sugeruje, że nowe domeny są pseudo-ADP-rybozylotransferazami, jednak nie można wykluczyć nietypowej aktywności ART lub alternatywnej aktywności enzymatycznej. Postawiono hipotezę, że nowa rodzina, w tym jej ludzki członek – LRRC9, może być zaangażowana w pierwotny mechanizm obronny, analogiczny do wrodzonego układu odpornościowego, i dzięki posiadaniu różnych domen funkcjonalnych może łączyć w obrębie jednego białka wykrywanie bodźców zewnętrznych (domena LRR) oraz mechanizmy sygnalizacyjne (domena ADP-rybozylotransferazy).

5.3. Publikacja 3. *Pan-kinome of* Legionella *expanded by a bioinformatics survey*

Patogenne bakterie *Legionella* są znane z dostarczania licznych białek efektorowych do komórki gospodarza w celu zakłócania i przejmowania procesów komórkowych na swoją korzyść. Pomimo intensywnych badań wiele efektorów pozostaje niescharakteryzowanych. Zmotywowani bogactwem repertuaru efektorów *Legionella* i ich często nietypową biochemią, a także kilkoma znanymi nietypowymi kinazami i odkrytymi niedawno pseudokinazami efektorowymi z *Legionella*, autorzy niniejszej pracy podjęli się badania *in silico* i eksploracji pankinomu rodzaju *Legionella* (tj. połączenia kinomów – pełnych zestawów genów kinaz białkowych – poszczególnych gatunków). W opisanym badaniu odkryto 13 nowych rodzin (pseudo)kinaz (wszystkie są potencjalnymi efektorami) przy użyciu niestandardowych podejść bioinformatycznych.

Katalog rodzin kinaz białkowych efektorowych i nieefektorowych w obrębie *Legionella*, wraz z 16 znanymi rodzinami, udostępniony został pod adresem: http://bioinfo.sggw.edu.pl/kintaro/. Zawiera on analizę i omówienie prawdopodobnych ról funkcjonalnych nowych przewidywanych kinaz.

Warto zauważyć, że niektóre z rodzin kinaz są również obecne w innych taksonach bakterii, w tym w innych patogenach, często filogenetycznie bardzo odległych od *Legionella*. Praca ta pokazuje pomysłowość natury w wyścigu zbrojeń patogen–gospodarz i oferuje przydatne źródło do badania mechanizmów infekcji.

6. Podsumowanie i wnioski

Dzięki kompleksowym analizom zidentyfikowano nowe rodziny ART, w tym 26 nowych rodzin u *Legionella*. Analogicznymi metodami zidentyfikowano również 13 nowych rodzin kinaz białkowych. Komplet uzyskanych wyników, wzbogacony o dane dotyczące znanych rodzin ART i kinaz, jest dostępny w dwóch publicznych bazach danych: astARTe (baza danych ART) i KINtaro (baza danych kinaz).

Uzyskane wyniki dowodzą, że zastosowana metodyka, oparta na podobieństwach strukturalnych, jest w przypadku białek enzymatycznych o dobrze zachowanym rdzeniu struktury wydajniejszą metodą od metod „klasycznych", bazujących na podobieństwach sekwencyjnych.

Prawdopodobnie nadrodziny ART oraz PKL, a także wiele innych nadrodzin enzymów będzie można rozbudowywać metodami opisanymi w niniejszej pracy. Bazy danych, takie jak astARTe i KINtaro, mogą zostać rozszerzone o dane dotyczące innych organizmów, co pozwoli na identyfikację dodatkowych rodzin enzymów oraz ich biologicznych ról w różnych kontekstach ekologicznych i chorobotwórczych. Tego typu badania mogą również pomóc w zrozumieniu oporności na antybiotyki, co jest kluczowe w erze wzrastającej liczby opornych patogenów. Dotychczas przeprowadzone badania skoncentrowane na bakterii *Legionella* oraz proteomie człowieka to wąski wycinek świata ART i PKL, a uzyskane wyniki pozwalają myśleć o bardziej rozbudowanych badaniach.

Odkrycie nowych rodzin enzymów ART i kinaz białkowych może mieć bezpośrednie znaczenie dla rozwoju nowych terapii medycznych. Enzymy te są często kluczowymi regulatorami w patogenezie różnych chorób, w tym infekcji bakteryjnych (jak w przypadku *Legionella*) oraz chorób przewlekłych, takich jak nowotwory czy choroby neurodegeneracyjne. Nowe rodziny mogą ujawnić wcześniej nieznane szlaki molekularne i mechanizmy chorobotwórcze, które staną się potencjalnymi celami terapeutycznymi.

Nowe rodziny ART mogą być zaangażowane w modyfikacje posttranslacyjne, które regulują aktywność białek w komórkach gospodarza podczas infekcji. Identyfikacja tych enzymów może przyczynić się do opracowania leków celujących w specyficzne modyfikacje, co może blokować zdolność patogenów do przeżycia i infekcji.

Nowe rodziny kinaz białkowych mogą ujawniać nowe mechanizmy regulacji szlaków sygnalizacyjnych istotnych w procesie rozwoju i progresji nowotworów. Dzięki temu mogą powstać nowe inhibitory kinaz, które celują w specyficzne białka zaangażowane w procesy nowotworzenia.

Podsumowując, uzyskane wyniki nie tylko zwiększają naszą wiedzę o strukturze i funkcji enzymów, ale także mogą przyczynić się do rozwoju nowych strategii terapeutycznych. Dzięki temu badania oparte na podobieństwach strukturalnych mają potencjał, by w przyszłości odgrywać kluczową rolę zarówno w podstawowych badaniach biologicznych, jak i w medycynie spersonalizowanej oraz rozwoju nowych leków.

7. Bibliografia

1. Ranganathan, S., Cannataro, M. & Khan, A. M. *Encyclopedia of Bioinformatics and Computational Biology* (Elsevier, 2025).

2. Higgs, P. D. & Attwood, T. K. *Bioinformatyka i ewolucja molekularna.* (Blackwell Publishing, 2010)

3. Hogeweg, P. The Roots of Bioinformatics in Theoretical Biology. *PLoS Comput Biol* **7**, e1002021 (2011).

4. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

5. Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* **49**, 463–481 (1951).

6. *Dayhoff, M.O. et al. Atlas of Protein Sequence and Structure. Vol. 1 (National Biomedical Research Foundation, Silver Spring [Md.], 1965).*

7. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441–448 (1975).

8. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **41**, D36-42 (2013).

9. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **53**, D609–D617 (2025).

10. Stoesser, G. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **30**, 21–26 (2002).

11. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

12. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

13. Lesk, A. M. *Introduction to Protein Science: Architecture, Function, and Genomics*. (Oxford University Press, Oxford, 2010).

14. Berg, J. M., Stryer, L. *Biochemia* (Wydawnictwo Naukowe PWN, Warszawa, 2022).

15. Mount, D. W. *Bioinformatics: Sequence and Genome Analysis* (Laboratory Press, Cold Spring Harbor, N.Y., 2004).

16. Xiong, J. *Podstawy bioinformatyki* (Wydawnictwa Uniwersytetu Warszawskiego, Warszawa, 2009).

17. Brown, T. A. *Genomy* (Wydawnictwo Naukowe PWN, Warszawa, 2019).

18. Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147–157 (1982).

19. Breaker, R. R. DNA enzymes. *Nat Biotechnol* **15**, 427–431 (1997).

20. Köhler, T. *et al.* DNAzymes as catalysts for L-tyrosine and amyloid β oxidation. *ACS Omega* **5**, 7059–7064 (2020).

21. Cooper, G. M. The Central Role of Enzymes as Biological Catalysts. In *The Cell: A Molecular Approach. 2nd edition* (Sinauer Associates, 2000).

22. Ribeiro, A. J. M., Tyzack, J. D., Borkakoti, N., Holliday, G. L. & Thornton, J. M. A global analysis of function and conservation of catalytic residues in enzymes. *The Journal of Biological Chemistry* **295**, 314 (2019).

23. Price, N. C. What is meant by 'competitive inhibition'? *Trends in Biochemical Sciences* **4**, N272–N273 (1979).

24. Wu, P., Clausen, M. H. & Nielsen, T. E. Allosteric small-molecule kinase inhibitors. *Pharmacol Ther* **156**, 59–68 (2015).

25. Strelow, J. M. A perspective on the kinetics of covalent and irreversible inhibition. *SLAS Discov* **22**, 3–20 (2017).

26. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res* **28**, 304–305 (2000).

27. Corda, D. & Di Girolamo, M. Functional aspects of protein mono-ADP-ribosylation. *EMBO J* **22**, 1953–1958 (2003).

28. Aravind, L., Zhang, D., de Souza, R. F., Anand, S. & Iyer, L. M. The natural history of ADP-ribosyltransferases and the ADP-ribosylation system. *Curr Top Microbiol Immunol* **384**, 3–32 (2015).

29. Cohen, M. S. & Chang, P. Insights into the biogenesis, function, and regulation of ADP-ribosylation. *Nat Chem Biol* **14**, 236–243 (2018).

30. Ueda, K. & Hayaishi, O. ADP-ribosylation. *Annual Review of Biochemistry* **54**, 73–100 (1985).

31. Sugimura, T. & Miwa, M. Poly(ADP-ribose). Historical perspective. *Molecular and Cellular Biochemistry* **138**, 5–12 (1994).

32. Munnur, D. & Ahel, I. Reversible mono-ADP-ribosylation of DNA breaks. *FEBS J* **284**, 4002–4016 (2017).

33. Koch-Nolte, F.(ed). *Endogenous ADP-Ribosylation* (Springer International Publishing, 2015). DOI

34. Koch-Nolte, F., Reche, P., Haag, F. & Bazan, F. ADP-ribosyltransferases: plastic tools for inactivating protein and small molecular weight targets. *J Biotechnol* **92**, 81–87 (2001).

35. Hassa, P. O., Haenni, S. S., Elser, M. & Hottiger, M. O. Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going? *MMBR* **70**, 789–829 (2006).

36. Laing, S., Unger, M., Koch-Nolte, F. & Haag, F. ADP-ribosylation of arginine. *Amino Acids* **41**, 257–269 (2011).

37. Klockgether, J. & Tümmler, B. Recent advances in understanding Pseudomonas aeruginosa as a pathogen. *F1000Res* **6**, 1261 (2017).

38. Fontana, P. *et al.* Serine ADP-ribosylation in *Drosophila* provides insights into the evolution of reversible ADP-ribosylation signalling. *Nat Commun* **14**, 3200 (2023).

39. Tan, J. *et al.* Molecular basis of threonine ADP-ribosylation of ubiquitin by bacterial ARTs. *Nat Chem Biol* **20**, 463–472 (2024).

40. Bartlett, E. *et al.* Interplay of histone marks with serine ADP-ribosylation. *Cell Rep* **24**, 3488–3502.e5 (2018).

41. Groslambert, J., Prokhorova, E. & Ahel, I. ADP-ribosylation of DNA and RNA. *DNA Repair (Amst)* **105**, 103144 (2021).

42. De Vos, M., Schreiber, V. & Dantzer, F. The diverse roles and clinical relevance of PARPs in DNA damage repair: Current state of the art. *Biochemical Pharmacology* **84**, 137–146 (2012).

43. Bai, P. Biology of poly(ADP-ribose) polymerases: the factotums of cell maintenance. *Molecular Cell* **58**, 947–958 (2015).

44. Zarkovic, G. *et al.* Characterization of DNA ADP-ribosyltransferase activities of PARP2 and PARP3: new insights into DNA ADP-ribosylation. *Nucleic Acids Research* **46**, 2417–2431 (2018).

45. Pascal, J. M. & Ellenberger, T. The rise and fall of poly(ADP-ribose): An enzymatic perspective. *DNA Repair* **32**, 10–16 (2015).

46. Gupte, R., Liu, Z. & Kraus, W. L. PARPs and ADP-ribosylation: recent advances linking molecular functions to biological outcomes. *Genes Dev* **31**, 101–126 (2017).

47. Gagné, J.-P. *et al.* Proteome-wide identification of poly(ADP-ribose) binding proteins and poly(ADP-ribose)-associated protein complexes. *Nucleic Acids Res* **36**, 6959–6976 (2008).

48. Jungmichel, S. *et al.* Proteome-wide identification of poly(ADP-ribosyl)ation targets in different genotoxic stress responses. *Molecular Cell* **52**, 272–285 (2013).

49. Zhang, Y., Wang, J., Ding, M. & Yu, Y. Site-specific characterization of the Asp- and Glu-ADP-ribosylated proteome. *Nat Methods* **10**, 981–984 (2013).

50. Bock, F. J., Todorova, T. T. & Chang, P. RNA regulation by poly(ADP-ribose) polymerases. *Mol Cell* **58**, 959–969 (2015).

51. Munnur, D. *et al.* Reversible ADP-ribosylation of RNA. *Nucleic Acids Research* **47**, 5658–5669 (2019).

52. Vyas, S., Chesarone-Cataldo, M., Todorova, T., Huang, Y.-H. & Chang, P. A systematic analysis of the PARP protein family identifies new functions critical for cell physiology. *Nat Commun* **4**, 2240 (2013).

53. Bock, F. J. & Chang, P. New directions in poly(ADP-ribose) polymerase biology. *FEBS J* **283**, 4017–4031 (2016).

54. Hottiger, M. O., Hassa, P. O., Lüscher, B., Schüler, H. & Koch-Nolte, F. Toward a unified nomenclature for mammalian ADP-ribosyltransferases. *Trends in Biochemical Sciences* **35**, 208–219 (2010).

55. Bazan, F. J. & Koch-Nolte, F. Sequence and Structural Links between Distant ADP-Ribosyltransferase Families. In *ADP-Ribosylation in Animal Tissues: Structure, Function, and Biology of Mono (ADP-ribosyl) Transferases and Related Enzymes* (eds. Haag, F. & Koch-Nolte, F.) 99–107 (Springer US, Boston, MA, 1997). DOI

56. Aravind, L. & de Souza, R. F. Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions. *Molecular BioSystems* **8**, 1661–1677 (2012).

57. Maresso, A. W., Deng, Q., Pereckas, M. S., Wakim, B. T. & Barbieri, J. T. Pseudomonas aeruginosa ExoS ADP-ribosyltransferase inhibits ERM phosphorylation. *Cell Microbiol* **9**, 97–105 (2007).

58. Komander, D. & Randow, F. Strange new world: bacteria catalyze ubiquitylation via ADP ribosylation. *Cell Host & Microbe* **21**, 127–129 (2017).

59. Suskiewicz, M. J., Prokhorova, E., Rack, J. G. M. & Ahel, I. ADP-ribosylation from molecular mechanisms to therapeutic implications. *Cell* **186**, 4475–4495 (2023).

60. Lüscher, B. *et al.* ADP-ribosyltransferases, an update on function and nomenclature. *The FEBS Journal* **n/a**, 289, 739-7410 (2021).

61. Cheek, S., Zhang, H. & Grishin, N. V. Sequence and structure classification of kinases. *J Mol Biol* **320**, 855–881 (2002).

62. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).

63. Manning, G., Plowman, G. D., Hunter, T. & Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* **27**, 514–520 (2002).

64. Scheeff, E. D. & Bourne, P. E. Structural evolution of the protein kinase-like superfamily. *PLoS Computational Biology* **1**, e49 (2005).

65. Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C. & Manning, G. Structural and functional diversity of the microbial kinome. *PLoS Biology* **5**, e17 (2007).

66. Taylor, M., Ross, K. & Bentham, R. Legionella, protozoa, and biofilms: interactions within complex microbial systems. *Microb Ecol* **58**, 538–547 (2009).

67. McDade, J. E. *et al.* Legionnaires' disease: isolation of a bacterium and demonstration of its role in other respiratory disease. *N Engl J Med* **297**, 1197–1203 (1977).

68. Fraser, D. W. *et al.* Legionnaires' disease: description of an epidemic of pneumonia. *N Engl J Med* **297**, 1189–1197 (1977).

69. Brenner, D. J., Steigerwalt, A. G. & McDade, J. E. Classification of the Legionnaires' disease bacterium: *Legionella pneumophila*, genus novum, species nova, of the family Legionellaceae, familia nova. *Ann Intern Med* **90**, 656–658 (1979).

70. Newton, H. J., Ang, D. K. Y., van Driel, I. R. & Hartland, E. L. Molecular pathogenesis of infections caused by Legionella pneumophila. *Clin Microbiol Rev* **23**, 274–298 (2010).

71. Mondino, S. *et al.* Legionnaires' disease: state of the art knowledge of pathogenesis mechanisms of Legionella. Annu Rev Pathol Mech Dis **15**, 439–466 (2020).

72. Kanatani, J. *et al.* Detection of *Legionella* species, the influence of precipitation on the amount of *Legionella* DNA, and bacterial microbiome in aerosols from outdoor sites near asphalt roads in Toyama Prefecture, Japan. *BMC Microbiology* **21**, 215 (2021).

73. Steiner, B., Weber, S. & Hilbi, H. Formation of the *Legionella*-containing vacuole: phosphoinositide conversion, GTPase modulation and ER dynamics. *Int J Med Microbiol* **308**, 49–57 (2018).

74. Horwitz, M. A. The Legionnaires' disease bacterium (*Legionella pneumophila*) inhibits phagosome-lysosome fusion in human monocytes. *J Exp Med* **158**, 2108–2126 (1983).

75. Horwitz, M. A. Formation of a novel phagosome by the Legionnaires' disease bacterium (*Legionella pneumophila)* in human monocytes. *J Exp Med* **158**, 1319–1331 (1983).

76. Chauhan, D. & Shames, S. R. Pathogenicity and Virulence of *Legionella:* Intracellular replication and host response. *Virulence* **12**, 1122–1144 (2021).

77. Bruckert, W. M. & Abu Kwaik, Y. Complete and ubiquitinated proteome of the-Legionella-containing vacuole within human macrophages. *J Proteome Res* **14**, 236–248 (2015).

78. Shames, S. R. Eat or be eaten: Strategies used by Legionella to acquire host-derived nutrients and evade lysosomal degradation. *Infect Immun* **91**, e00441-22 (2023).

79. Price, C. T. D., Richards, A. M. & Abu Kwaik, Y. Nutrient generation and retrieval from the host cell cytosol by intra-vacuolar *Legionella pneumophila*. *Front Cell Infect Microbiol* **4**, 111 (2014).

80. Fonseca, M. V. & Swanson, M. S. Nutrient salvaging and metabolism by the intracellular pathogen *Legionella pneumophila. Front Cell Infect Microbiol* **4**, 12 (2014).

81. Schunder, E. *et al.* Amino acid uptake and metabolism of *Legionella pneumophila* hosted by *Acanthamoeba castellanii*. *J Biol Chem* **289**, 21040–21054 (2014).

82. Ge, J. & Shao, F. Manipulation of host vesicular trafficking and innate immune defence by *Legionella* Dot/Icm effectors. *Cellular Microbiology* **13**, 1870–1880 (2011).

83. Grishin, A. M., Beyrakhova, K. A. & Cygler, M. Structural insight into effector proteins of Gram-negative bacterial pathogens that modulate the phosphoproteome of their host. *Protein Science: A Publication of the Protein Society* **24**, 604 (2015).

84. Black, M. H. *et al.* A *Legionella* effector ADP-ribosyltransferase inactivates glutamate dehydrogenase. *J Biol Chem* **296**, 100301 (2021).

8. Publikacje wchodzące w skład rozprawy wraz z właściwymi oświadczeniami współautorów

A survey of ADP-ribosyltransferase families in the pathogenic *Legionella*

Authors: Marianna Krysińska[1], Marcin Gradowski[1], Bartosz Baranowski[2], Krzysztof Pawłowski[3,4] and Małgorzata Dudkiewicz[1]

Affiliation:

[1] Warsaw University of Life Sciences — SGGW, Warsaw, Poland

[2] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland

[3] University of Texas Southwestern Medical Center, Dallas, TX, USA

[4] Howard Hughes Medical Institute, Dallas, TX, USA

Correspondence: Krzysztof Pawłowski (krzysztof.pawlowski@utsouthwestern.edu)

Małgorzata Dudkiewicz (malgorzata_dudkiewicz@sggw.edu.pl)

Abstract

Background. ADP-ribosyltransferases (ARTs) are a superfamily of enzymes implicated in various cellular processes, including pathogenic mechanisms. The *Legionella* genus, known for causing Legionnaires' disease, possesses diverse ART-like effectors. This study explores the proteomes of 41 *Legionella* species to bioinformatically identify and characterise novel ART-like families, providing insights into their potential roles in pathogenesis and host interactions.

Methods. We conducted a comprehensive bioinformatic survey of 41 *Legionella* species to identify proteins with significant sequence or structural similarity to known ARTs. Sensitive sequence searches were performed to detect candidate ART-like families. Subsequent validation, including structure prediction of such families, was achieved using artificial intelligence-driven tools, such as AlphaFold. Comparative analyses were performed to assess sequence and structural similarities between the novel ART-like families and known ARTs.

Results. Our analysis identified 63 proteins with convincing similarity to ARTs, organised into 39 ART-like families, including 26 novel families. Key findings include:

- DUF2971 family: exhibits sequence similarity to DarT toxins and other DNA-acting ARTs.
- DUF4291 family: the largest newly identified family shows structural and sequence similarity to the diphtheria toxin, suggesting the ability to modify proteins.

Most members of the novel ART families are predicted effectors. Although experimental validation of the predicted ART effector functions is necessary, the novel ART-like families identified present promising targets for understanding *Legionella* pathogenicity and developing therapeutic strategies. We publish a complete catalogue of our results in the astARTe database: http://bioinfo.sggw.edu.pl/astarte/.

Keywords: ADP-ribosyltransferase, Legionella, bacterial effectors, protein function prediction, protein structure prediction, protein sequence analysis, astARTe

Introduction

ADP-ribosyltransferases (ARTs) are enzymes that chemically modify proteins, nucleic acids and small molecules by covalently attaching an ADP-ribose derived from oxidised nicotinamide adenine dinucleotide (NAD^+) via N-, O- or S-glycosidic bonds (Corda, 2003; Aravind et al., 2015; Cohen & Chang, 2018). ADP-ribosyltransferases are typically classified as monoARTs or poly(ADP-ribose) polymerases (PARPs), catalysing the addition of a single ADP-ribose moiety or polymers of ADP-ribose to its targets. Further, the ADP-ribose chain catalysed by PARPs can be either linear or branched (Ueda & Hayaishi, 1985; Sugimura & Miwa, 1994; Aravind et al., 2015; Munnur & Ahel, 2017).

ADP-ribosyltransferases are common in nature and are found in all domains of life and in viruses. The characteristic feature of the ART superfamily is a conserved structural core formed by a split beta-sheet formed by six or seven antiparallel beta strands and surrounded by alpha helices. As a result, the NAD^+ molecule is sandwiched between the two halves of the split beta sheet (Aravind & de Souza, 2012; Aravind et al., 2015; Cohen & Chang, 2018). Almost all ADP-ribosyltransferases have additional protein domains accompanying the ART catalytic domain that target enzymes to their substrates and specific cell locations (Vyas et al., 2013; Bock & Chang, 2016; Cohen & Chang, 2018). In contrast to the structural similarity, sequence conservation between ART families is poor, which is also manifested by the diverse amino acid composition of the active sites. Classifying by active site types, the ART superfamily contains four main clades: HHh, RSE, HYE, and the so-called atypical clade (Aravind et al., 2015).

Mono-ADP-ribosylation is believed to have originally appeared in bacteria as a defence mechanism against viruses or bacteria (Aravind et al., 2015; Koch-Nolte, 2015). Poly-ADP-ribosylation emerged in Eukaryotes as a process involved in DNA-repair, modulation of chromatin structure and programmed cell death (Morales et al., 2014).

During infection, ADP-ribosylation of host proteins by bacterial effectors can change the apoptotic potential of the cell, disrupt the organisation of the actin cytoskeleton, alter the organisation of cell membranes, and interfere with the cellular immune response (Maresso et al.,

2007; Hottiger et al., 2010; Aravind et al., 2015; Komander & Randow, 2017; Klockgether & Tümmler, 2017; Cohen & Chang, 2018). Other toxins ADP-ribosylate nucleotides in single-stranded DNA (ssDNA). This leads to ADP-ribosylation of the origin of chromosomal replication of DNA and hence inhibition of cell growth (Jankevicius et al., 2016). A similar mechanism is used by bacteria for self-defence during phage infection. For example, DarT from *M. tuberculosis* selectively ADP-ribosylates thymidine nucleotides within phage ssDNA which blocks phage DNA replication (Ritter et al., 2003).

Legionella is a genus comprising free-living, biofilm-associated, or host-associated bacteria (Taylor, Ross & Bentham, 2009). *Legionella* infection in humans can cause severe atypical pneumonia called legionellosis or a milder version called Pontiac fever (Fraser et al., 1977; McDade et al., 1977; Brenner, Steigerwalt & McDade, 1979; Newton et al., 2010; Mondino et al., 2020; Kanatani et al., 2021). Currently, the genus *Legionella* comprises more than 70 different species of bacteria, about half of which have been found to be pathogenic to humans. Additionally, the majority are considered potential human pathogens (Mercante & Winchell, 2015; Chambers et al., 2021; Kanatani et al., 2021). After entering the host cell, *Legionella* creates a special vacuole in which it can survive and reproduce called the Legionella Containing Vacuole (LCV) (Samrakandi et al., 2002). Effector proteins translocated by *Legionella* into the host cell can profoundly alter the host cell's behaviour, which greatly facilitates *Legionella* replication and pathogenesis. These alterations include safeguarding the LCV from degradation, manipulating the endomembrane system, dampening the host immune response, and exploiting the hosts resources (Segal, Feldman & Zusman, 2005; Isberg, O'Connor & Heidtman, 2009; Burstein et al., 2016; Gomez-Valero et al., 2019). The most thoroughly-characterized *Legionella* species, *L. pneumophila*, translocates over 330 effectors (Burstein et al., 2016; Wexler et al., 2022). Three recently described *L. pneumophila* effectors with ART functions are lpg0181/Lart1 (Black et al., 2021), lpg0080 (Fu et al., 2022) and the SidE family of all-in-one Ub ligases. The first was identified as an ADP-ribosylating factor for a specific class of NAD+-dependent glutamate dehydrogenase (GDH) enzymes found in fungi and protists, which includes many natural hosts of *Legionella* (Black et al., 2021). The second (lpg0080) was identified as a mono-ADP-ribosyltransferase that localises to mitochondria in host cells and ADP-ribosylates

the ADP/ATP translocase, which impairs its activity (Fu et al., 2022). The third (SidE family) shows atypical activity by attaching phosphoribosyl ubiquitin (PR-Ub) to serine residues on substrates, using the phosphodiesterase and mono-ADP-ribosyltransferase domains. SidE proteins modify small Rab GTPases, disrupting vesicle movement and cell membrane dynamics, preventing the maturation of phagosomes into bacterial lysosomes, and thus allowing *Legionella* to establish a replication niche (Akturk et al., 2018).

In this article, basing on our expertise in identifying novel enzyme families (Black et al., 2021; Wyżewski et al., 2021), and motivated by many known bacterial effectors and toxins being ARTs (Black et al., 2021; Fu et al., 2022), we undertook a bioinformatics search and survey of novel ART-like domains in 41 *Legionella* proteomes. We identified 26 novel ART-like families, six of which we describe in detail. Presenting a catalogue of ART-like catalytic domains in *Legionella,* we compare the known and the novel families, predict biological functions for the novel ones, establish their evolutionary history and occurrence across the bacterial world.

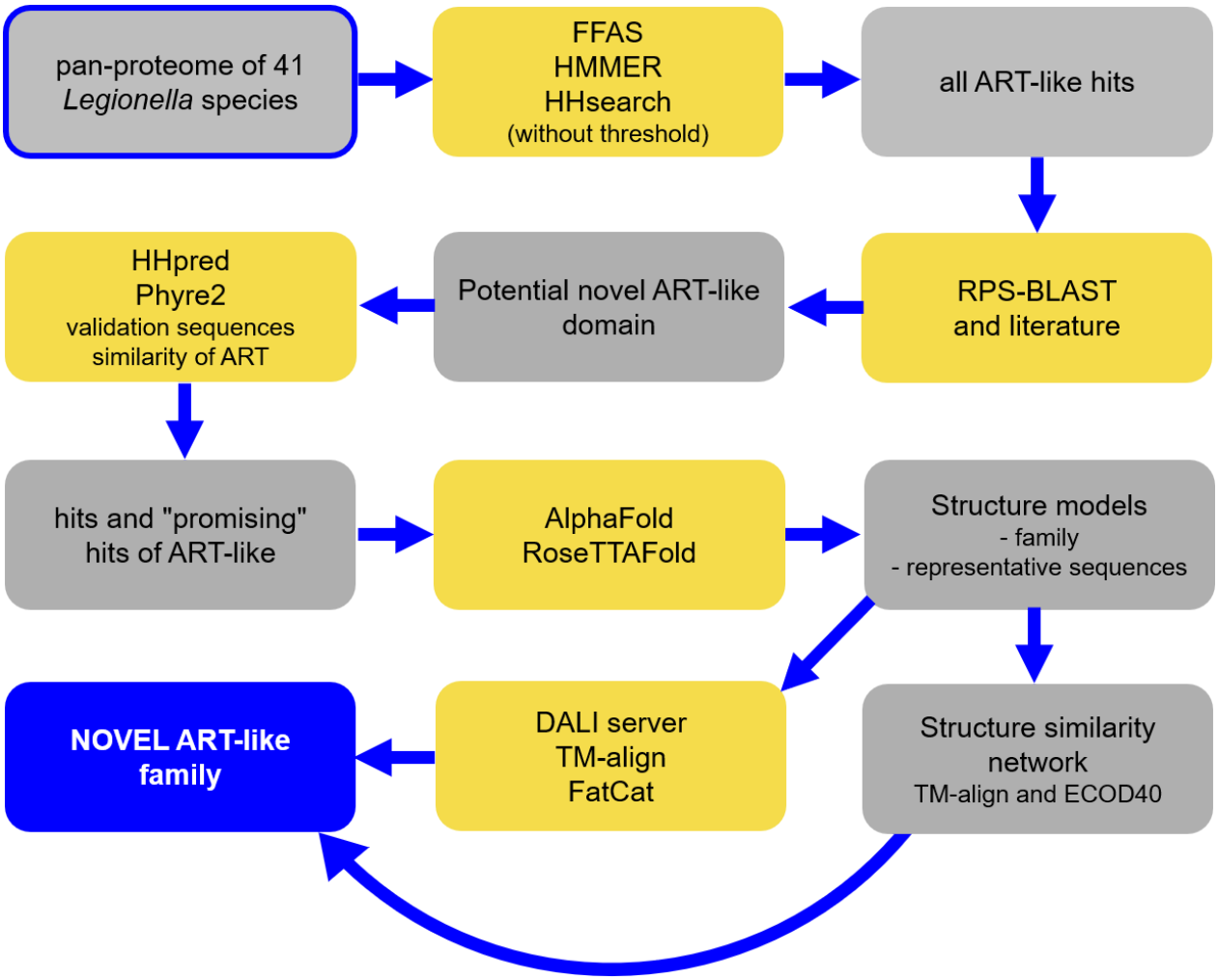


Figure 1. Methods used to identify new ARTs in *Legionella* sp.; grey rectangles – data, yellow rectangles – tools used.

Materials and methods

Sequence data

Proteomes for 41 species of Legionella were provided by the study by Burstein et al. (Burstein et al., 2016). To reduce the computational burden, the sequences were clustered by sequence similarity in two steps using the CD-HIT algorithm with sequence identity cut-offs 70% and 50% (Li, Jaroszewski & Godzik, 2001, 2002; Li & Godzik, 2006; Huang et al., 2010). Next, 21,616 resulting protein sequences from 41 *Legionella* proteomes were cut into fragments of 300 amino

acids with an overlap of 100 amino acids. This method facilitates the detection of domains in multi-domain proteins, while using an overlap can help detect domains at fragment boundaries.

Remote homology detection

For distant similarity recognition to ART families, four methods were used: 1) the profile-profile alignment and fold recognition algorithm – FFAS (Xu et al., 2014) (searching the COG, Hsapiens, PDB, SCOP, ECOD and Pfam databases), 2) homology detection and structure prediction method HHpred (Gabler et al., 2020) and HHsearch pipeline (databases searched: Pfam, PDB, SCOP) that uses hidden Markov model HMM-to-HMM comparison; 3) a similar method, Phyre2 (Kelley et al., 2015), which additionally models 3D structure of query and compares it with 3D models library, and 4) homology detection by comparing a profile-HMM to either a single sequence or a database of sequences – HMMER algorithm (Eddy, 2011). Standard parameters and significance thresholds were selected except for the HHpred algorithm, where several parameters have been modified: MSA (Multiple Sequence Alignment) generation method (all options were used); Alignment Mode (local:norealign; local:realign; global:realign); Min probability in hitlist (%) used 10% or 20%; Min coverage of MSA hits (%) used 10% or 20%; MSA generation iteration number was 0, 3, 5 or 8 and E-value cutoff was 1e-3, 0.05 or 0.1 (see Fig. 1 and Suppl. Tables S1-S5).

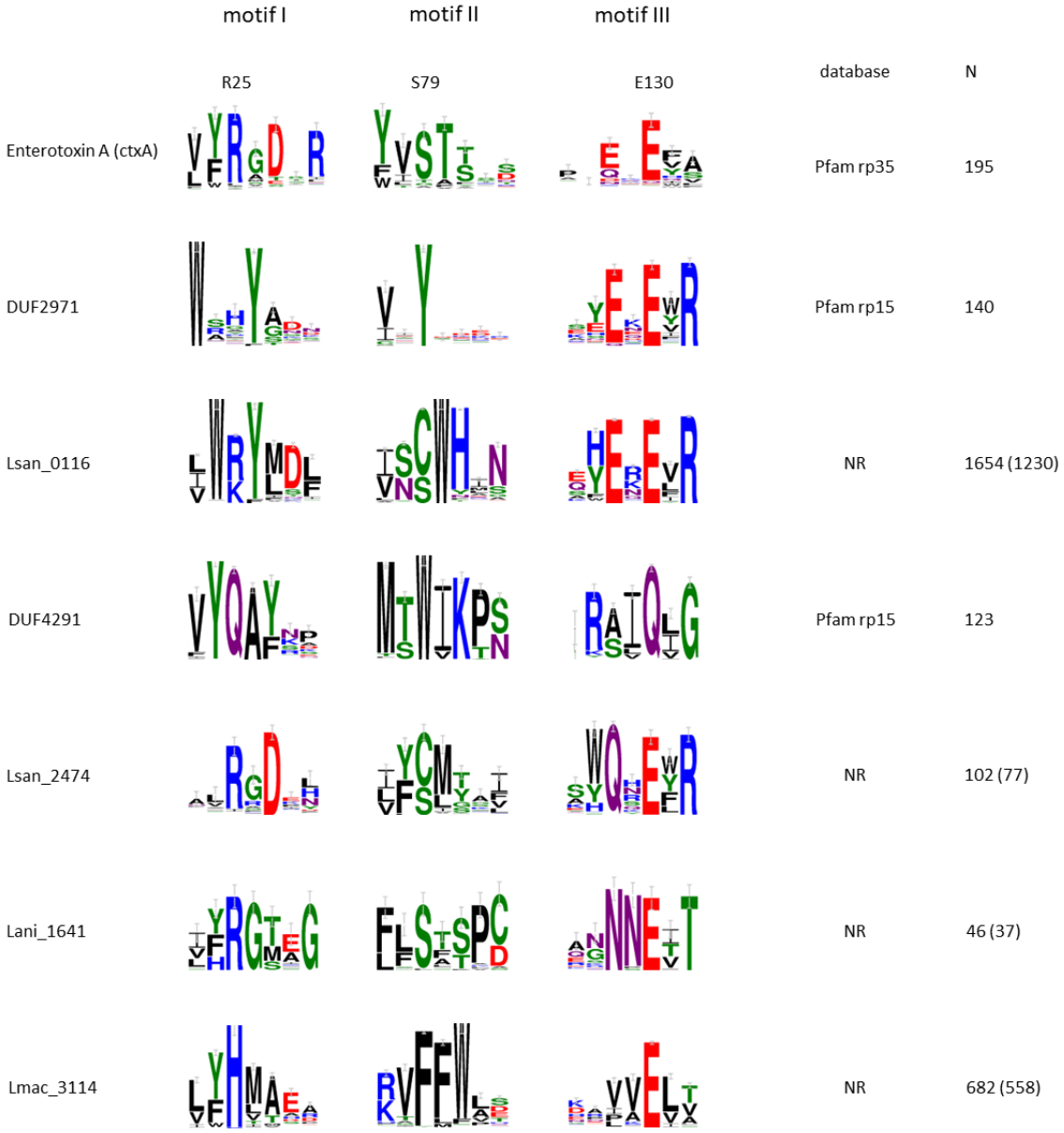


Figure 2. The tripartite active site signature [R/H]x[D/T] - [S/Y][T/x][S/x] - [Q/E]x[Q/E/D] is well conserved in most new ART families. Active site motif sequence logos for selected families. An archetypal ADP-ribosyltransferase, Enterotoxin a, ctxA, is shown at the top. Numbering of the residues (top row) according to ctxA sequence (O'Neal et al., 2004). The DB column indicates the source database of ART sequences used for logos. N is the number of homologous sequences from the BLAST search (E-value = 1e-4). In parentheses - numbers of homologous sequences after CD-HIT clustering at 99%

sequence identity.

Multiple sequence alignments and sequence logos

Members of novel families were manually collected using BLAST (Altschul et al., 1997; Boratyn et al., 2012) and non-redundant sequence database (NR) (Sayers et al., 2021) (E-value = 1e-4 or with default options). In cases, when BLAST found less than 5 homologs, PSI-BLAST was used (3 or 5 iterations with default options) (Boratyn et al., 2012). Multiple sequence alignments were made using the MAFFT algorithm with default settings (Katoh et al., 2002; Katoh, Rozewicki & Yamada, 2019). Next, the sequence logos were prepared using WebLogo3 (Crooks et al., 2004). Colouring according to amino acid chemistry was used. For the logos, the alignments were processed with an in-house script that removes alignment columns that contain gaps in the reference sequence.

Structure modelling and comparison

The structures of representative proteins from new ART-like families were modelled with the use of two different tools: AlphaFold (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2023). Comparisons of structures were performed with FATCAT (Li et al., 2020), Dali server (Holm, 2020) and TM-align (Zhang & Skolnick, 2005) (Suppl. Table S6 and data made available in the astARTe online database).

Visual clustering of families (sequence analysis of families relations, Fig. 5)

The CLANS program (Frickey & Lupas, 2004) was used to visualise the relationships between clusters of ADP-ribosyltransferases families. The collection of the known ADP-ribosyltransferase domains was obtained from the Pfam database, as the sets of sequences from families of ADP-ribosyl clan CL0084 – Pfam 35.0 (Mistry et al., 2021) (see Suppl. Table S9B). Then the sequence sets from ART families not yet included in the ADP-ribosyl Pfam clan were added. Next, collected data were supplemented with sequence sets representing the following ART-like domain families not included in the Pfam database (see Suppl. Table S9B) and, in the end, we added all 26 newly predicted ART families with their homologs collected by one of the three methods with cut-off threshold of 99%: 1) BLAST, NR, E-value = 1e-4; 2) PSI-BLAST, two

iterations, NR, default; 3) PSI-BLAST, three iterations, NR, default (see Suppl. Table S9C). Additionally, two families DUF2971 (PF11185, rp15 sequence set) and DUF4291, rp15 set were included (see all details in Suppl. Table S9A).

Next, the whole collection of putative and known domain sequences were prepared for the CLANS procedure by clustering each family separately with CD-HIT and selecting representatives at a 70% (known domains) or 99% (putative domains) sequence identity threshold (Huang et al., 2010). CLANS algorithm was run with following parameters: BLOSUM45 (scoring matrix) and E-value = 1 or E-value = 10e-2.

HMMs of known families for searching sequence databases and profile databases

Sequence sets for HMM construction are collected in a similar manner as the data for CLANS analysis. The most important difference: all sequence sets for families present in the Pfam database were downloaded as rp75 sequence sets from Pfam 34.0 (Mistry et al., 2021) (see all details in Suppl. Table S9B).

In the next step, we rejected short sequences (less than 50 amino acids) and applied a sequence similarity cut-off threshold of 80% (CD-HIT). The ClustalO program (Sievers & Higgins, 2018) was used to build multiple sequence alignments. HMMs were prepared by the hmmbuild program from the HMMER software package (Eddy, 2011).

Species dendrogram and estimating numbers of homologs

Dendrogram was made from alignment of 16S rRNA sequences for the species in the family *Legionellaceae*. The sequences were downloaded from NCBI ("Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information") and SILVA database (Glöckner et al., 2017). The alignment was performed via the ngphylogeny.fr server (Dereeper et al., 2008; Lemoine et al., 2019) using the Muscle 3.8.31 algorithm with default parameters (Edgar, 2004). A dendrogram was constructed using the PhyML method with the Approximate Likelihood-Ratio Test (aLRT): SH-like (Guindon et al., 2010). Homologs for all families were collected using the hmmsearch algorithm (E-value = 1e-4) (Eddy, 2011) by using

the HMMs of known and new ART and ART-like families as queries against the RefSeq database (Sayers et al., 2021) (see all details in Suppl. Table S9C).

Species dendrogram was visualised using the iTol server (Letunic & Bork, 2021).

Structure similarity network

Sequences for each of the new families were collected (see Suppl. Table S9D) and protein domains were modelled using ColabFold (Mirdita et al., 2022) (an implementation of AlphaFold (Jumper et al., 2021) using the fast homology search of MMseqs2 (Steinegger & Söding, 2017; Mirdita, Steinegger & Söding, 2019)) or ESMFold (Lin et al., 2023). The resulting 432 models and 59 ART structures from the ECOD40 (version 2022/09/12, develop286) database were compared (all to all) using the TM-align algorithm (Zhang & Skolnick, 2005). Based on the obtained TM-score values (nTM-score, score normalised against a smaller structure; with threshold better than 0.3, and processed: $ln(nTM\text{-}score) \times (-1)$) for each pair of structures (so that the new score is in the range 0-1), a structure similarity network was constructed in CLANS algorithm (see Suppl. Table S9D). We checked the quality of the models by analysing the pLDDT and pTM parameters for each model (see Suppl. Table S9E). The same set of sequences was used for the sequence similarity network (Fig. 4).

NAD docking to models

We made models (ColabFold) for reference sequences from the 6 new families discussed in this article. NAD docking was performed using the HADDOCK 2.4 (High Ambiguity Driven protein-protein DOCKing) web service (van Zundert et al., 2016; Honorato et al., 2021) package with automatic (CPORT (Vries & Bonvin, 2011)) and manual indication of the binding site residues. Visualisation of the results was done in UCSFChimera (Pettersen et al., 2004) using ConSurf server (Ashkenazy et al., 2010) for mapping MSA conservation into reference sequences models for 6 families. Structures superposition and the closest structural homologues for models were obtained using the DALI server. Additionally, proteins representative of the families described were modelled with AlphaFold3 (Abramson et al., 2024) including NAD as a ligand.

The high-quality models obtained (pLDDT>90, pTM>90) confirmed the docking results – location and position of the NAD (data not shown).

Additional analyses

The subcellular localization was predicted by SubCons (Salvatore, Shu & Elofsson, 2018). Prediction of the presence of signal peptides was made in SignalP 6.0 (Community, 2021). Predictions of effector function were done using EffectiveDB (Arnold et al., 2009; Eichinger et al., 2016) and BastionX (Wang et al., 2021) with standard parameters. Analysis of genomic neighbourhoods was done using the ProFaNa tool (Baranowski & Pawłowski, 2023) and we used the BioCyc portal for detailed analysis of the operons ((Karp et al., 2017)). The occurrence of additional domains was analysed using Batch CD-Search (RPS-BLAST) (Marchler-Bauer et al., 2011).

Alternative approach to the detection of distantly related homologues

Alternatively, by using RoseTTAFold structure models for Pfam families, we found an ART-like family in *Legionella* that was not detectable by sequence similarity. Here, the pipeline included pairwise TM-align structure alignments between RoseTTAFold models provided in the Pfam 35.0 database and a set of reference ART domain structures from the ECOD database (ECOD40).

Definition of new families

To define what is meant by "novel family" we applied several criteria. A novel family is a set of proteins not described in the literature as ADP-ribosyltransferases and not annotated such, with similarity to known ARTs not detectable using standard sequence analysis methods – RPS-BLAST or Pfam HMM. Additionally, a family designated as "novel", based on CLANS clustering analysis, must be sufficiently different from known and other novel ART families.

Results

Search for ART superfamily members in *Legionella*

We started our survey with the pan-proteome from 41 *Legionella* species (see Methods). After data pre-processing (see Methods), 21,616 sequences were analysed by FFAS03, HMMER, and

HHsearch for sequence similarity to known protein domains. The raw results were filtered by a Python text-mining script that retrieved all ART-like hits using Pfam, PDB, and SCOP identifiers as well as keywords in protein names. In this way, 63 potential ART-like domains were identified exhibiting any similarity to known ART-like domains, including statistically non-significant similarity (Suppl. Tables S2-S4). In the next step, we applied RPS-BLAST validation and literature searches, and known ADP-ribosyltransferases were filtered out. The known hits included 21 proteins from 11 families: RES (Pfam identifier PF08808), PARP (PF00644), ART (PF01129), Dot_icm_IcmQ (PF09475), DarT (PF14487), DUF952 (PF06108), lpg0080, Lart1 and the poorly studied AbiGi family (PF10899), presumably part of the type IV toxin-antitoxin system, as well as the FRG family identified recently by Aravind (Aravind et al., 2015; Burroughs & Aravind, 2020) as an ART-like family present in *Legionella* (Suppl. Tables S1-S2). Forty-two potential ART-like FFAS-HHsearch-HHMER hits were not automatically identified as ART-like and were verified by other distant sequence similarity search methods (Phyre2, HHpred and CLANS analysis) and de novo structure modelling using the RoseTTAFold and AlphaFold methods supplemented with structural comparisons (FATCAT and Dali servers) (Suppl. Tables S5-S6). After discarding 3 false positive ARTs, remaining 39 atypical ART proteins were grouped in twenty-six novel families (see Suppl. Tables S6, S7 and S11).

We identified unequivocal active site signatures in 15 families. In 7 families, we found the presence of conservative substitutions in the predicted active sites (see Fig. 2 and Tab. 1). In the remaining 4 families, the active site is partly non-conserved which suggests they are pseudoenzymes, i.e. proteins that are homologous to active enzymes but presumably lack catalytic activity because of mutations to critical active site amino acids. We classified six of the new families into the HYE clade, one into the HHh clade, one into the atypical clade, and the rest (18 families) into the RSE clade (Fig. 3).

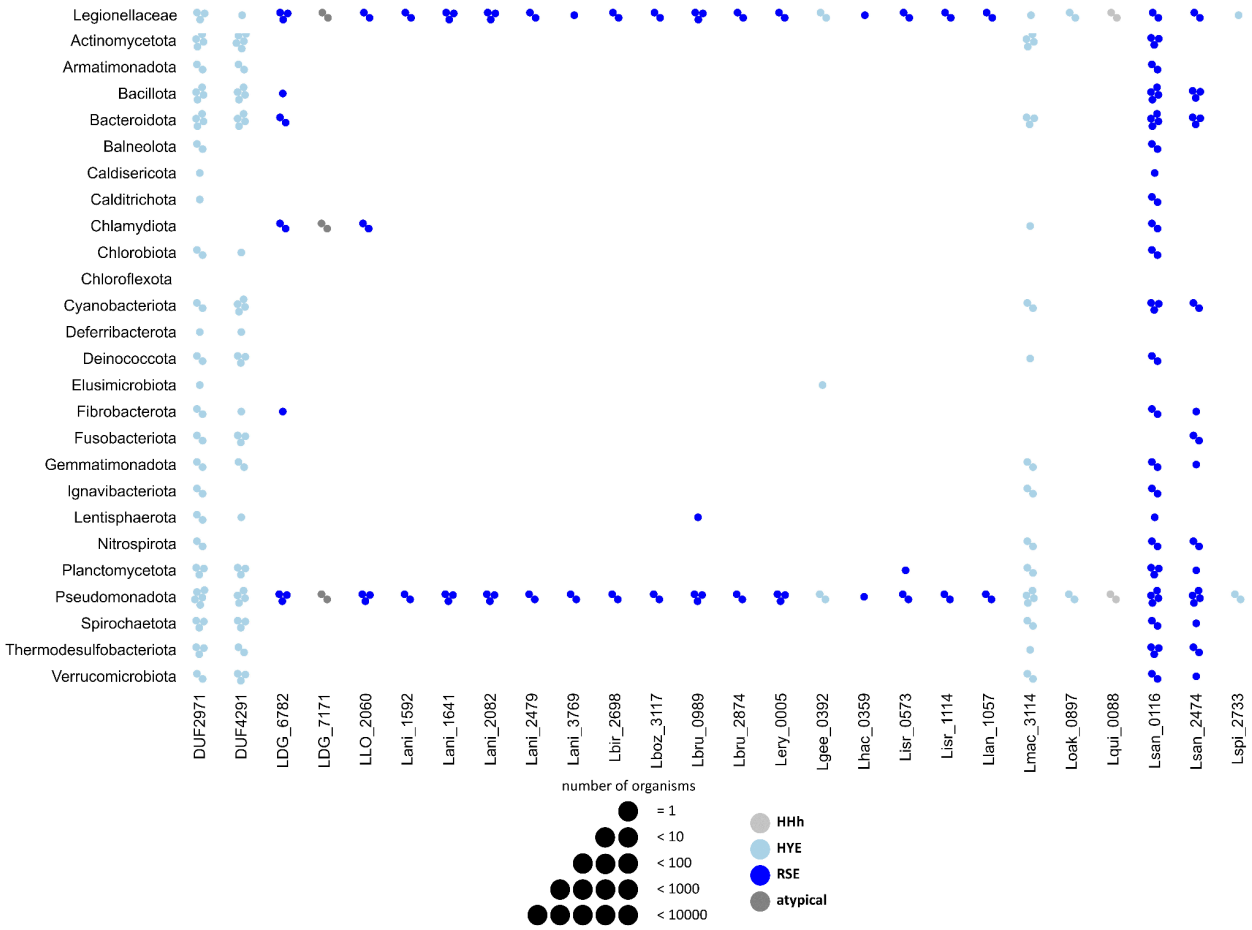


Figure 3. Taxonomic distribution of members of novel ART families in bacterial phyla. The family *Legionellaceae* was considered separately. The number of bacterial strains in which homologues of new families were found is shown on a logarithmic scale (see legend of the figure). Colours reflect ART clade membership.

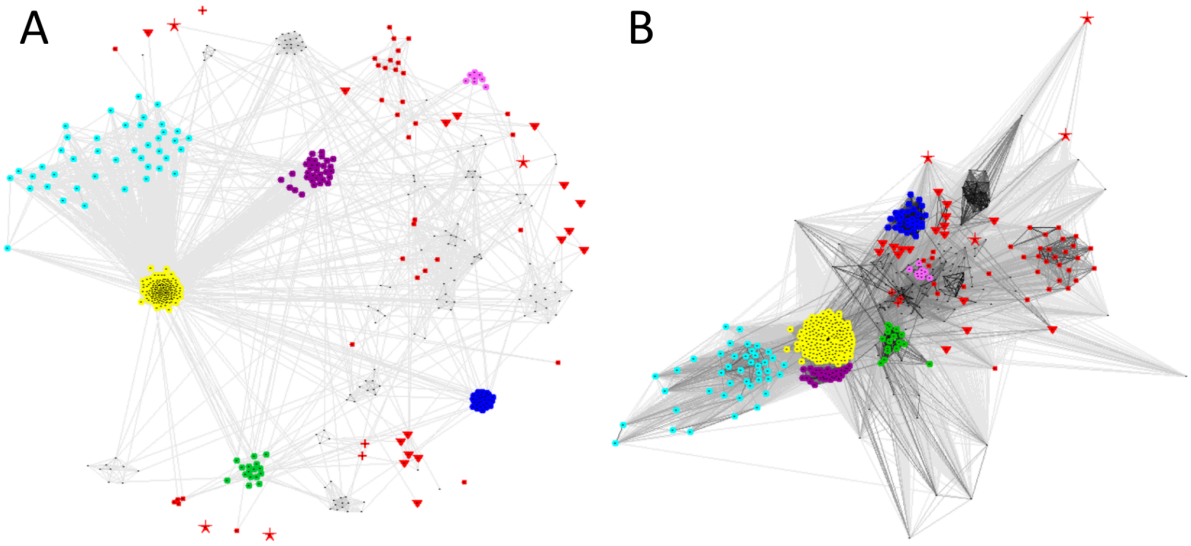


Figure 4. Sequence and structure similarity relationships between novel and known ART families. The CLANS graph shows: (A) the sequence similarities obtained using pairwise BLAST comparisons, taking into account significant and borderline significance similarities up to the E-value of 1 (BLOSUM45); (B) the structure similarities obtained using TM-align all-to-all comparisons (with TMscore threshold better than 0.3). Novel families discussed in detail shown: DUF2971 – turquoise, Lsan_0116 – yellow, DUF4291 – blue, Lsan_2474 – purple, Lani_1641 – pink and Lmac_3114 – green. Known ART families (as defined in the ECOD40 dataset) are marked as red (crosses – clade HHh, dots – clade RSE, triangles – clade HYE, asterisks – atypical clade).

CLANS sequence and structure similarity graphs capture distant relationships between the novel *Legionella* ART-like families and fifty-nine representative known ART domains from ECOD40 database (Cheng et al., 2014) (see Fig. 4 and Fig. 5). The CLANS graph is created by comparing sequences using pairwise alignments from BLAST or pairwise structure comparisons using TM-align. The CLANS sequence and structure similarity networks confirm the characteristics of ART-like superfamily of proteins, i.e., high sequence divergence (Fig. 4A) – even within a single clade – and a highly conserved tertiary structure core (Fig. 4B), also between clades. Even well-characterised ART proteins from the ECOD40 database do not always show sequence similarity high enough to be included in the graph, even with a very relaxed threshold: E-value = 1 (Fig. 4A). It is also apparent that there is more similarity within clades than between them, both when comparing sequences and structures. The small HHh clade, the minimal version of the

ART-like fold (only six strands, without any extended inserts), locates centrally in the structural CLANS graph. Our analysis showed that the novel ART-like families: 1) form separate clusters of proteins and 2) generally do not group together with established, well-studied ARTs, thus confirming their distinct family status within the ART-like superfamily.

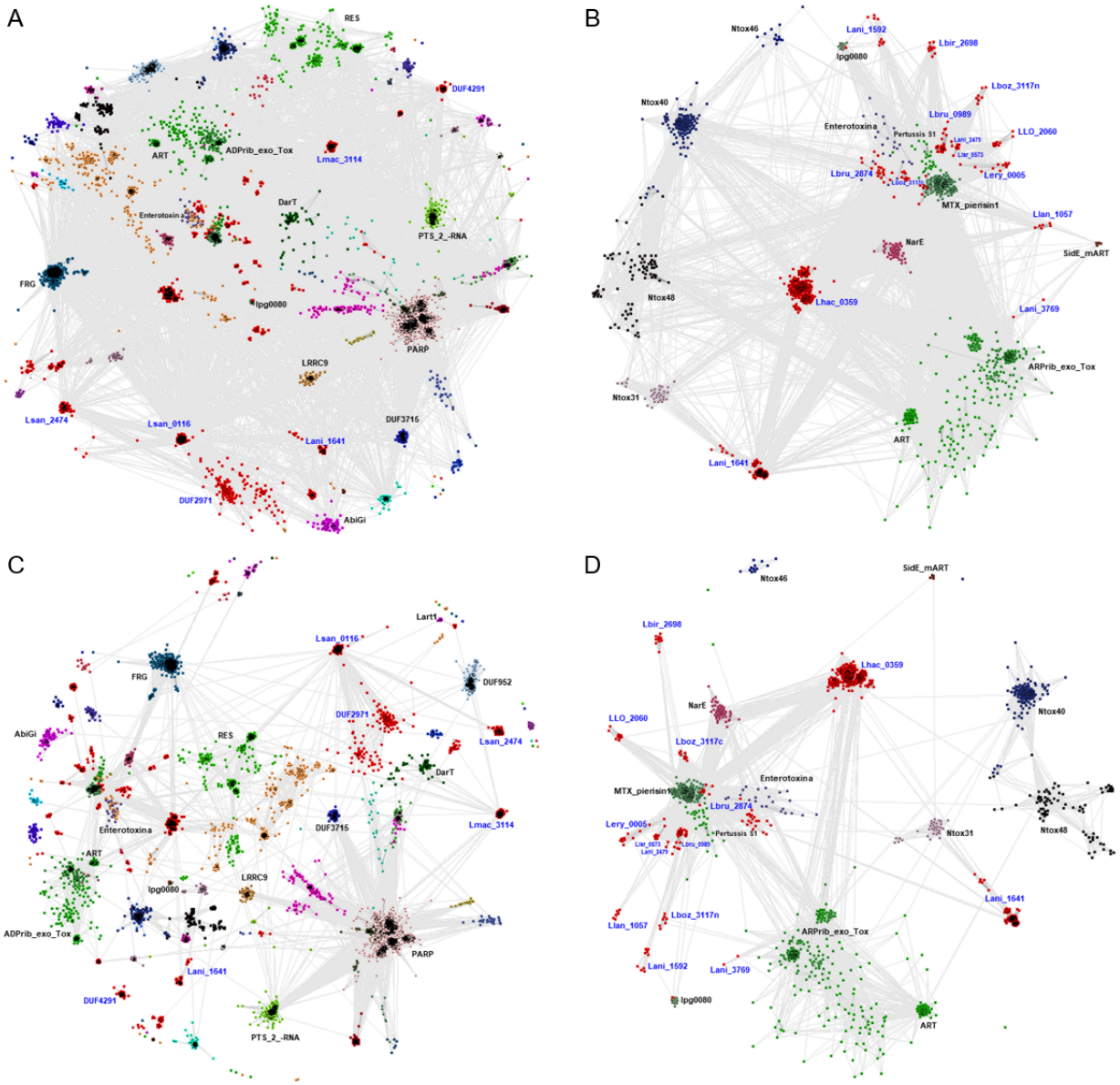


Figure 5. Sequence similarity relationships between new and known ART families. A and C: all ARTs families; B and D: subset of families that are closely similar to *E. coli* heat-labile enterotoxin (LT). The CLANS graph shows the sequence similarities obtained using pairwise BLAST comparisons, taking into account significant and borderline significance similarities: up to the E-value of 1 (A and B) and up to the E-value of 0.01 (C and D). Novel families of ARTs are marked in red. Colouring by families. Labelling by Pfam identifiers of selected known ART families and the gene symbols representing novel families not yet described in the Pfam database.

In-depth sequence similarity analyses (Fig. 5) compare all novel ART families (including those not described in detail in this article) with known ART-like families. The separation of the novel families (indicated in red) is clearly visible. Some of the ARTs show similarity to *E. coli* heat-labile enterotoxin (LT) (Fig. 5B and Fig. 5D).

Out of the twenty-six novel *Legionella* ART-like families, we characterised in more detail the six most interesting ones, considering breadth of their taxonomic distribution, numbers of representatives, features of operons and the conservation of key motifs. The results from this study are compiled into an ART database available at http://bioinfo.sggw.edu.pl/astarte/. This resource combines the new ART families with already known ones. In total, we present data on 82 ART families. We provide sequences, HMMs, sequence logos, 3D structure models, and brief descriptions.

The DUF2971 family

The putative effector protein lpg1268 from *Legionella pneumophila* subsp. *pneumophila* and homologs from 13 other *Legionella* species are predicted as a novel ADP-ribosyltransferase family. The DUF2791 domain (DUF, domain of unknown function) shows statistically significant sequence similarity to known ARTs, especially AbiGi, DarT, and Tox-ART-HYD1 families (see Fig. 2, Fig. 4 and Fig. 5). DarT is the toxin element of a toxin-antitoxin (TA) system. It is an enzyme that specifically ADP-ribosylates thymidine nucleotides on ssDNA in a sequence-specific manner (Jankevicius et al., 2016; Schuller et al., 2021).

The ADP-ribosyltransferase catalytic core appears to be well conserved in this family, although the histidine residue in the catalytic motif I is poorly conserved. There is a strictly conserved tyrosine residue in motif II and two conserved glutamic acid residues in the ExE motif III.

The DUF2971 family is present in 5981 species (UniProt database), mainly in Bacteria (5913 species), Archaea (56 species), Eukaryota (6 hits, mostly in in fungi) and in *Caudoviricetes* - tailed bacteriophages (12). Among the DUF2971-possessing bacteria, one numerous group consists of soil-living saprophytic *Clostidriales* that ferment plant polysaccharides (82 species). Another large group are ubiquitous bacteria of the order *Enterobacteriales*, abundant in the human large intestine, also on the skin and in the oropharynx, or free-living in water. Most are opportunistic pathogens that infect the organisms with the weakened immune system, like *Salmonella*, *Escherichia coli*, *Klebsiella* or *Shigella*. Notably, DUF2971 homologues can be found in strains pathogenic to humans in *Escherichia coli* O157: H7 (EHEC), *Salmonella enterica*, *Vibrio cholerae*.

The appearance of the active site and the structure of the protein (the core is built of 6 beta sheets with inserts between) leaves no doubt this family should be included in the HYE clade (see Fig. 2, Suppl. Fig. 1 and Suppl. Fig. 2). The structural divergence from the HYE clade seen in the CLANS analysis (see Fig. 4) is due to unusually large helical inserts between β-strands 1 and 2, as well as between β-strands 4 and 5, that do not disrupt the core of the structure (see Suppl. Fig. 1).

About 18% of proteins having a DUF2971 domain may function as bacterial effectors (Fig. 8 and Suppl. Table S7). Predicted DUF2971 effectors are found in well-known pathogenic strains of diverse bacteria: *S. enterica*, *V. parahaemolyticus*, *V. cholerae*, *L. pneumophila* (Suppl. Table S7), suggesting that DUF2971 may play a role in their pathogenic mechanisms.

The DUF2971 domain proteins are, in a few instances, found in operons together with peptidase C26 and low affinity iron permeases. We propose that these proteins may interact in response to oxidative stress, regulation of iron metabolism and protein degradation. Under stress, such as iron excess, peptidase C26 can degrade poly-gamma-glutamate, which binds iron ions, reducing their toxicity. The degradation of poly-gamma-glutamates provides glutamic acid, a precursor to

important metabolites, including glutathione, which protects against oxidative stress. Peptidase C26 can influence the availability of metabolites and iron ions by modulating the activity of ADP-ribosyltransferases in response to stress or cell damage. Iron permease regulates the transport of iron ions into the cell, and degradation of poly-gamma-glutamate by peptidase C26 affects intracellular iron levels, which is crucial for homeostasis and protection against oxidative stress. The operon-coordinated response to oxidative stress and regulation of iron metabolism promotes bacterial adaptation to changing environmental conditions.

In another interesting operon variant, we found DUF2971 together with Abi_C, KfrA_N, and a dermonecrotic toxin of the papain-like fold (see Fig. 7). The grouping of these genes in a single operon suggests a coordinated response to various forms of stress, such as bacteriophage infections, competitive pressure, and the regulation of plasmid replication. ADP-ribosyltransferase (ART) may be involved in stress response and DNA repair. The Abi_C domain induces abortive infection, which leads to the death of the infected cell and prevention of bacteriophage spread. KfrA_N suggests involvement in the control of plasmid replication, which is crucial for the propagation of resistance genes. Papain-like fold toxins can destroy competing bacteria or protect against pathogens, which is vital in competitive environments. The integration of DNA repair mechanisms, bacteriophage defence, plasmid control, and toxin production indicate a complex system of defence and adaptation, enhancing the bacteria's chances of survival in changing conditions.

Lsan_0116 – an effector family common to bacteria and archaea

The Lsan_0116 novel ART-like family, named after the cognate *Legionella santicrucis* protein, shows a noticeable similarity (13% sequences identity by HHpred) to the previously described DUF2971 (see Fig. 4, CLANS diagram). However, the two families present different types of active sites: Lsan_0116 presents an RSE motif, while DUF2971 presents a modified HYE motif (see Fig. 2). Therefore, we consider them to be separate families, even though distant homologs of the two families overlap. We discovered it mainly in bacteria with it being present in most major bacterial taxa (see Fig. 3), and also sparsely in archaeons.

The active site logo (Fig. 2) shows an arginine or lysine residue in motif I and two glutamic acid residues from motif III. Motif II, which is responsible for the groove for binding the nicotinamide ring and ribose and catalysis (Bell & Eisenberg, 1996), is the most unusual. In this family, motif II has the form of [SC]WH instead of the canonical STS. The protein structure presented by the AlphaFold model is also unusual (see Suppl. Fig. 1 and Suppl. Fig. 2) – the active site is formed by the first, third and seventh beta sheets (1-3-7 instead of the typical 1-2-5 layout). The structural divergence from the RSE clade is evident in the CLANS analysis (see Fig. 4) and is due to unusual sequence insertions within the catalytic core sequence.

Majority of the Lsan_0116 family (80%) are predicted effectors, in contrast to the related DUF2971 family which has only 18% likely effectors (Fig. 8 and Suppl. Table S7).

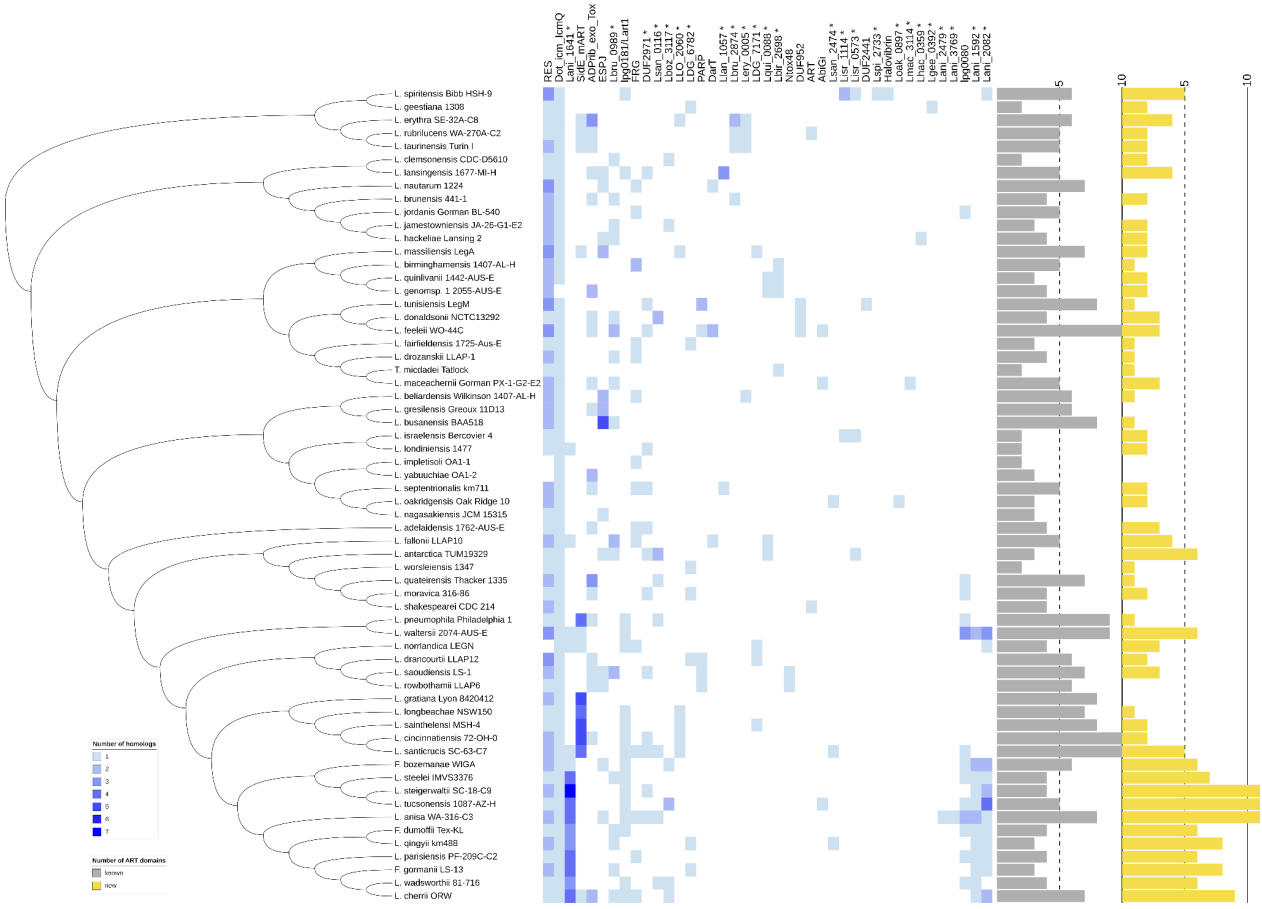


Figure 6. Distribution of ART families in *Legionella* strains. Numbers of homologues of novel and known ADP-ribosyltransferase families in selected members of the family *Legionellaceae*. Histograms on the right: numbers of members of known (grey) and novel (yellow) ART families identified in each species.

New families are marked with asterisks above the heatmap. The DUF4291 family was not included in the diagram because *Legionella* homologs were not found in the RefSeq database (see Methods).

The DUF4291 family – the largest atypical novel ART family, present also outside *Bacteria*

In addition to sequence searches for novel ARTs, we performed a structural search among RoseTTAFold models of all families in the Pfam database (see Methods). A single domain of unknown function (DUF4291), present in *Legionella*, stood out as significantly similar by structure to known ARTs. Based on structural similarity, conserved sequence motifs in the beta sheets 1, 2 and 5, an ART-like active site could be proposed (Fig. 2).

According to our survey, the DUF4291 is a novel effector family with similarity to ARTs. According to the Pfam database annotations, there are 2 conserved motifs in this uncharacterized family. Structural alignment suggests that these motifs are atypical counterparts of the motifs I and II of the ART catalytic site (QAY and WVK correspond to HxT and STS, respectively).

Although the predicted active site in this family has a very unusual appearance, the catalytic motif III includes a "classical" glutamine. The new family is strikingly well conserved. The DUF4291 domain shows statistically significant structure similarity to HYE clade of ARTs, especially Pfam families Exotox-A_cataly and RES. We could not identify significant sequence similarity to any known family of ARTs. (Suppl. Fig. 1 and Suppl. Fig. 2). Also, the correct docking of NAD to the obtained model suggests that the DUF4291 domain may have ART activity (Suppl. Fig. 2).

The DUF4291 family is present in 4540 species, mainly in *Bacteria* (3887 organisms) and in *Eukaryota* (646 species). This family is observed in *Legionellaceae* (it has only been identified in the *Tatlockia* genus which is closely related to the genus *Legionella* (Tindall, 2020; Saini & Gupta, 2021)) and in other, well-studied bacteria e.g., *Escherichia coli* O103:H25, *Salmonella enterica*, *Xanthomonas oryzae*, *Myxococcus xanthus*, *Klebsiella aerogenes*. In bacteria, most hits were identified in bacterial phytopathogens and animal pathogens, including human ones or in opportunistic pathogens, and in harmless commensals of mammals and bacteria that occur in water or soil. We also found it common in bacteria widely used to produce antibiotics and other therapeutics (e.g. *Streptomycetaceae*), in photosynthesising *Cyanobacteria* and in the human

microbiome (e.g. *Firmicutes*). DUF4291 was also found in abundance in organic compound-degrading bacteria of the order *Myxococcales* and in plant-symbiotic nitrogen-fixing rhizobia.

In *Eukaryotes,* this family is present mainly in *Fungi* (605 hits), particularly *Dikarya*. It is widespread in *Aspergillaceae.* It also occurs in saprophytic fungi of the genus *Fusarium*. In *Metazoa,* the DUF4291 family was found in some taxa *Animalia,* among others in segmented annelid worms, urchins, lancelets and molluscs, but is absent from arthropods, tunicates and vertebrates. In addition, DUF4291 is found in *Alveolata*, *Amoebozoa*, *Rhodymeniophycidae* and *Viridiplantae*. Proteins from the new family are occasionally found in *Archaea* and *Viruses*.

Genomic neighbourhood analysis of DUF4291 family proteins (Fig. 7) revealed the common presence of TetR_N and NUDIX domains (in 28% and 18% of analysed neighbourhoods, respectively). TetR_N is a domain found in several bacterial and archaeal transcriptional regulators. The NUDIX hydrolase domains are widespread and usually function as pyrophosphohydrolases. Some NUDIX proteins degrade potentially mutagenic, oxidised nucleotides while others control the levels of metabolic intermediates and signalling compounds. Slightly less frequent are Macro and ADP_ribosyl_GH domains, co-occurring in 9-10% of DUF4291 neighbourhoods. Macro and ADP-ribosylglycohydrolase (ADP_ribosyl_GH) domains are elements of ADP-ribosylation signalling and complete the “writer-reader-eraser” triad. Macros can act as "readers" or "erasers", whereas ADP-ribosylhydrolyses are “erasers” of ADP-ribosylation marks. Another known ART domain – RNA 2'-phosphotransferase (PTS_2-RNA) was observed almost as often (10%) in the genomic neighbourhoods of DUF4291. Taken together, these observations strongly support involvement of DUF4971 in ADP-ribosylation signalling and transcription and/or RNA biochemistry.

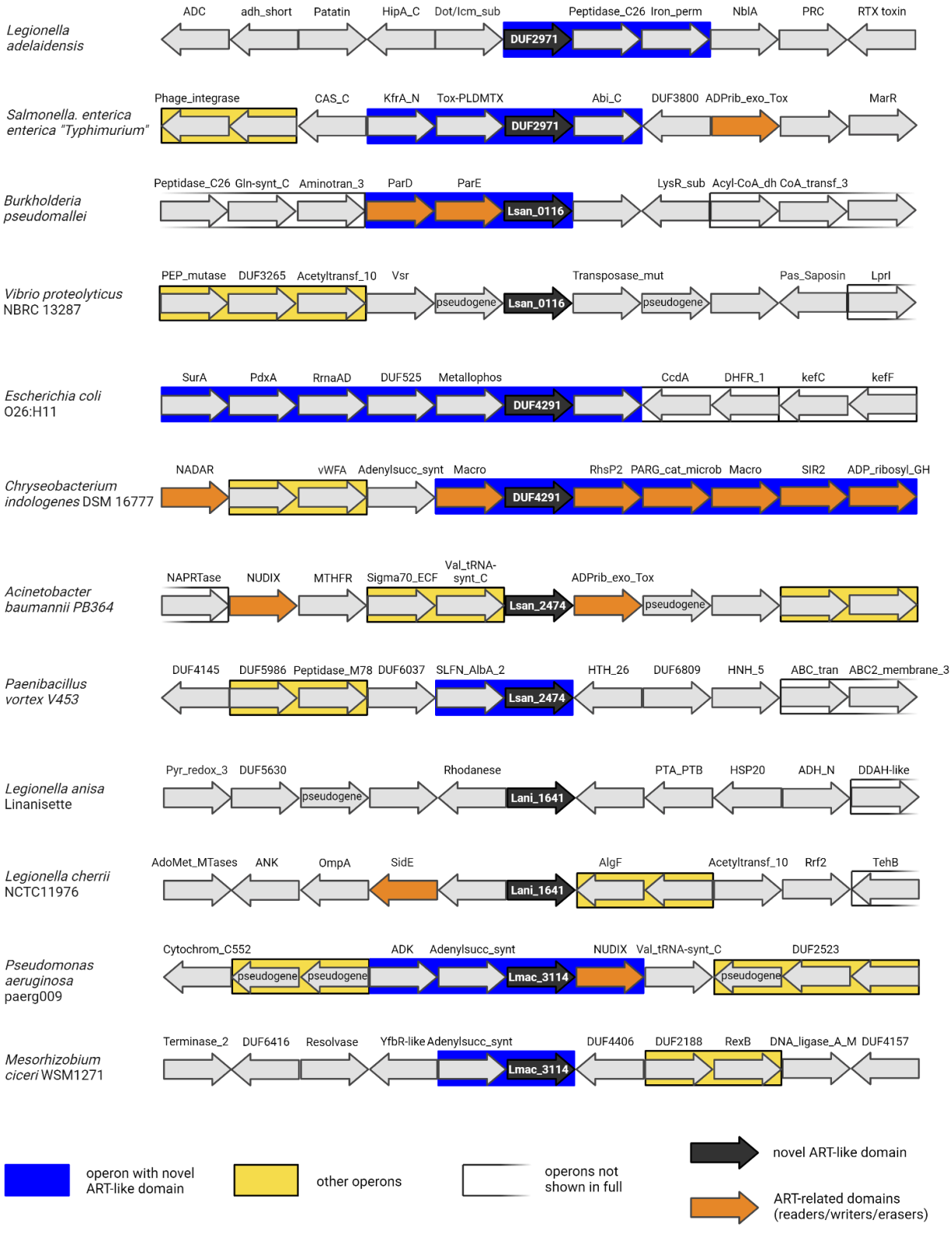


Figure 7. Genomic neighbourhoods of selected novel ART-like domains. Pfam/InterPro protein domain short names or gene names used. The known or predicted operon information was obtained from the BioCyc database. Created with BioRender.com

It is estimated that more than 40% of proteins having a DUF4291 domain may be effectors (see Fig. 8 and Suppl. Table S7). However, due to the prediction algorithm using the presence of eukaryotic-like domains (ELDs), there is a risk of overprediction.

Lsan_2474, a new ART-like domain, similar to AbiGi

This novel family, named after the *Legionella santicrucis* Lsan_2474 protein, is a distant homologue R-S-E clade families, AbiGi (8% sequence identity) and DUF2971. FFAS sequence alignments identified three conserved active site motifs typical of the R-S-E clade (R-SxS-ExE), however, motif II is often mutated to CxS. The presence of the novel family was confirmed only in Bacteria (133 organisms) from diverse taxa: *Gammaproteobacteria, Betaproteobacteria, Epsilonproteobacteria*, *Terrabacteria* group, *Acidobacteria*, and *Bacteroidetes*, e.g., *Vibrio cholerae, Pseudomonas syringae, Yersinia enterocolitica* and a virus (Siphoviridae). There are bacterial effectors in the Lsan_2474 family and we estimate that about 78% of proteins having a Lsan_2474 domain may be effectors (Fig. 8 and suppl. S7).

Lani_1641, a new ART-like domain, similar to Ntox31 and Enterotoxin_a

Lani_1641 is a small, novel ART-like family named after *L. anisa* protein Lani_1641. Although the active site presents the RSE motif, sequence comparisons (Fig. 4 and Fig. 5) do not place this family close to the RSE clade families. The matching of the protein structure model to known ART structures shows strong similarity to Pertussis toxin, Scabin and PARPs with 11-18% sequence identity and significant DALI Z-scores between 5.5 and 11.1. The CLANS analysis also places this family in the RSE clade, and the protein model suggests the presence of a catalytic core composed of seven beta sheets.

Members of this family were detected only in *Legionellaceae*. It is estimated that about 65% of proteins having a Lani_1641 domain may be effectors (Fig. 8 and Suppl. Table S7). The

ART-like Lani_1641 domains are not accompanied by other functional and structural domains in proteins. Also, no multigene operons involving the Lani-1641 domain were identified (Fig. 7).

Lmac_3114 new ART-like domain, similar to DarT

The novel Lmac_3114 family, named after *Legionella maceachernii* protein Lmac_3114 is a distant homologue of DarT domain – from the H-Y-E clade of ART. In Lmac_3114, the catalytic motif II [Y-x-x] contains a conserved tryptophan residue. DarT family members act as the toxins in toxin-antitoxin (TA) system, by specifically modifying thymidine nucleotides on ssDNA.

Its closest relatives identified by FFAS, HHpred and Phyre2 also include the ART family RNA 2`-phosphotransferase. Sequence alignments identified three conserved active site motifs (Hxx, FFW, and xxE). Structural comparisons do not make it possible to determine precisely to which clade this new family should be assigned. Greatest similarity is found to the DarT family from the HYE clade and PTS_2-RNA from the HHh clade, however, the presence of a conserved glutamic acid residue in the third motif leads us to assign it to the HYE clade. For both clades, the presence of six beta-sheets in the catalytic core is characteristic, which is reflected in the Lmac_3114 model.

This family is detected only in bacteria from *Gammaproteobacteria*, *Alphaproteobacteria*, *Betaproteobacteria*, *Deltaproteobacteria*, *Nitrospirae*, *Terrabacteria* group, *Acidobacteria*, PVC group and *Bacteroidetes*, e.g., *Pseudomonas aeruginosa*, *P. syringae, Xanthomonas oryzae* and *Burkholderia pseudomallei.*

Genomic neighbourhood analysis of Lmac_3114 family proteins provided evidence that the adenylosuccinate synthetase domain is extremely common in its vicinity (co-occurrence in 60% of cases) and, in at least some cases, forms an operon with the ART-like domain (Fig. 7). The presence of genes encoding ADP-ribosyltransferase and adenylosuccinate synthase in the operon suggests their functional association, e.g. co-operation in specific metabolic or regulatory processes. We have identified such an operon in nitrogen-fixing soil bacteria (*Rhizobium azooxidifex* and *Mesorhizobium ciceri*). A more elaborate version of the operon is found in *Pseudomonas aeruginosa* – there the ART-like domain is found together with the AAA domain, adenylosuccinate synthetase and NUDIX domain, a hydrolase that cleaves nucleoside

diphosphates linked to another moiety. The presence in the common operon of adenylosuccinate synthase and ART domains suggests regulation of ART activity by changes in purine synthesis, e.g. indirect modulation of ADP-ribosyltransferase activity in either cellular regulation or stress response by regulating NAD+ availability.

Analysis of potential NAD binding sites

To evaluate the compatibility of the novel ART families with NAD binding, we used HADDOCK to dock NAD to reference protein structure models for families discussed in detail in this article. For comparison, we performed re-docking for a crystallographic complex consisting of NAD bound to eukaryotic mono-ADP-ribosyltransferase ART2.2 (PDB code 1OG3) (Ritter et al., 2003). Among the six novel ART families, the best NAD docking scores ranged from -60.2 to -31.5 while the score for redocking to the NAD bound crystal structure, the score was -49.6. The best scores (under -60.0) were obtained for Lsan2474 and DUF4291 representatives. In all cases, manual identification of amino acids forming the putative NAD pocket based on sequence conservation and structural alignments to known ARTs resulted in a more favourable docking than automatic detection of ligand-binding residues. In almost all cases (except Lmac_3114), the docking procedure resulted in the NAD ligand in a correct position, i.e. resembling the NAD poses observed in known ART crystal structures (Suppl. Fig. 1). Intriguingly, for Lmac_3114, the NAD molecule is located in a non-standard location, however, the predicted active site residues interact with the bound NAD molecule which may suggest an atypical catalytic mechanism, different from most ART families. Mapping sequence conservation onto the protein surfaces shows that highly conserved residues group within and near in the predicted NAD binding pockets. Close-up view of the NAD sites focusing on conserved residue side chains shows that the poses of NAD molecule are very similar for all models, with adenine ring stacking out (again except of Lmac_3114).

The active site signatures indicated by sequence logos (Fig. 2) were confirmed by structural models and NAD docking results as likely catalytically relevant residues involved in binding the NAD molecule (Suppl. Fig. 1, Suppl. Fig. 2 and Suppl. Table S8). We obtained similar results using the AlphaFold3 server (Abramson et al., 2024) to model representative proteins from each

family with NAD+ (data not shown). These observations support our hypothesis that the newly discovered ART-like families can bind NAD and be functional ADP-ribosyltransferases.

To gain insight into potential functions of the novel ART-like families, we predicted the content of secreted proteins and effectors in each family by analysing presence of signal peptides, the amino acid composition of the C-terminus of the sequence, the binding site of the secretion chaperone protein and the presence of eukaryotic-like domains (see Methods and Suppl. Table S7). Among the families discussed in detail in this article, all are predicted to include secreted effector proteins. Particularly prominent is the Lsan_0116 family, in which we predict that more than 80% of the proteins will be effectors (Fig. 8).

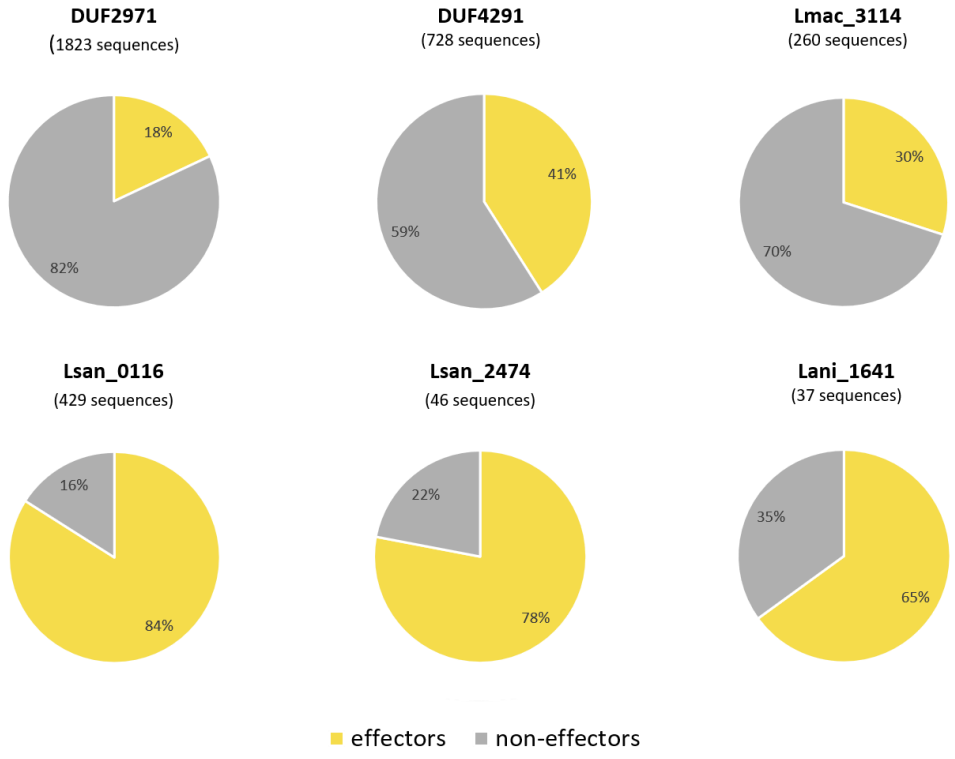


Figure 8. Predicted share of effector proteins in the novel ART families. Predictions made using SignalP 6.0, EffectiveDB and BastionX. Data for all identified families available in the supplement (Suppl. Table S7).

Discussion

In our bioinformatic examination of 41 *Legionella* species, we have documented 63 groups of *Legionella* orthologues with convincing sequence or structure similarity to ADP-ribosyltransferases, that can be organised into 39 ART-like families. Among these families, 26 families are novel ART superfamily members (see Table 1). One of the novel families (DUF2971) is represented in *L. pneumophila.* The newly identified ART families, initially found by sequence searches, have been further corroborated by artificial intelligence-driven structural predictions.

Table 1. Novel ADP-ribosyltransferase families identified in the *Legionella* pan-proteome

Families are identified by Pfam names (DUF) or by symbols of representative genes (Burstein et al., 2016).

family name	clade	1st motif RxD/HxT	2nd motif STS/Yxx	3rd motif Ex[EQD]	likely substrate	occurrence
DUF2971	HYE	Hyx	Yxx	ExE	protein or bacteriophage DNA	Bacteria, Archaea, Viruses
DUF4291	(?) HYE	QA[YF]	WIK	[AS]IQ	protein	Bacteria, Eukaryota, Archaea, Viruses
Lani_1592	RSE	Rx[DH]	[GS]x[TSC]	ExE	protein	*Legionellaceae* only
Lani_1641	RSE	RG[TMS]	Sx[ST]	NNE	protein or auto-ADP-ribosylation	*Legionellaceae* only
Lani_2082	RSE	RxM	STS	ExE	protein or auto-ADP-ribosylation	*Legionellaceae* only
Lani_2479	RSE	Rxx	STS	[ED]x[NE]	protein	*Legionellaceae* only
Lani_3769	RSE	R??	STS	ExE	protein	*Legionellaceae* only
Lbir_2698	RSE	Rxx	[CSD]FS	xx[EQ]	protein	Bacteria
Lboz_3117	RSE	RGD	STS	ExE	protein	Bacteria
Lbru_0989	RSE	R[SA]D	C[ML][TS]	?	protein	Bacteria
Lbru_2874	RSE	RGD	ST[ST]	ExE	protein	*Legionellaceae* only
LDG_6782	RSE	?	STS	ExE	DNA	Bacteria, Eukaryota
LDG_7171	atypical	HxT	Cx[GC]	xxD	DNA	Bacteria
Lery_0005	RSE	RGD	STS	ExE	protein	Bacteria
Lgee_0392	HYE	HGT	Yxx	[IV]HE	protein	Bacteria
Lhac_0359	RSE	RGx	SW[TS]	ExE	protein or auto-ADP-ribosylation	*Legionellaceae* only
Lisr_0573	RSE	RxD	Cx[ST]	xxE	protein	Bacteria
Lisr_1114	RSE	RxD	[GS]T[TS]	ExE	protein	*Legionellaceae* only
Llan_1057	RSE	Rx[IVD]	ST[ST]	ExE	protein or DNA	*Legionellaceae* only
LLO_2060	RSE	RGD	STS	ExE	protein	Bacteria
Lmac_3114	HYE	Hxx	FFW	xxE	DNA	Bacteria
Loak_0897	HYE	HxT	Yxx	ExD	protein	*Legionellaceae* only
Lqui_0088	HHh	HxT	HxT	xSY or xxS	RNA	*Legionellaceae* only
Lsan_0116	RSE	[RK]Y[ML]	[CS]WH	ExE	protein or bacteriophage DNA	Bacteria, Archaea, Viruses
Lsan_2474	RSE	Kxx	[CS]Mx	QxE	RNA or DNA	Bacteria
Lspi_2733	HYE	HxS	Yxx	xxE	protein or DNA	*Legionellaceae* only

The DUF2971 family shows similarity to DarT proteins and scabin, a DNA-acting ADP-ribosyltransferases (Lyons et al., 2016) suggests that DUF2971 effectors could modify host or invader's DNA.

The largest newly identified ART-like family, DUF4291, including approximately 5 000 members, shows structural and sequence similarity to the diphtheria toxin family of mono-ART toxins, e.g. *P. aeruginosa* exotoxin A, *C. diphtheriae* diphtheria toxin and cholix toxin from *V. cholerae* (Jørgensen et al., 2008). These ARTs are important virulence factors belonging to the class of exotoxins secreted by pathogenic bacteria and cause diphtheria, cholera and pneumonia, respectively. All of them ADP-ribosylate elongation factor 2 in host cells, so it is plausible to hypothesise that the DUF4291 family modifies proteins rather than nucleic acids.

The ART complement of the *Legionella* genus showcases diverse effector ARTs, possibly corresponding to adaptability to various hosts. For example, the extensively studied *Legionella pneumophila*, possesses genes known to facilitate infection in various hosts, which exhibit both similarities and distant relationships with eukaryotic ARTs, possibly acquired through horizontal gene transfer.

Although the predicted effector functions remain to be validated, these novel ART-like families present attractive candidates for experimental studies into mechanisms of pathogenicity due to their likely crucial roles in the pathogen's survival. While this manuscript was being finalised, a thorough study presented a detailed survey of structural domains in all the effectors of *L. pneumophila* (Patel et al., 2024). While our study addresses only ART-like proteins, it is complementary to work of Patel et al. because we surveyed 41 species instead of one and did not limit the exploration to effectors.

In summary, our survey of the *Legionella* ADP-ribosyltransferase world provides insights into the pathogen's toolkit for infection and offers a broader perspective on nature's remarkable adaptability and ingenuity in the use of the evolutionarily successful ART-like superfamily. The results from this study are available at http://bioinfo.sggw.edu.pl/astarte/.

Acknowledgements

We thank Drs A. Muszewska, V. Tagliabracci and B. Mayro for critical reading of the manuscript and helpful comments.

Author contributions

MK: conceptualised the project, performed the experiments, analysed the data, prepared figures and tables, authored or reviewed drafts of the paper, and approved the final draft

BB: performed the experiments, analysed the data, and approved the final draft.

MG: performed the experiments, analysed the data, prepared supplemental tables, and approved the final draft.

KP: conceptualised the project, co-supervised the project, authored or reviewed drafts of the paper, and approved the final draft.

MD: conceptualised the project, co-supervised the project, performed the experiments, analysed the data, prepared supplemental figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Funding and additional information

K.P. was supported by the Polish National Science Centre grant 2019/33/B/NZ2/01409. M. G. was supported by the Polish National Science Centre grant 2019/35/N/NZ2/02844. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung C-C, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Žemgulytė A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Žídek A,

Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*:1–3. DOI: 10.1038/s41586-024-07487-w.

Akturk A, Wasilko DJ, Wu X, Liu Y, Zhang Y, Qiu J, Luo Z-Q, Reiter KH, Brzovic PS, Klevit RE, Mao Y. 2018. Mechanism of phosphoribosyl-ubiquitination mediated by a single Legionella effector. *Nature* 557:729–733. DOI: 10.1038/s41586-018-0147-6.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402.

Aravind L, de Souza RF. 2012. Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions. *Molecular BioSystems* 8:1661–1677. DOI: 10.1039/c2mb05487f.

Aravind L, Zhang D, de Souza RF, Anand S, Iyer LM. 2015. The Natural History of ADP-Ribosyltransferases and the ADP-Ribosylation System. *Current topics in microbiology and immunology* 384:3–32. DOI: 10.1007/82_2014_414.

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes H-W, Horn M, Rattei T. 2009. Sequence-Based Prediction of Type III Secreted Proteins. *PLoS Pathogens* 5:e1000376. DOI: 10.1371/journal.ppat.1000376.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research* 38:W529-533. DOI: 10.1093/nar/gkq399.

Baek M, Anishchenko I, Humphreys IR, Cong Q, Baker D, DiMaio F. 2023. Efficient and

accurate prediction of protein structure using RoseTTAFold2. :2023.05.24.542179. DOI: 10.1101/2023.05.24.542179.

Baranowski B, Pawłowski K. 2023. Protein family neighborhood analyzer-ProFaNA. *PeerJ* 11:e15715. DOI: 10.7717/peerj.15715.

Bell CE, Eisenberg D. 1996. Crystal structure of diphtheria toxin bound to nicotinamide adenine dinucleotide. *Biochemistry* 35:1137–1149. DOI: 10.1021/bi9520848.

Black MH, Osinski A, Park GJ, Gradowski M, Servage KA, Pawłowski K, Tagliabracci VS. 2021. A Legionella effector ADP-ribosyltransferase inactivates glutamate dehydrogenase. *The Journal of Biological Chemistry* 296:100301. DOI: 10.1016/j.jbc.2021.100301.

Bock FJ, Chang P. 2016. New directions in poly(ADP-ribose) polymerase biology. *The FEBS Journal* 283:4017–4031. DOI: 10.1111/febs.13737.

Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. 2012. Domain enhanced lookup time accelerated BLAST. *Biology Direct* 7:12. DOI: 10.1186/1745-6150-7-12.

Brenner DJ, Steigerwalt AG, McDade JE. 1979. Classification of the Legionnaires' disease bacterium: Legionella pneumophila, genus novum, species nova, of the family Legionellaceae, familia nova. *Annals of Internal Medicine* 90:656–658. DOI: 10.7326/0003-4819-90-4-656.

Burroughs AM, Aravind L. 2020. Identification of Uncharacterized Components of Prokaryotic Immune Systems and Their Diverse Eukaryotic Reformulations. *Journal of Bacteriology* 202. DOI: 10.1128/JB.00365-20.

Burstein D, Amaro F, Zusman T, Lifshitz Z, Cohen O, Gilbert JA, Pupko T, Shuman HA, Segal

G. 2016. Uncovering the Legionella genus effector repertoire - strength in diversity and numbers. *Nature genetics* 48:167–175. DOI: 10.1038/ng.3481.

Chambers ST, Slow S, Scott-Thomas A, Murdoch DR. 2021. Legionellosis Caused by Non-Legionella pneumophila Species, with a Focus on Legionella longbeachae. *Microorganisms* 9:291. DOI: 10.3390/microorganisms9020291.

Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H, Grishin NV. 2014. ECOD: An Evolutionary Classification of Protein Domains. *PLOS Computational Biology* 10:e1003926. DOI: 10.1371/journal.pcbi.1003926.

Cohen MS, Chang P. 2018. Insights into the biogenesis, function, and regulation of ADP-ribosylation. *Nature Chemical Biology* 14:236–243. DOI: 10.1038/nchembio.2568.

Community NPB. 2021.SignalP 6.0 predicts all five types of signal peptides using protein language models. *Available at http://bioengineeringcommunity.nature.com/posts/signalp-6-0-predicts-all-five-types-of-signal-peptides-using-protein-language-models* (accessed March 8, 2022).

Corda D. 2003. NEW EMBO MEMBER'S REVIEW: Functional aspects of protein mono-ADP-ribosylation. *The EMBO Journal* 22:1953–1958. DOI: 10.1093/emboj/cdg209.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: A Sequence Logo Generator. *Genome Research* 14:1188–1190. DOI: 10.1101/gr.849004.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M, Claverie J-M, Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* 36:W465–W469. DOI:

10.1093/nar/gkn180.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Computational Biology* 7:e1002195. DOI: 10.1371/journal.pcbi.1002195.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792–1797. DOI: 10.1093/nar/gkh340.

Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. 2016. EffectiveDB—updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Research* 44:D669–D674. DOI: 10.1093/nar/gkv1269.

Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE, Shepard CC, Brachman PS. 1977. Legionnaires' disease: description of an epidemic of pneumonia. *The New England Journal of Medicine* 297:1189–1197. DOI: 10.1056/NEJM197712012972201.

Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics (Oxford, England)* 20:3702–3704. DOI: 10.1093/bioinformatics/bth444.

Fu J, Zhou M, Gritsenko MA, Nakayasu ES, Song L, Luo Z-Q. 2022. Legionella pneumophila modulates host energy metabolism by ADP-ribosylation of ADP/ATP translocases. *eLife* 11:e73611. DOI: 10.7554/eLife.73611.

Gabler F, Nam S-Z, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V. 2020. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics* 72:e108. DOI: 10.1002/cpbi.108.

Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, Bruns G, Yarza P, Peplies J, Westram R, Ludwig W. 2017. 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology* 261:169–176. DOI: 10.1016/j.jbiotec.2017.06.1198.

Gomez-Valero L, Rusniok C, Carson D, Mondino S, Pérez-Cobas AE, Rolando M, Pasricha S, Reuter S, Demirtas J, Crumbach J, Descorps-Declere S, Hartland EL, Jarraud S, Dougan G, Schroeder GN, Frankel G, Buchrieser C. 2019. More than 18,000 effectors in the *Legionella* genus genome provide multiple, independent combinations for replication in human cells. *Proceedings of the National Academy of Sciences* 116:2265–2273. DOI: 10.1073/pnas.1808016116.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* 59:307–321. DOI: 10.1093/sysbio/syq010.

Holm L. 2020. DALI and the persistence of protein shape. *Protein Science* 29:128–140. DOI: 10.1002/pro.3749.

Honorato RV, Koukos PI, Jiménez-García B, Tsaregorodtsev A, Verlato M, Giachetti A, Rosato A, Bonvin AMJJ. 2021. Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Frontiers in Molecular Biosciences* 8.

Hottiger MO, Hassa PO, Lüscher B, Schüler H, Koch-Nolte F. 2010. Toward a unified nomenclature for mammalian ADP-ribosyltransferases. *Trends in Biochemical Sciences* 35:208–219. DOI: 10.1016/j.tibs.2009.12.003.

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and

comparing biological sequences. *Bioinformatics* 26:680–682. DOI: 10.1093/bioinformatics/btq003.

Isberg RR, O'Connor T, Heidtman M. 2009. The Legionella pneumophila replication vacuole: making a cozy niche inside host cells. *Nature reviews. Microbiology* 7:13–24. DOI: 10.1038/nrmicro1967.

Jankevicius G, Ariza A, Ahel M, Ahel I. 2016. The Toxin-Antitoxin System DarTG Catalyzes Reversible ADP-Ribosylation of DNA. *Molecular Cell* 64:1109–1116. DOI: 10.1016/j.molcel.2016.11.014.

Jørgensen R, Purdy AE, Fieldhouse RJ, Kimber MS, Bartlett DH, Merrill AR. 2008. Cholix Toxin, a Novel ADP-ribosylating Factor from Vibrio cholerae*. *Journal of Biological Chemistry* 283:10671–10678. DOI: 10.1074/jbc.M710008200.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589. DOI: 10.1038/s41586-021-03819-2.

Kanatani J, Watahiki M, Kimata K, Kato T, Uchida K, Kura F, Amemura-Maekawa J, Isobe J. 2021. Detection of Legionella species, the influence of precipitation on the amount of Legionella DNA, and bacterial microbiome in aerosols from outdoor sites near asphalt roads in Toyama Prefecture, Japan. *BMC Microbiology* 21:215. DOI:

10.1186/s12866-021-02275-2.

Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, Ong WK, Paley SM, Subhraveti P. 2017. The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics* 20:1085–1093. DOI: 10.1093/bib/bbx085.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30:3059–3066. DOI: 10.1093/nar/gkf436.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* 20:1160–1166. DOI: 10.1093/bib/bbx108.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10:845–858. DOI: 10.1038/nprot.2015.053.

Klockgether J, Tümmler B. 2017. Recent advances in understanding Pseudomonas aeruginosa as a pathogen. *F1000Research* 6:1261. DOI: 10.12688/f1000research.10506.1.

Koch-Nolte F (ed.). 2015. *Endogenous ADP-Ribosylation*. Springer International Publishing. DOI: 10.1007/978-3-319-10771-4.

Komander D, Randow F. 2017. Strange New World: Bacteria Catalyze Ubiquitylation via ADP Ribosylation. *Cell Host & Microbe* 21:127–129. DOI: 10.1016/j.chom.2017.01.014.

Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, Gascuel O. 2019. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic*

Acids Research 47:W260–W265. DOI: 10.1093/nar/gkz303.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* 49:W293–W296. DOI: 10.1093/nar/gkab301.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659. DOI: 10.1093/bioinformatics/btl158.

Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17:282–283. DOI: 10.1093/bioinformatics/17.3.282.

Li W, Jaroszewski L, Godzik A. 2002. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 18:77–82. DOI: 10.1093/bioinformatics/18.1.77.

Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. 2020. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Research* 48:W60–W64. DOI: 10.1093/nar/gkaa443.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379:1123–1130. DOI: 10.1126/science.ade2574.

Lyons B, Ravulapalli R, Lanoue J, Lugo MR, Dutta D, Carlin S, Merrill AR. 2016. Scabin, a Novel DNA-acting ADP-ribosyltransferase from Streptomyces scabies. *The Journal of*

Biological Chemistry 291:11198–11215. DOI: 10.1074/jbc.M115.707653.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* 39:D225–D229. DOI: 10.1093/nar/gkq1189.

Maresso AW, Deng Q, Pereckas MS, Wakim BT, Barbieri JT. 2007. Pseudomonas aeruginosa ExoS ADP-ribosyltransferase inhibits ERM phosphorylation. *Cellular Microbiology* 9:97–105. DOI: 10.1111/j.1462-5822.2006.00770.x.

McDade JE, Shepard CC, Fraser DW, Tsai TR, Redus MA, Dowdle WR. 1977. Legionnaires' disease: isolation of a bacterium and demonstration of its role in other respiratory disease. *The New England Journal of Medicine* 297:1197–1203. DOI: 10.1056/NEJM197712012972202.

Mercante JW, Winchell JM. 2015. Current and Emerging Legionella Diagnostics for Laboratory and Outbreak Investigations. *Clinical Microbiology Reviews*. DOI: 10.1128/CMR.00029-14.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nature Methods* 19:679–682. DOI: 10.1038/s41592-022-01488-1.

Mirdita M, Steinegger M, Söding J. 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35:2856–2858. DOI:

10.1093/bioinformatics/bty1057.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research* 49:D412–D419. DOI: 10.1093/nar/gkaa913.

Mondino S, Schmidt S, Rolando M, Escoll P, Gomez-Valero L, Buchrieser C. 2020. Legionnaires' Disease: State of the Art Knowledge of Pathogenesis Mechanisms of *Legionella*. *Annual Review of Pathology: Mechanisms of Disease* 15:439–466. DOI: 10.1146/annurev-pathmechdis-012419-032742.

Morales JC, Li L, Fattah FJ, Dong Y, Bey EA, Patel M, Gao J, Boothman DA. 2014. Review of Poly (ADP-ribose) Polymerase (PARP) Mechanisms of Action and Rationale for Targeting in Cancer and Other Diseases. *Critical reviews in eukaryotic gene expression* 24:15–28.

Munnur D, Ahel I. 2017. Reversible mono-ADP-ribosylation of DNA breaks. *The FEBS Journal* 284:4002–4016. DOI: 10.1111/febs.14297.

Newton HJ, Ang DKY, van Driel IR, Hartland EL. 2010. Molecular Pathogenesis of Infections Caused by Legionella pneumophila. *Clinical Microbiology Reviews* 23:274–298. DOI: 10.1128/CMR.00052-09.

Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. *Available at https://www.ncbi.nlm.nih.gov/nucleotide/* (accessed March 8, 2022).

O'Neal CJ, Amaya EI, Jobling MG, Holmes RK, Hol WGJ. 2004. Crystal structures of an intrinsically active cholera toxin mutant yield insight into the toxin activation mechanism.

Biochemistry 43:3772–3782. DOI: 10.1021/bi0360152.

Patel DT, Stogios PJ, Jaroszewski L, Urbanus M, Sedova M, Semper C, Le C, Takkouche A, Ichii K, Innabi J, Patel DH, Ensminger A, Godzik A, Savchenko A. 2024. Global atlas of predicted functional domains in Legionella pneumophila Dot/Icm translocated effectors. :2024.05.09.593423. DOI: 10.1101/2024.05.09.593423.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25:1605–1612. DOI: 10.1002/jcc.20084.

Ritter H, Koch-Nolte F, Marquez VE, Schulz GE. 2003. Substrate binding and catalysis of ecto-ADP-ribosyltransferase 2.2 from rat. *Biochemistry* 42:10155–10162. DOI: 10.1021/bi034625w.

Saini N, Gupta RS. 2021. A robust phylogenetic framework for members of the order Legionellales and its main genera (Legionella, Aquicella, Coxiella and Rickettsiella) based on phylogenomic analyses and identification of molecular markers demarcating different clades. *Antonie van Leeuwenhoek* 114:957–982. DOI: 10.1007/s10482-021-01569-9.

Salvatore M, Shu N, Elofsson A. 2018. The SubCons webserver: A user friendly web interface for state‐of‐the‐art subcellular localization prediction. *Protein Science : A Publication of the Protein Society* 27:195–201. DOI: 10.1002/pro.3297.

Samrakandi MM, Cirillo SLG, Ridenour DA, Bermudez LE, Cirillo JD. 2002. Genetic and Phenotypic Differences between Legionella pneumophila Strains. *Journal of Clinical Microbiology* 40:1352–1362. DOI: 10.1128/JCM.40.4.1352-1362.2002.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. 2021. Database resources of the national center for biotechnology information. *Nucleic Acids Research* 50:D20–D26. DOI: 10.1093/nar/gkab1112.

Schuller M, Butler RE, Ariza A, Tromans-Coia C, Jankevicius G, Claridge TDW, Kendall SL, Goh S, Stewart GR, Ahel I. 2021. Molecular basis for DarT ADP-ribosylation of a DNA base. *Nature* 596:597–602. DOI: 10.1038/s41586-021-03825-4.

Segal G, Feldman M, Zusman T. 2005. The Icm/Dot type-IV secretion systems of Legionella pneumophila and Coxiella burnetii. *FEMS Microbiology Reviews* 29:65–81. DOI: 10.1016/j.femsre.2004.07.001.

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* 27:135–145. DOI: 10.1002/pro.3290.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 35:1026–1028. DOI: 10.1038/nbt.3988.

Sugimura T, Miwa M. 1994. Poly(ADP-nbose). Historical perspective. *Molecular and Cellular Biochemistry* 138:5–12.

Taylor M, Ross K, Bentham R. 2009. Legionella, protozoa, and biofilms: interactions within complex microbial systems. *Microbial Ecology* 58:538–547. DOI: 10.1007/s00248-009-9514-z.

Tindall BJ. 2020. Taking a Closer Look at the Valid Publication and Authorship of Legionella

bozemanae Brenner et al. 1980, Fluoribacter bozemanae Garrity et al. 1980, Legionella pittsburghensis Pasculle et al. 1980, Legionella micdadei Hébert et al. 1980 and Tatlockia micdadei (Hébert et al. 1980) Garrity et al. 1980. *Current Microbiology* 77:146–153. DOI: 10.1007/s00284-019-01793-7.

Ueda K, Hayaishi O. 1985. ADP-Ribosylation. *Annual Review of Biochemistry* 54:73–100.

Vries SJ de, Bonvin AMJJ. 2011. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. *PLOS ONE* 6:e17695. DOI: 10.1371/journal.pone.0017695.

Vyas S, Chesarone-Cataldo M, Todorova T, Huang Y-H, Chang P. 2013. A systematic analysis of the PARP protein family identifies new functions critical for cell physiology. *Nature communications* 4:2240. DOI: 10.1038/ncomms3240.

Wang J, Li J, Hou Y, Dai W, Xie R, Marquez-Lago TT, Leier A, Zhou T, Torres V, Hay I, Stubenrauch C, Zhang Y, Song J, Lithgow T. 2021. BastionHub: a universal platform for integrating and analyzing substrates secreted by Gram-negative bacteria. *Nucleic Acids Research* 49:D651–D659. DOI: 10.1093/nar/gkaa899.

Wexler M, Zusman T, Linsky M, Lifshitz Z, Segal G. 2022. The Legionella genus core effectors display functional conservation among orthologs by themselves or combined with an accessory protein. *Current Research in Microbial Sciences* 3:100105. DOI: 10.1016/j.crmicr.2022.100105.

Wyżewski Z, Gradowski M, Krysińska M, Dudkiewicz M, Pawłowski K. 2021. A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution. *PeerJ* 9:e11051. DOI: 10.7717/peerj.11051.

Xu D, Jaroszewski L, Li Z, Godzik A. 2014. FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* 30:660–667. DOI: 10.1093/bioinformatics/btt578.

Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* 33:2302–2309. DOI: 10.1093/nar/gki524.

van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, Melquiond ASJ, van Dijk M, de Vries SJ, Bonvin AMJJ. 2016. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* 428:720–725. DOI: 10.1016/j.jmb.2015.09.014.

From: BMC Genomics <bmcgenomics@biomedcentral.com>
Date: Friday, August 8, 2025 at 2:36 PM
To: Krzysztof Pawlowski <Krzysztof.Pawlowski@UTSouthwestern.edu>
Subject: BMC Genomics: Decision on your manuscript

Ref: Submission ID 142c620f-28da-4cfd-9b9f-331e6a1f8304

Dear Dr Pawłowski,

Re: "A survey of ADP-ribosyltransferase families in the pathogenic Legionella"

We're delighted to let you know that your manuscript has been accepted for publication in BMC Genomics.

Editor comments

"The authors have addressed all comments made by the reviewers and myself, and the manuscript is now ready for publication. Congratulations! Looking ahead, authors should treat code and data sharing as a fundamental requirement for research in genomics, bioinformatics, and related fields."

-Andres Iriarte

Prior to publication, our production team will check the format of your manuscript to ensure that it conforms to the journal's requirements. They will be in touch shortly to request any necessary changes, or to confirm that none are needed.

Checking the proofs

Once we've prepared your paper for publication, you will receive a proof. At this stage, for the main text, only errors that have been introduced during the production process, or those that directly compromise the scientific integrity of the paper, may be corrected.

As the corresponding (or nominated) author, you are responsible for the accuracy of all content, including spelling of names and current affiliations.

To ensure prompt publication, your proofs should be returned within two working days.

Publication policies

Acceptance of your manuscript is conditional on all authors agreeing to our publication policies at: https://urldefense.com/v3/__https://www.springernature.com/gp/policies/editorial-policies__;!!MznTZTSvDXGV0Co!HgC7rUsgH6kT61j-ZtD_vswuQWROHlNHFNadfihbwj9e-viP_eTjuj652pj0-OGe8SyxnOqXmUJARKiP61ZCDLpTTWb10JBrFRK8m6UQ0i8Mjw$

Licence to Publish and Article Processing Charge

As the corresponding author of an accepted manuscript, your next steps will be to complete an Open Access Licence to publish on behalf of all authors, confirm your institutional affiliation, and arrange payment of your article-processing charge (APC). You will shortly receive an email with more information.

Once again, thank you for choosing BMC Genomics, and we look forward to publishing your article.

Kind regards,

Sneha Kanade

Editor

BMC Genomics

Warszawa, 1.09.2025 r.

Marianna Krysińska
marianna_krysinska@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Gradowski M, Baranowski B, Pawłowski K, Dudkiewicz M. 2025. A survey of ADP-ribosyltransferase families in the pathogenic *Legionella*. BMC Genomics 26, 915** mój indywidualny udział w jej powstaniu polegał na opracowaniu koncepcji projektu, przeprowadzeniu większości eksperymentów, w tym obliczeń i analiz dotyczących: wykrywania odległych homologów, podobieństw sekwencyjnych i strukturalnych, modelowania białek, otoczeń genomicznych, analiz predykcji efektorów; kompleksowym przeanalizowaniu danych uzyskanych we wszystkich eksperymentach, przygotowaniu grafik, wykresów i tabel do głównej części artykułu, napisaniu pierwszej wersji artykułu, przygotowaniu danych i tabel do suplementu, współredagowaniu artykułu oraz zatwierdzeniu ostatecznej wersji manuskryptu.

Podpis

Warszawa, 1.09.2025

dr Marcin Gradowski
marcin_gradowski@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Gradowski M, Baranowski B, Pawłowski K, Dudkiewicz M. 2025. A survey of ADP-ribosyltransferase families in the pathogenic *Legionella*. BMC Genomics 26, 915** mój indywidualny udział w jej powstaniu polegał na przeprowadzeniu części eksperymentów dotyczących porównań strukturalnych, napisaniu oprogramowania do analizy sieci podobieństw strukturalnych, przygotowaniu części tabeli uzupełniającej (suplement) i zatwierdzeniu ostatecznej wersji artykułu.

Podpis

Warszawa, 29.09.2025 r.

dr Bartosz Baranowski
bartosz.piotr.baranowski@gmail.com

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Gradowski M, Baranowski B, Pawłowski K, Dudkiewicz M. 2025. A survey of ADP-ribosyltransferase families in the pathogenic *Legionella*. BMC Genomics 26, 915** mój indywidualny udział w jej powstaniu polegał na przeprowadzeniu części obliczeń dotyczących otoczeń genomicznych i przeanalizowaniu uzyskanych danych oraz zatwierdzeniu ostatecznej wersji manuskryptu.

Podpis

Dallas, 30.10.2025

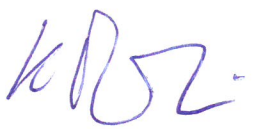
dr hab. Krzysztof Pawłowski
Krzysztof.Pawlowski@utsouthwestern.edu

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Gradowski M, Baranowski B, Pawłowski K, Dudkiewicz M. 2025. A survey of ADP-ribosyltransferase families in the pathogenic *Legionella*. BMC Genomics 26, 915** mój indywidualny udział w jej powstaniu polegał na udziale w opracowaniu koncepcji projektu, współnadzorowaniu projektu, współautorstwie roboczych wersji artykułu oraz zatwierdzeniu ostatecznej wersji artykułu.



Podpis

Warszawa, 29.07.2025 r.

dr hab. Małgorzata Dudkiewicz
małgorzata_dudkiewicz@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Gradowski M, Baranowski B, Pawłowski K, Dudkiewicz M. 2025. A survey of ADP-ribosyltransferase families in the pathogenic *Legionella*. BMC Genomics 26, 915** mój indywidualny udział w jej powstaniu polegał na udziale w opracowaniu koncepcji projektu, współnadzorowaniu projektu, przeprowadzeniu niektórych eksperymentów – obliczeń dotyczących dokowania i przeanalizowaniu uzyskanych danych, przygotowaniu dodatkowych rysunków (suplement), pisaniu i zrecenzowaniu roboczych wersji artykułu oraz zatwierdzeniu jego ostatecznej wersji.

Podpis

A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution

Zbigniew Wyżewski[1,*], Marcin Gradowski[2,*], Marianna Krysińska[2], Małgorzata Dudkiewicz[2] and Krzysztof Pawłowski[2,3,4]

[1] Institute of Biological Sciences, Cardinal Stefan Wyszynski University in Warsaw, Warszawa, Poland
[2] Department of Biochemistry and Microbiology, Warsaw University of Life Sciences - SGGW, Warszawa, Poland
[3] Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX, United States
[4] Department of Translational Medicine, Lund University, Lund, Sweden
[*] These authors contributed equally to this work.

ABSTRACT

The presence of many completely uncharacterized proteins, even in well-studied organisms such as humans, seriously hampers full understanding of the functioning of the living cells. ADP-ribosylation is a common post-translational modification of proteins; also nucleic acids and small molecules can be modified by the covalent attachment of ADP-ribose. This modification, important in cellular signalling and infection processes, is usually executed by enzymes from the large superfamily of ADP-ribosyltransferases (ARTs). Here, using bioinformatics approaches, we identify a novel putative ADP-ribosyltransferase family, conserved in eukaryotic evolution, with a divergent active site. The hallmark of these proteins is the ART domain nestled between flanking leucine-rich repeat (LRR) domains. LRRs are typically involved in innate immune surveillance. The novel family appears as putative novel ADP-ribosylation-related actors, most likely pseudoenzymes. Sequence divergence and lack of clearly detectable "classical" ART active site suggests the novel domains are pseudoARTs, yet atypical ART activity, or alternative enzymatic activity cannot be excluded. We propose that this family, including its human member LRRC9, may be involved in an ancient defense mechanism, with analogies to the innate immune system, and coupling pathogen detection to ADP-ribosyltransfer or other signalling mechanisms.

Subjects Biochemistry, Bioinformatics, Biophysics, Evolutionary Studies, Molecular Biology
Keywords ADP-ribosyltransferases, Evolution, Protein domains, Pseudoenzymes, Protein structure and function prediction

INTRODUCTION

ADP-ribosyltransferases (ARTs) are enzymes catalyzing the transfer of ADP-ribose from oxidized nicotinamide adenine dinucleotide (NAD^+) to different acceptor molecules. Thus, they are responsible for chemical modification of various targets such as proteins, nucleic acids and small molecules. These enzymes are highly conserved in evolution and widespread in nature. They are common to all three domains of life: the Archaea, the Bacteria and the Eukarya. Multiplicity of ADP-ribosylated substrates is reflected in a variety

Submitted 24 July 2020
Accepted 11 February 2021
Published 10 March 2021

Corresponding authors
Małgorzata Dudkiewicz, malgorzata.dudkiewicz@gmail.com
Krzysztof Pawłowski, Krzysztof_Pawlowski@sggw.edu.pl, Krzysztof.Pawlowski@utsw.edu

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj.11051**

of functions performed by ARTs (*Aravind et al., 2015*; *Cohen & Chang, 2018*; *Munnur & Ahel, 2017*; *Otto et al., 2005*; *Teloni & Altmeyer, 2016*).

Prokaryotic ADP-ribosyltransferases often play a role of bacterial toxins or effectors. Several studies show that ADP-ribosylation of the host target molecules by bacterial ARTs is correlated with the development of infection (*Aktories et al., 2011*; *Bhogaraju et al., 2016*; *Boyer et al., 2006*; *Kalayil et al., 2018*; *Simon, Aktories & Barbieri, 2014*). At the intracellular scale, ADP-ribosylation may change apoptotic potential of infected cell (*Klockgether & Tümmler, 2017*) and/or disturb organization of cellular membranes and actin cytoskeleton (*Kagan & Roy, 2002*; *Komander & Randow, 2017*; *Maresso et al., 2007*) whereas at systemic level, it can disturb cell-mediated immune response (*Bhogaraju et al., 2016*; *Klockgether & Tümmler, 2017*), or increase the permeability of barriers limiting bacterial spread (epithelium and endothelium) (*Boyer et al., 2006*; *Klockgether & Tümmler, 2017*) and consequently contribute to serious structural and functional disorders of the host tissues and organs (*Chiu et al., 2009*; *Munro et al., 2010*). Some ARTs act on small molecules, e.g., the bacterial Arr enzymes that ADP-ribosylate the antibiotic rifamycin (*Baysarowich et al., 2008*).

Eukaryotic ADP-ribosyltransferases play important roles in both physiological and pathophysiological processes, contributing to changes in chemical properties of the ADP-ribose acceptors. Proteins, post-translationally modified by ARTs, can also lose the capacity to interact with their ligands or acquire the ability to bind new ones. ADP-ribosylation of enzymes may cause substantial changes in their catalytic activity (*Jwa & Chang, 2012*; *Liu & Yu, 2015*). ADP-ribose moieties may function as signals that direct modified acceptors to ubiquitin-dependent proteolysis (*Aravind et al., 2015*; *Cohen & Chang, 2018*). Influencing the half-life of the target molecules, ADP-ribosyltransferases determine their intracellular level and thus affect their activity (*Cohen & Chang, 2018*). Moreover, ADP-ribose moieties may form molecular scaffolds with a negative charge. Such structures play a role in recruitment of positively charged proteins, favoring specific intermolecular interactions (*Leung et al., 2012*). Poly-ADP-ribosylation of DNA-binding proteins (e.g., core histones (*Javle & Curtin, 2011*; *Weaver & Yang, 2013*), chromatin remodeling enzymes (e.g., histone demethylase KDM5B (*Krishnakumar & Kraus, 2010*), DEK protein (*Gamble & Fisher, 2007*) and DNA repair factors (*Kim et al., 2015*) influences genome organization (*Weaver & Yang, 2013*) and expression (*Gamble & Fisher, 2007*; *Krishnakumar & Kraus, 2010*) as well as effective DNA repair (*Javle & Curtin, 2011*; *McCann, 2019*). Recently, growing evidence shows that PARPs also directly ADP-ribosylate mRNA (*Kim et al., 2020*). Other examples of nucleic acid ADP-ribosylation are provided by the bacterial toxin DarT that acts on ssDNA (*Lawaree et al., 2020*) and the toxin scabin, acting on mononucleosides, nucleotides, and both single-stranded and dsDNA (*Lyons et al., 2018*).

The members of the poly-ADP-ribosyltransferase (PARP) family, PARP1 and PARP2, are examples of ARTs that modify proteins interacting with DNA. PARP1 and PARP2 are localized in the nucleus where they are involved in many cellular processes (*Choi et al., 2016*; *Liang et al., 2013*; *Riccio, Cingolani & Pascal, 2016*; *Szántó et al., 2014*). PARP1 and PARP2 are engaged in DNA repair, that is elimination of DNA damage such as deamination, hydroxylation and methylation as well as for repair of single strand breaks.

PARP1 and other PARPs also play a role in maintaining telomere stability (*Bai, 2015*; *De Vos, Schreiber & Dantzer, 2012*). PARPs are known as EMA/FDA approved drug targets, PARP inhibitors are used in treatment of prostate, breast and ovarian cancers. (*Curtin, 2005*; *Curtin & Szabo, 2013*; *Kamel et al., 2018*; *Kummar et al., 2012*).

In general, ADP-ribosyltransferases are responsible for regulation of intracellular and extracellular signal transduction. Therefore, their activity determines viability and proliferation potential of cells, DNA stability, immune system reactivity and thus the proper functioning of eukaryotic organisms (*Bai, 2015*; *Corda & Di Girolamo, 2002*; *De Vos, Schreiber & Dantzer, 2012*; *Yamamoto et al., 2017*). On the other hand, ADP-ribosyltransferases may also be involved in development of pathological states such as neurodegenerative disorders, diabetes, atherosclerosis, cataract and cancer (*Barreiro & Gea, 2018*; *Gunderson & Moore, 2015*; *Kamboj et al., 2013*; *Malyuchenko et al., 2015*; *Mangerich & Bürkle, 2012*; *Rossi, Ghosh & Bohr, 2010*). Similarly to protein phosphorylation, ADP-ribosylation is a reversible post-translational modification, performed by a trio of "writers" (ARTs), "readers" (e.g., macro domain proteins) and "erasers" (e.g., some macro domains, ADP-ribose hydrolases, and NUDIX phosphodiesterases) (*O'Sullivan et al., 2019*).

According to the Pfam database, the ART clan (superfamily) comprises 23 families of domains, fourteen of which, ADPrib_exo_Tox, ART, DarT, DUF2441, DUF3990, DUF952, Enterotoxin_a, Exotox-A_cataly, PARP, Pertussis_S1, PTS_2-RNA, RES, RolB_RolC and TNT, include eukaryotic members (*El-Gebali et al., 2019*). However, some of these 14 families are predominantly bacterial with just a handful of eukaryotic members, e.g. ADPrib_exo_Tox and DarT. PARPs, the best studied family of ADP-ribosyltransferases, are responsible for modification of target structures by covalently adding polymeric chains of ADP-ribose moieties instead of transferring only one moiety. Despite low conservation of amino acid sequence, ART clan members are characterized by a common spatial structure comprising a split β-sheet and two helical regions surrounding it. The "split" separates the β-sheet into two halves, each composed of three β-strands (4-5-2 and 1-3-6, respectively) (*Aravind et al., 2015*; *Cohen & Chang, 2018*). Aravind and colleagues divide ARTs into three main clades: the H-H-h clade, the H-Y-[EDQ] clade and the R-S-E clade (*Aravind et al., 2015*; *Cohen & Chang, 2018*) that are characterized by the different configurations of active site amino acid residues. In the H-H-h clade domains, the catalytic centre comprises two histidines and one hydrophobic residue supplied by β-strands 1, 2 and 5, respectively. The H-Y-[EDQ] clade domains, including PARPs, are characterized by active site composed of histidine, tyrosine and glutamate/aspartate/glutamine whereas in the R-[ST]-E clade domains, active site triad comprises arginine, a polar residue (serine/threonine) and glutamate (*Aravind et al., 2015*). Three families of the ART clan, PARP, PTS_2-RNA and ART, are present in humans (Table 1), and contain 16, 1 and 4 representatives, respectively. The NEURL-4 ART-like domain family (*De Souza & Aravind, 2012*), not included in Pfam database, is the fourth human ART-like family. The full catalogue of ADP-ribosylation enzymes is likely far from completion, as one may expect from recent discoveries of novel ART domains in effectors from pathogenic bacteria that perform non-canonical ubiquitination (*Akturk et al., 2018*; *Kalayil et al., 2018*; *Kim et al., 2018*), novel ART/macro pairs in bacterial toxin/antitoxin systems (*Jankevicius et al., 2016*) and novel human macro

Table 1 List of known human ART-like domains. Full list of ART-like domains from human, including the novel DUF3715 domains in TASOR, TASOR2 and TEX15.

Name	Gene names*	family classification for ART Domain (Pfam)		Triad motif	R/H-G-T/S motif in β-strand 1	S-X-S/Y-X-X motif in β-strand 2	X-X-E motif at front edge of β-strand 5	ADP-ribosylationactivity: mono (M), oligo (O), poly (P)	Cellular localization	ADP-ribosylation target: protein (P), DNA (D), RNA (R), and auto-ADP-ribosylation (A)	Total length (ART-like domain length)
TRPT1	TRPT1	PTS_2-RNA	PF01885	H-H-V	HGT	HLA	NGV	M	mitochondrion, nucleus, endoplasmic reticulum, cytosol	R	253 (181)
ART1	ART1	ART	PF01129	R-S-E	RGV	SAS	EEE	M	sarcoplasmic reticulum membrane, plasma membrane, ER membrane and lumen, glycosylphosphatidylinositol (GPI)-anchored ectoenzymes	P	327 (222)
ART3	ART3 TMART	ART	PF01129	K-L-V	RTS	SAK	ERI	M	cell membrane, extracellular region, extracellular exosome	P	389 (222)
ART4	ART4 DO DOK1	ART	PF01129	G-S-E	YRT	STS	KKE	M	cell membrane, extracellular region, extracellular exosome, glycosylphosphatidylinositol (GPI)-anchored proteins	P	314 (222)
ART5	ART5 UNQ575/PRO1137	ART	PF01129	R-S-E	RGV	SSS	ERE	M	secreted	P	291 (226)
PARP1	**PARP1 ADPRT PPOL**	PARP	PF00644	H-Y-E	HGS	YFA	YNE	P	nucleus	D, P	1014 (227)
PARP2	**PARP2 ADPRT2 ADPRTL2**	PARP	PF00644	H-Y-E	HGS	YFA	YNE	P	nucleus	D, P, A	583 (228)
PARP3	**PARP3 ADPRT3 ADPRTL3**	PARP	PF00644	H-Y-E	HGT	YFA	QSE	M	nucleus	D, P, A	533 (221)
PARP4	**PARP4 ADPRTL1 KIAA0177 PARPL**	PARP	PF00644	H-Y-E	HGS	YFS	DDE	M	nucleus, exosomes, cell membrane, spindle	P	1724 (205)
PARP5A (tankyrase-1)	TNKS PARP5A PARPL TIN1 TINF1 TNKS1	PARP	PF00644	H-Y-E	HGS	YFA	YAE	O, P	nucleus, telomeres, Golgi apparatus, cytoplasm	P, A	1327 (206)
PARP5B (tankyrase-2)	TNKS2 PARP5B TANK2 TNKL	PARP	PF00644	H-Y-E	HGS	YFA	LAE	O, P	nucleus, telomeres, Golgi apparatus, cytoplasm	P, A	1166 (206)
PARP6	PARP6	PARP	PF00644	H-Y-I	HGS	YLS	GEI	M	cytoplasm	P, A	630 (227)
PARP8	PARP8	PARP	PF00644	H-Y-I	HGS	YLS	GNI	M	nucleus, endoplasmic reticulum	P, A	854 (228)
PARP16	PARP16 ARTD15 C15orf30	PARP	PF00644	H-Y-Y	HGS	YLT	PKY	M	cell membrane, endoplasmic reticulum	P, A	322 (186)

(*continued on next page*)

Table 1 (*continued*)

Name	Gene names*	family classification for ART Domain (Pfam)		Triad motif	R/H-G-T/S motif in β-strand 1	S-X-S/Y-X-X motif in β-strand 2	X-X-E motif at front edge of β-strand 5	ADP-ribosylationactivity: mono (M), oligo (O), poly (P)	Cellular localization	ADP-ribosylation target: protein (P), DNA (D), RNA (R), and auto-ADP-ribosylation (A)	Total length (ART-like domain length)
PARP15	*PARP15 BAL3*	PARP	PF00644	H-Y-L	HGT	YFA	PKL	M	cytoplasm	R, P, A	678 (197)
PARP14	*PARP14 BAL2 KIAA1268*	PARP	PF00644	H-Y-L	HGT	YFA	PSL	M	cytoplasm, nucleus	P, A	1801 (197)
PARP10	*PARP10*	PARP	PF00644	H-Y-I	HGT	YFA	PSI	M	nucleus, cytoplasm	R, P, A	1025 (220)
PARP13	*ZC3HAV1 ZC3HDC2 PRO1677*	PARP	PF00644	Y-Y-V	YAT	YFA	PSV	inactive	cytoplasm	–	902 (187)
PARP7	*TIPARP PARP7*	PARP	PF00644	H-Y-I	HGT	YFA	PQI	M	nucleus, cytoplasm	P, A	657 (209)
PARP12	*PARP12 ZC3HDC1*	PARP	PF00644	H-Y-I	HGT	YFA	PSI	M	cytoplasm	P, A	701 (215)
PARP11	*PARP11 C12orf6*	PARP	PF00644	H-Y-I	HGT	YFA	PKI	M	nucleus (in mice)	R, P, A	338 (216)
PARP9	*PARP9 BAL, BAL1*	Undetected **/insignificant		Q-Y-T	QQV	YFT	PET	M	nucleus, cell membrane, cytoplasm, mitochondrion	ubiquitin	854 (223)
TASOR	TASOR	DUF3715 (partial alignment)	PF12509	L-Y-Q	LMV	YLS	LTQ	inactive	nucleus, chromosome	–	1670 (221)
TASOR2	TASOR2	DUF3715 (partial alignment)	PF12509	*V-T-E*	*VVA*	*TLD*	*LLE*	not determined	cytoplasm, nucleoplasm	not determined	2430 (199)
TEX15	TEX15	DUF3715 (partial alignment)	PF12509	*L-Y-S*	*LAL*	*YMF*	*VLS*	not determined	nucleus, cytoplasm	not determined	2789 (197)
NEURL4	NEURL4	Undetected**		*H-L-E*	*HGS*	*LLS*	*ELE*	not determined	cytoplasm, cytoskeleton	not determined	1562 (121)
LRRC9	LRRC9	Undetected**		*Y-V-K*	*YVF*	*EFL (SIS)*	*QCK*	not determined	not determined	not determined	1453 (233)

Notes.

*PARP family members are sorted by subgroup (font format) on the basis of similarities in amino acid sequence, intron positions and associated protein domains (*Otto et al., 2005*).

**For these domains, Pfam family assignments cannot be made using standard Pfam HMM tool.

Italicized motifs were identified based on HHpred and FFAS03 alignments to canonical PARPs, in some cases the two methods were not in agreement.DUF3715 domains correspond to a part of TASOR-like ART domains.

domains (*Dudkiewicz & Pawłowski, 2019*). Other examples of novel ADP-ribosylation players are provided by the viral macro domains, present in many dangerous viruses, including the SARS and SARS-CoV-2 coronaviruses (*Saikatendu et al., 2005*), and by the recently characterized novel ART-like domain (DUF3715) in the human TASOR protein involved in gene silencing (*Douse et al., 2020*; *Tchasovnikarova et al., 2015*).

In this paper, building up on experience in identification of novel enzyme families (*Dudkiewicz, Lenart & Pawłowski, 2013*; *Dudkiewicz & Pawłowski, 2019*; *Dudkiewicz et al., 2012*; *Pawłowski et al., 2006*; *Sreelatha et al., 2018*), we identified structural similarity between ART-like catalytic domains and a family of eukaryotic uncharacterized protein domains present in homologs of the leucine-rich repeat containing protein 9 (LRRC9). For clarity, we will call this domain LRRC9-ART. Human LRRC9 is annotated in the UniProt database as protein of unknown function with experimental evidence at transcript level. LRRC9 locus is located on chromosome 14 (q23.1). Gene and protein names refer to the twenty-two leucine-rich repeats located between positions 97 and 246 and 717-1365. According to The Human Protein Atlas (*Uhlen et al., 2015*), expression of LRRC9 mRNA is enhanced in brain, pituitary gland and testis.

We investigated sequence similarities between the novel and the known ADP-ribosyltransferases. We reconstructed phylogenetic relationships between sequences within LRRC9 domain family and other ART clan families. We also explored sequence variability in the novel ART-like domain in the context of predicted three-dimensional structures and proposed their likely biological functions.

MATERIALS AND METHODS

The FFAS03 (*Xu et al., 2014*), HHpred (*Zimmermann et al., 2017*) and Phyre2 (*Kelley et al., 2015*) servers were used to determine distant sequence similarities of LRRC9 central domain to proteins of known structures from public databases. Standard parameters and significance thresholds were selected.

The representative set of LRRC9 ART-like domain sequences was collected by submitting the ART-like domain of the human LRRC9 protein (UniProtKB: Q6ZRR7.2, positions 389-628) to two iterations of JackHMMER (also with standard parameters) ran on the Reference Proteomes database.

The MAFFT program (*Katoh, Rozewicki & Yamada, 2019*) was used to build the multiple sequence alignments of the novel domains and PARP catalytic domains obtained from the rp75 set from the Pfam database (*El-Gebali et al., 2019*). In-house scripts were used to merge family-wise multiple sequence alignments according to a FFAS pairwise alignment of representatives of the two families. The WebLogo server (*Crooks et al., 2004*) was used to visualize results as sequence logos.

The CLANS (*Frickey & Lupas, 2004*) algorithm was used to visualize close and distant similarities between the putative and the known ADP-ribosyltransferase domains. Sequence similarities up to BLAST *E*-value of 1, $1e^{-2}$ or $1e^{-4}$ and the BLOSUM45 substitution matrix were used. In order to acquire collection of the known ADP-ribosyltransferase domains, the following sets of sequences were obtained from Pfam database: ADPrib_exo_Tox

rp15, ADPRTs_Tse2 rp15, Anthrax-tox_M rp75, Arr-ms rp15, ART rp15, ART-PolyVal rp15, AvrPphF-ORF-2 rp15, Diphtheria_C rp75, Dot_icm_IcmQ rp35, DUF2441 rp15, DUF952 rp15, Enterotoxin_a rp15, Exotox-A_cataly rp75, NADase_NGA rp35, PARP rp15, Pertussis_S1 rp15, PTS_2-RNA rp15, RolB_RolC rp55 and TNT rp15. Next, collected data were supplemented with three sequence sets representing the following ART-like domain families not included in Pfam database: ESPJ, NEURL4 and DUF3715/TASOR. Sets of ESPJ and SidE domain sequences were obtained by submitting sequences from UniProtKB and Protein NCBI databases (UniProtKB: W0AL45, positions 52-214 and refseq ID: YP_094288.1, positions 686-912, respectively) to two iterations of JackHMMER with standard parameters, ran on UniProtKB database. The input sequence of NEURL4 (UniProtKB: F2TYZ7, positions 102-275) was subject to the same procedure, but extended to three iterations and ran on the Reference Proteomes database. Set of DUF3715 sequences was obtained by submitting the TASOR ART domain (UniProtKB: Q9UK61.3, positions 92-337) to two iterations of JackHMMER with standard parameters. Next, the whole collection of putative and known domain sequences were prepared for the CLANS procedure by clustering with CD-HIT and selecting representatives at a 70% sequence identity threshold (*Huang et al., 2010*).

In order to determine phylogenetic spread of the novel ART domains, a set of LRRC9 ART-like domain homologs was obtained using JackHMMER search seeded with input sequence UniProtKB: Q6ZRR7.2, positions 389-628. Then, Phylogeny.fr platform (*Dereeper et al., 2008*) was utilized to construct phylogenetic tree. A multiple sequence alignment was generated using the MUSCLE program (*Edgar, 2004*). Next, advanced mode of the platform was utilized to build phylogenetic tree using the maximum likelihood PhyML program (*Guindon et al., 2009*) with standard settings (WAG amino-acid substitution model, four substitution rate categories, aLRT test for bootstrap support). Phylogenetic relationships between homologs were visualized as a dendrogram using the iTOL server (*Letunic & Bork, 2016*).

Three three-dimensional structure models for the human LRRC9-ART domain were built using homology modelling method Modeller9v21 (*Webb & Sali, 2017*), I-Tasser server (*Yang et al., 2015*) and Robetta server (*Hiranuma et al., 2020*). Simple homology model (Modeller) was based on human PARP10 structure as single template (PDB ID: 3HKV, identity 16%). The I-Tasser multi template model (*Yang et al., 2015*) were based on best PDB hits identified automatically by the server, top three of them were human tankyrase 2 (PDB ID:3MHK, 3KR7 identity 17%), PARP10 (PDB ID 3HKV, identity 16%) and human poly(ADP-ribose) polymerase 14 (PDB ID 3SMJ, identity 16%). The Robetta model was constructed using a deep learning-based method, trRosetta (*Yang et al., 2020*).

Putative NAD binding pockets in all modelled structures and chosen native PARP structures were first found and described using DoGSiteScorer fully automatic algorithm for pocket and druggability prediction available at ProteinsPlus server (*Volkamer et al., 2012*) and independently identified using the COFACTOR algorithm (*Roy & Zhang, 2012*), an option delivered by the I-Tasser (Iterative Threading ASSEmbly Refinement) server for protein structure and function prediction (*Yang et al., 2015*) based on threading the query structural model through the BioLiP protein function database. COFACTOR identifies

potential functional sites by local and global structure matches and suggests annotations. Putative poses of NAD ligand in the LRRC9-ART trRosetta model were predicted by docking using AutodockVina module in UCSF Chimera (*Trott & Olson, 2010*). Structures were visualized and analyzed using UCSF Chimera tools (*Huang et al., 2014*).

The three-domain model for full-length human LRRC9 protein was built using AIDA (Ab Initio Domain Assembly Server) (*Xu et al., 2015*) based on structures modelled using Modeller9v21 separately for three identified domains: N-terminal LRR region with helical fragment (PDB code of modelling template: 3OJA, 15% identity), ART-like domain (template: 3HKV), and C- terminal LRR region domain (template: 4LSX, 21% identity).

Conservation values of sequence positions in multiple sequence alignment created for the human LRRC9-ART domain homologues (for mapping onto the modelled structures), were derived from Jalview alignment editor (*Waterhouse et al., 2009*) and were automatically calculated as the number of conserved physicochemical properties for each column of the alignment (*Livingstone & Barton, 1993*).

The LRR repeats were identified using the LRRFinder tool (*Offord, Coffey & Werling, 2010*). Gene Ontology analysis of human genes encoding LRR-containing proteins (i.e., their cellular/extracellular localization, molecular function and association with biological processes) was performed with the use of Panther Classification System (*Mi et al., 2017*). The domain organization of LRRC9 proteins was visualised using the DOG2.0 tool (*Ren et al., 2009*).

For analysis of protein and mRNA expression, the Proteomics-db, PAXdb, neXtProt and Protein Atlas databases were used (*Samaras et al., 2020*; *Thul & Lindskog, 2018*; *Wang et al., 2015*; *Zahn-Zabal et al., 2020*). The BioGRID Interaction Database was used to obtain information about possible place of LRRC9 in human protein-protein interaction networks (*Oughtred et al., 2019*). For prediction of the subcellular localization of LRRC9, the DeepLoc server was used (*Almagro Armenteros et al., 2017*).

RESULTS

Assignment of the central domain of LRRC9 to the ADP-ribosyl clan

A sequence similarity screen of human proteins for remote homologs of ADP-ribosyltransferases (ARTs) using the FFAS03 method indicated that weak ART similarity might exist in the central region of the LRRC9 protein (Figs. 1, 2). In order to examine in detail this similarity, we used FFAS03, HHpred and Phyre2 servers. All of them are dedicated to protein structure prediction, allowing detection of non-obvious sequence similarity between very distant protein homologs. These three independent bioinformatics tools revealed statistically significant albeit distant sequence similarity of a region of human LRRC9 protein to representatives of the PARP catalytic domain family between positions 392 and 625 (Table 2). Such remote sequence similarity cannot be the basis for drawing a definite conclusion about an enzymatic function of the putative ART domain.

The relationship of the novel LRRC9 ART-like family, found in many eukaryotic lineages (Fig. 3), to the known ART families can be visualized using the sequence-based CLANS clustering approach (Fig. 4). CLANS analysis was performed at three different *E*-value

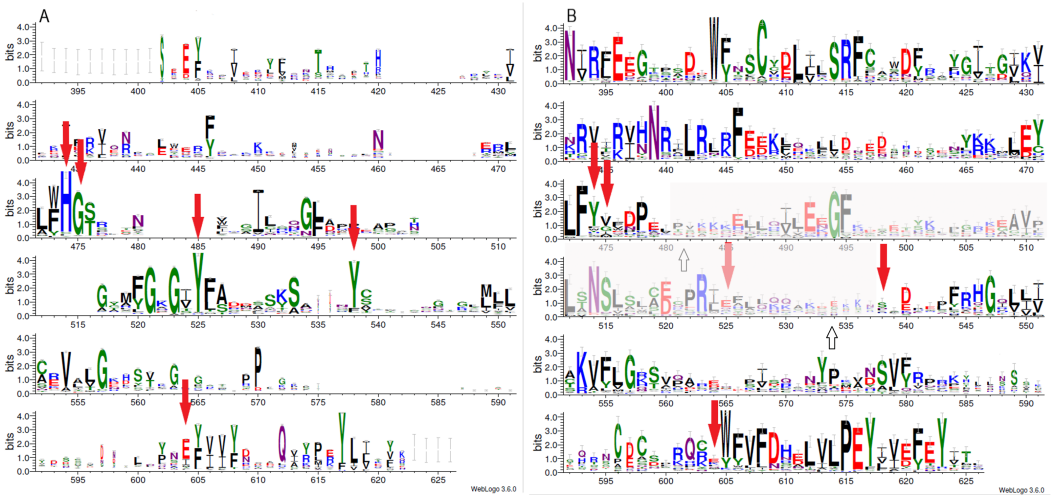


Figure 1 Sequence variability represented as sequence logos for PARP family (A) and LRRC9 ART-like domains (B). Logos were created from two separate MAFFT alignments for PARP family and LRRC9 sequences using WebLogo3 server. The two alignments were merged based on a FFAS03D alignment obtained for human PARP10 and LRRC9-ART domains. Both alignments were trimmed by removing positions with gaps in the human LRRC9 ART domain. Sequence numbering is according to human LRRC9. The catalytically important PARP residues His, Gly, Tyr, Tyr and Glu [matched to positions 474, 475, 525, 538 and 604, respectively, of the human LRRC9 protein] marked with red arrows. Region where FFAS and HHpred alignments between PARP and LRRC9 are inconsistent (between two white arrows) has been faded.

Full-size DOI: 10.7717/peerj.11051/fig-1

LRRC9_HUMAN

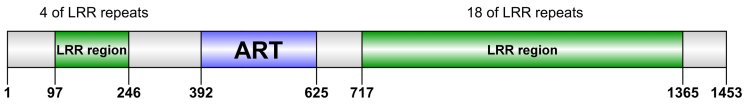


Figure 2 Domain organization of LRRC9 proteins. Sequence numbering as per human LRRC9.

Full-size DOI: 10.7717/peerj.11051/fig-2

Table 2 Structure predictions for the LRRC9 ART-like domain. Structure and distant sequence similarity predictions for human LRRC9 ART-like region (residues 389-628) obtained using different bioinformatics tools. For each method, only the first hit shown.

Bioinformatic tool for structure prediction	Top hit: PDB code, name	Statistical significance for top hit	Region of query aligned to the hit	Sequence identity
FFAS03	**4DVI** Tankyrase 1 with IWR2	Z-score =-12.500	427-625	20%
HHpred	**5LX6,** Human PARP10 (ARTD10), catalytic fragment in complex with PARP inhibitor Veliparib [*Homo sapiens*]	*E*-value =6.2e−23	392-624	15%
Phyre2	**3HKV**, Human poly(ADP-ribose) polymerase 10, catalytic fragment in complex with an inhibitor 3-aminobenzamide; chain A	Confidence= 100%	395-624	16%

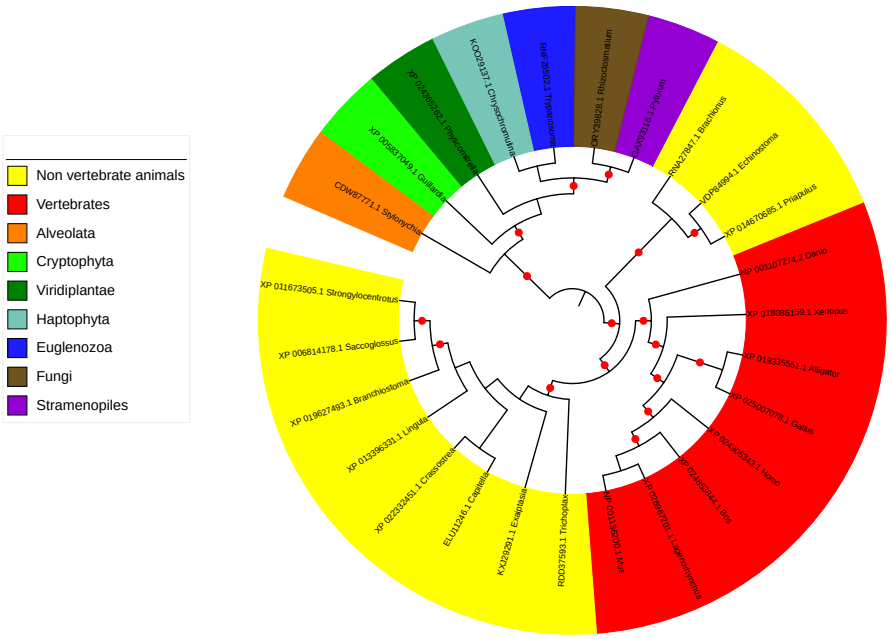


Figure 3 **Phylogenetic tree and taxonomic spread of LRRC9 ART-like domains.** A maximum likelihood phylogenetic tree for representative LRRC9 ART-like domains. NCBI protein identifiers and genus names shown. Colored by taxon. Red circles denote branches with bootstrap support of at least 75%.
Full-size DOI: 10.7717/peerj.11051/fig-3

levels for BLAST hits for all known families containing ADP-ribosyltransferase domains. In agreement with the FFAS03, HHpred and Phyre results, CLANS results suggest a closer relation between the novel ART family and the PARP domains, noticeable at *E*-value threshold 1, 0.01 and even at 0.0001. This, together with sequence conservation analysis allowed us to hypothesize that LRRC9-ART belongs to the H-Y-[EDQ] clade. The CLANS clustering analysis also suggests that LRRC9-ART should be regarded as a separate family within the ART-like clan/superfamily. The same applies to the DUF3715 family that was recently shown to possess an ART-like structure (*Douse et al., 2020*). Representatives of both families clearly cluster away from other ART-like clan members and the two novel families are also away from each other. Notably, detection of very distant sequence similarities depends on accurate domain boundary prediction. This may have hindered the recognition of the ART-like domain in TASOR whereas the domain of unknown function DUF3715 as defined in the Pfam database was missing approx. 50 residues at its N-terminus.

Putative active site of the novel ART-like family

Catalysis of poly-ADP-ribosylation performed by the members H-Y-[EDQ] clade (including PARP domain family) involves three non-consecutive conserved amino acid residues: His, Tyr and an acidic one (Glu or Asp) that can also be substituted by Gln. For consistency, these motifs will be called motifs I, II and III, respectively. Histidine, occurring in the conserved motif I Hx[ST] (*Aravind et al., 2015*) within β-strand 1, is responsible for binding the 2-OH group of the adenosine ribose of NAD^+ and NH_2

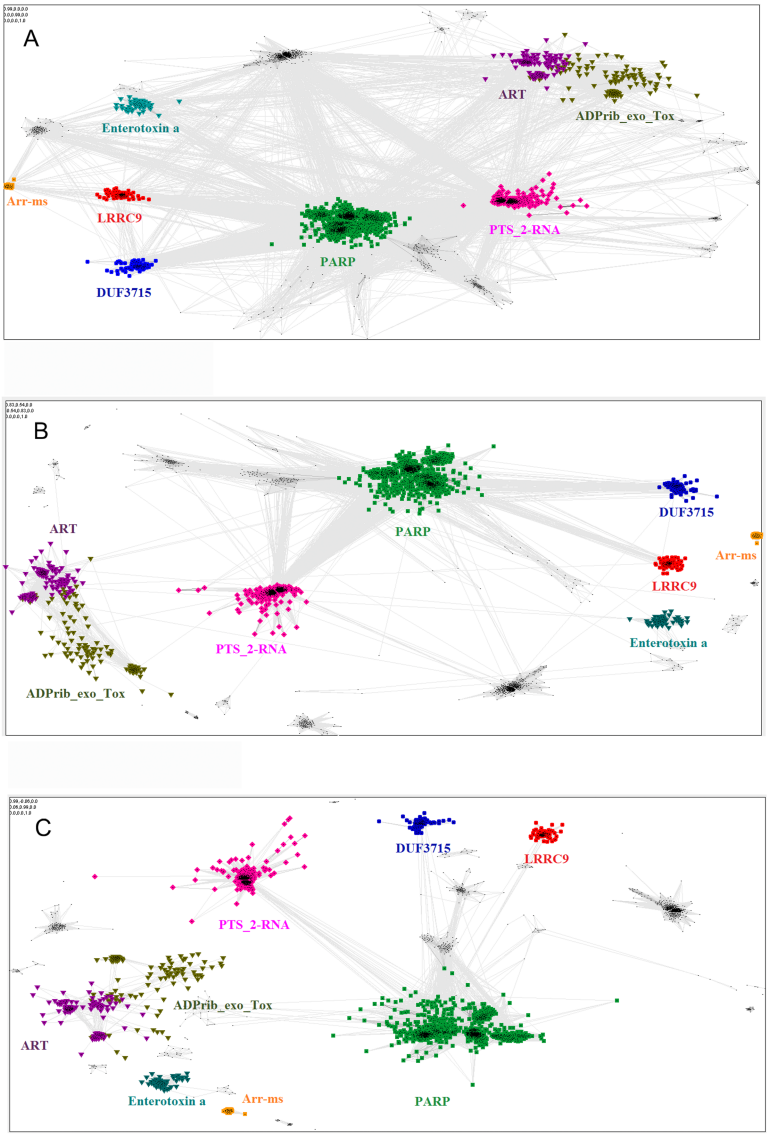


Figure 4 Close and distant sequence similarities between the putative and the known ADP-ribosyltransferase domains, visualized using CLANS algorithm. The graph groups sequences (graph nodes) according to BLAST-derived similarities (edges). (A) Sequence similarities up to BLAST *E*-value of 1, i.e., including distant, non-significant sequence similarities. (B) Up to *E*-value 1e-2. (C) up to *E*-value 1e-4.

Full-size DOI: 10.7717/peerj.11051/fig-4

group of the nicotinamide amide *via* hydrogen bonds (*Cohen & Chang, 2018*). Tyrosine (motif II), localized within β-strand 2 (*Aravind et al., 2015*), interacts with nicotinamide moiety whereas glutamate of β-strand 5 (first position in [QE]x[QED] motif according to Aravind 2018) seems to play a role in maintaining stability of the furanosyl oxocarbenium intermediate (*Cohen & Chang, 2018*).

Multiple sequence alignment comparison allowed us to identify few similarities and substantial differences between sequences of PARP catalytic domains and ART-like

domains, which we identified in human LRRC9 protein sequence in its central part, from position 392 to 625. The tyrosine residue (motif II), one of the catalytic triad elements in PARP, was found to be replaced by a weakly conserved Glu (E525) in human LRRC9-ART. Also, His and Ser residues (from the motif I, His-x-Ser, positions 474–476 in the logo in Fig. 1), highly conserved among PARP enzymatic domain members, were not conserved in LRRC9. The His residue was most often substituted by Tyr. Glutamate, the third element of the catalytic triad (motif III, position 604 in the sequence logo), was replaced by a poorly conserved Lys (Fig. 1). In PARP1, PARP2, PARP10, PARP12 and PARP15 sequences, there is also another important Tyr residue (Tyr932 in human PARP10) that is responsible for nicotinamide stacking (*Karlberg et al., 2015*). Some authors (*Karlberg et al., 2015*) include this second Tyr in the group of conserved residues being the hallmark of PARP ART-domains (*Karlberg et al., 2015*), hence we will use the term "catalytic tetrad HYYE", representing the non-contiguous catalytic motif H-x(n1)-Y-x(n2)-Y-x(n3)-E. This tyrosine is conserved in PARP (at the logo position 538, Fig. 1A) and not conserved in the LRRC9 ART-like domain (Fig. 1B). Lack of most of the catalytic amino acid residues, evolutionarily conserved in PARP catalytic domains, suggests that LRRC9 ART-like domains may be pseudoenzymes. Interestingly, the FFAS03 sequence alignment between human LRRC9 ART-like and PARP10 ART domains, especially in its central part, is not unequivocal. We obtained different alignments using FFAS03 and HHpred servers and even in FFAS results, there are some suboptimal alignment paths to be considered (Figs. S1 and S2). According to the FFAS alignment, PARP region flanked by two tyrosines belonging to the catalytic tetrad HYYE is aligned to a poorly conserved region 525–538 of the LRRC9 ART-like domain (Fig. 1A), but three positions upstream and downstream of this fragment there are two distinct motifs of LRRC9-ART domain: PR[IL] and [DE]xxxFRHG, respectively. This suggests that sequence-based alignments may be inaccurate in this region and one of both above-mentioned LRRC9 conserved motifs may in fact be involved in the active site. Also, the strongly conserved D-(5)-PEY-(2)-EFEY motif in LRRC9 (logo positions 609-623), although not aligned to PARP active site, may be speculated to be functionally important.

Structural analysis of the homology model

Because distant homology detection methods did not offer us consensus for template selection, we decided to create model based on three different approaches: classical one-template homology modeling (Modeller), reassembling structural fragments from threading templates using Monte Carlo simulations (I-Tasser approach, best model with C-score 0.5) and deep learning based methods which use covariance signals obtained from sequence alignments to predict inter-residue distances (trRosetta, best model confidence: 0.70). In the next step, we used DogSiteScorer to analyse the structure models to identify putative ligand binding pockets and compare them to the pockets in known PARP structures (Table S1, Fig. S3) To illustrate uncertainty of structural arrangements of putative ligand-binding residues identified by FFAS03 alignments, we used the homology model (PARP10/3HKV template-based) and the trRosetta (deep learning) model which according to ligand pocket analysis are more likely to accommodate a ligand. Comparison of the

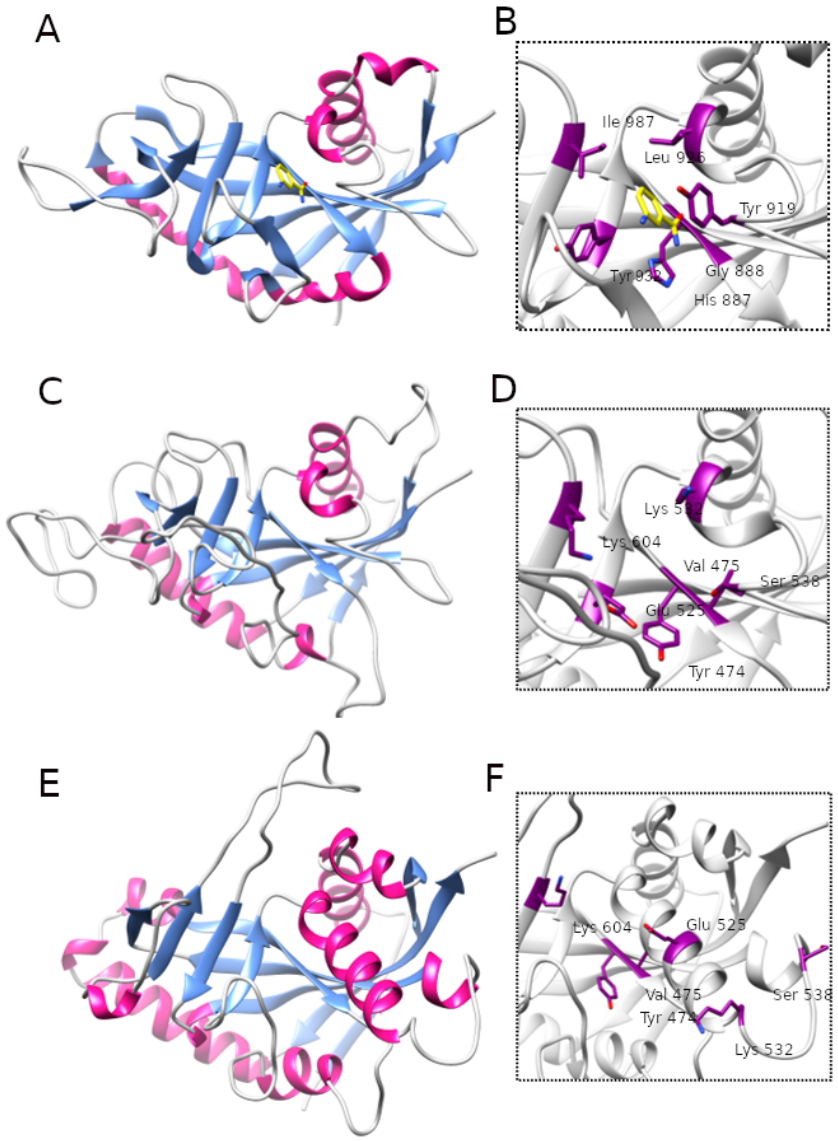


Figure 5 Comparison of ART-like domains and their active site regions. Overall structures of ART domains (A, C, E) and close-up active site regions (B, D, F) of: PARP10, PDB ID: 3HKV (A, B), human LRRC9-ART modelled on the PARP10 template (C, D), and human LRRC9-ART structure modelled using TrRosetta (E, F). Catalytic PARP residues and their counterparts in the modelled structures are depicted as sticks. The ligand shown in panels (A, B) is 3-aminobenzamide, a PARP inhibitor (yellow). Coloring by secondary structures.

Full-size DOI: 10.7717/peerj.11051/fig-5

template 3HKV structure and these two alternative models is presented in Fig. 5. To answer the question if the observed distant sequence similarity between LRRC9-ART and PARP families can translate into functional similarity, we focused first on the known structural determinants of ADP-ribosyltransferase activity (*Karlberg et al., 2015*) and relating these to LRRC9-ART features. In classical PARPs, like PARP1 and PARP2, possessing a poly-ADP-ribosylation (PAR) activity, besides the canonical H-x(n1)-Y-x(n2)-Y-x(n3)-[EI]

tetrad mentioned above, there are two other key positions associated with structural and functional roles of these proteins: glycine residue with nicotinamide anchoring function and an additional catalytic amino acid, lysine 903 (according to PARP1 numbering) (*Karlberg et al., 2015*). In PARP10 and PARP12 with mono-ADP-ribosylation (MAR) activity, instead of this catalytic lysine, there is a leucine or tyrosine residue. Investigating profile-profile alignments for human LRRC9 and PARP10 ART domains one can easily notice that only the first of the canonical tetrad residues is conserved, although substituted (H->Y). Thus, histidine and glycine responsible for nicotinamide anchoring and binding are replaced by tyrosine and valine (Figs. 5B, 5C). In PARPs, the glycine makes two conserved H-bonds to the nicotinamide via main chain carbonyl oxygen and amine hydrogen. These should not be disrupted by valine substitution as the side chain points outward from the active site. Interestingly, in LRRC9-ART at position 532 there is a weakly conserved Lys, as in PARP family members with poly-ADP-ribosylation (PAR) activity. However, instead of a second Tyr (logo position 538) and Glu/Ile residues (604), there are poorly conserved Ser and Lys.

To identify possible ligand pockets in modelled structures we used DoGSiteScorer. For comparison, we also analysed classic PARP structures where ligand binding sites are known and described. We selected PARP1 (ART domain with poly-ADP-ribosylation activity), PARP10 (mono-ADP-ribosylation activity) and two examples of inactive PARP/ART-like domains: PARP13 and TASOR. In case of native structures, DoGSiteScorer correctly identified pockets in the vicinity of PARP hallmark H-Y-Y-E tetrad or its counterpart as the best scoring ones. In case of models, pockets detected in the neighborhood of H-Y-Y-E aligned residues were the best scoring ones for I-Tasser and trRosetta models (Table S1). Overall shape, volume and surface of modelled pockets were very different, but the best score (close to the score obtained for native NAD binding pocket in PARP1) and parameters mostly resembling active PARPs were identified in trRosetta model of LRRC9-ART (Fig. S3). Thus, this model appears to best capture the putative ligand binding site in LRRC9.

Additionally, to investigate possible ligands for the LRRC9-ART domain we used COFACTOR. COFACTOR is a functional annotation tool based on threading of the protein structure query through a database of ligand-protein binding interactions by local and global structure matches to identify functional sites (*Zhang, Freddolino & Zhang, 2017*). COFACTOR suggests the best candidates fitting into the ligand binding site of a given structure. Adenosine monophosphate (AMP) turned out to be among the best scoring ligands for the LRRC9-ART homology model while NAD, a typical ligand for ART domains, was absent from the predicted ligand list. AMP molecule, which differs from NAD by the lack of nicotinamide mononucleotide moiety, is smaller and probably fits better in the modelled LRRC9-ART ligand binding groove. The docked AMP molecule pose was deduced from homology to crystal structure of the catalytic domain of *Pseudomonas aeruginosa* exotoxin A (PDB ID: 1DMA) which catalyzes the transfer of ADP ribose from NAD to elongation factor-2 in eukaryotic cells and hence inhibits protein synthesis. These docking results do not suggest that AMP is a physiological ligand of LRRC9-ART, rather, they provide a likely binding pose of the AMP moiety of NAD. For trRosetta model, COFACTOR found tankyrase-1 complexed with a PARP inhibitor, P4L, as the best ligand

binding site analogue. Here, NAD molecule was not present on the list of ligands complexed with identified binding site structural homologues.

To examine the possibility of NAD binding, we attempted to dock the full NAD molecule into the LRRC9-ART hypothetical active site in the trRosetta (deep learning) structure model. The ligand was constructed in UCSF Chimera build tool and then it was docked using Autodock Vina procedure without specific constraints to the binding groove identified by DoGSiteScorer (Fig. S4B). The docking procedure located NAD in the best scoring pocket in few different poses (Fig. S4A), the best one with a good docking score of −6.8 (Fig. S4C). This suggests we cannot exclude the LRRC9-ART domain, at least as modelled using trRosetta approach, can bind NAD. However, docking a ligand to a structure model built using very distant templates has to be seen as speculative.

Replacing the catalytic Glu residue by Ile is characteristic for PARP family members having mono-ADP-ribosylation (MAR) activity (*Karlberg et al., 2015*). Among FFAS/HHpred hits for LRRC9-ART domain, proteins with MAR, as well as PAR activity and with no confirmed ADP-ribosylation activity appeared. In the sequence logos of ART and LRRC9-ART, there are no common conserved motifs between the two domain families. In LRRC9, the second tyrosine of the HYYE tetrad is replaced by glutamine 525, which is noticeably less conserved. Another highly conserved motif of the PARP family, YxEY[VI][VI][FY], containing catalytic Glu of the HYYE tetrad does not align to a relatively similar motif visible in the LRRC9-ART domain EY[VI][VI]E[FY]EY (logo positions 616-623). As mentioned earlier, these motifs may correspond to each other, and the apparent misalignment may be just an artifact.

Figure S4 shows the trRosetta model of LRRC9-ART structure, with NAD molecule docked in the predicted active site. Figure S4B focuses on the same residues as Fig. 5 but colored according to conservation in the LRRC9-ART family. Remarkably, only tyrosine 474 and valine 475 (out of the residues aligned to the catalytic tetrad) are quite strongly conserved in the whole family. The rest of LRRC9 amino acids aligned to the classical HYYE tetrad residues in the ligand-binding hydrophobic pocket are not highly conserved. This approach of docking a ligand to a structure model built using remote sequence similarity cannot be regarded as authoritative; it is only preliminary and can give a general idea of ligand placement, but no detailed information on its interactions. The model presented in surface representation (Fig. S4C) shows the conserved residues are grouped at the bottom of the putative ligand binding groove.

The ART-like domain described here is only one region of approx. 200 residues within the whole LRRC9 protein molecule, which is 1453 amino acids long, with clear similarity to leucine-rich repeat-containing domains in its N-and C-terminal parts, and a very long helical "spine" in the center. We were tempted to construct the full-length molecule model, based on homology to LRR-containing templates and using the already modelled ART-like domain structure. The results are shown in Fig. S5. In the full-length structure model proposed here, a large groove is surrounded by LRR "horseshoes" from three sides and an ART-like domain from the fourth. The size of the groove seems to suggest a possibility of binding a large molecule, e.g., a protein rather than a small cofactor. It is tempting to speculate that LRRC9, binding some "bait", rearranges its conformation and activates the

enzymatic ART-like domain, thereby triggering a cellular signal. However, the full-length model is only an illustration of possible domain arrangements.

Taxonomic distribution of the LRRC9-like proteins

The LRRC9 protein is widespread in several of the major eukaryotic lineages. It is found in *Alveolata* and *Stramenopiles* (from the TSAR supergroup *Burki et al., 2020*), *Cryptophyta*, *Haptophyta, Viridiplantae, Fungi and Metazoa (Opisthokonta)* and *Euglenozoa* while absent from several other main lineages, e.g., *Hemimastigophora*, CRuMS, *Metamonada*, *Malawimonadida*, *Ancyromonadida*, *Ancoracysta*, *Picozoa* (Fig. 3). Thus, this protein appears to be an ancestral eukaryotic feature, that was subject to losses in a number of lineages.

ART-like domain is present in all vertebrate classes as well as in 11 non-vertebrate animal phyla (*e.g., Cnidaria, Echinodermata, Brachiopoda, Mollusca, Annelida, Hemichordata, Chordata*). However, notable is its absence in several animal phyla, some of which include important model organisms (e.g., *Arthropoda*, *Nematoda*, *Porifera*; see Fig. 3). Also, strikingly, although LRRC9-ART domains are found in some *Viridiplantae*, e.g., mosses, green algae, liverworts, and club mosses, they are absent from flowering plants with the exception of the water lily *Nymphaea colorata*. Notably, water lilies belong to *Magnoliophyta*, a taxon thought to have diverged earliest from the lineage leading to most extant flowering plants (*Zhang et al., 2020*). Altogether, this taxonomic spread suggests an ancient function for LRRC9, common to many diverse eukaryotes.

DISCUSSION

Because the regions of LRRC9-ART domain corresponding to ADP-ribosyltransferase active sites are poorly conserved, the most straightforward conclusion appears to be that LRRC9-ART is a pseudoenzymatic domain. However, as exemplified by the recent cases of two pseudokinases identified by us, SelO and SidJ, apparent pseudoenzymes may in fact perform enzymatic functions somewhat similar to those performed by their "non-pseudo" cousins (*Black et al., 2019*; *Sreelatha et al., 2018*). Also, many inactive pseudoenzymes engage their non-pseudo cousins and modulate their activity, e.g., allosterically (*Adrain, 2020*). Thus, we discuss the significance of the LRR repeats flanking the ART-like domain while speculating that the central domain either performs by itself an ADP-ribosylation-related catalytic function, or modulates such function performed by yet another molecule.

Apart from the ART-like domain, LRRC9 proteins comprise a variable number of leucine-rich repeats (LRRs) flanking the ART domain (Fig. 2). In all the homologs analysed, the LRR-ART-LRR domain architecture is conserved in evolution. The N-terminal region possesses 1-6 LRR repeats while the C-terminal region possesses 11-20 such repeats. The N-terminal repeats and C-terminal repeats are most similar to the LRR repeats from proteins such as Leucine Rich Repeat Containing 23 (LRC23) and TLR4 Interactor with leucine-rich Repeats (TRIL), respectively, however, degree of similarity between LRR motifs from different proteins is generally low and makes it difficult to draw specific functional conclusions. The LRR motifs typically have a length of about 20-30 amino acid residues and are rich in leucines or other aliphatic residues (typically seven per motif,

localized at positions 2, 5, 7, 12, 16, 21 and 24 within the consensus LRR sequence) (*Kobe & Deisenhofer, 1994*). The LRR motifs may form various combinations of two or three secondary structures, usually involving two stretches of a β-strand, e.g., two β-strands and an α-helix. The LRR repeats typically fold into a horseshoe-like conformation composed of helices present on its convex face and parallel β-sheet localized on the concave one (*Enkhbayar et al., 2004*). LRR motifs determine functionality of LRR-containing proteins, enabling them to make protein-protein interactions (*Kobe & Deisenhofer, 1994*) as well as to bind various target structures such as small molecule hormones, lipids and nucleic acids (*Helft et al., 2011*). The non-globular shape of LRR motifs may facilitate the contact between LRR-containing proteins and the target structures, increasing the interaction area (*Kobe & Deisenhofer, 1994*). For example, ribonuclease inhibitor is able to bind pancreatic ribonucleases and inhibit their enzymatic activity. This interaction may affect RNA turnover in the context of angiogenesis (*Kobe & Deisenhofer, 1993*).

Employing the LRR motifs, LRR proteins perform various functions. They may play a role in signal transduction and cell development as well as they are able to participate in RNA splicing and effective response to DNA damage (*Kobe & Deisenhofer, 1995*). Many LRR-containing proteins are adhesive molecules that perform important functions in processes such as regulation of collagen-fibril formation, osteogenesis, myelination and platelet adhesion at a site of vascular injury. Many proteins with LRR motifs play a role in signal transduction as specific receptors. For example, in response to LPS stimulation, LRR-containing CD14 protein induces tyrosine phosphorylation of intracellular proteins to stimulate antibacterial activity of macrophages. Other receptors exhibit specificity to gonadotrophins such as luteinizing hormone, chorionic gonadotrophin and follicle-stimulating hormone (*Bluhm et al., 2004*).

The LRR domains, structurally conserved in evolution, occur widely within plant, invertebrate and vertebrate proteins responsible for innate immunity. Due to their ability to mediate protein-protein interactions, LRR repeats determine signaling function of many pattern recognition receptors (PRRs), such as Toll-like receptors (TLRs) and NOD-like receptors (NLRs), by conditioning their affinity to various ligands, e.g., viral, bacterial, fungal and parasite antigens (*Ng & Xavier, 2011*; *Ng et al., 2011*). On the other hand, ADP-ribosylation is known to be involved in regulation of the host immune response (*Fehr et al., 2020*; *Rosado et al., 2013*). It may direct intracellular signaling to parthanatos (type of programmed death) (*Fatokun, Dawson & Dawson, 2014*; *Greenwald & Pierce, 2019*; *Hoch & Polo, 2019*), whereas such scenario implies induction of inflammatory response to infection. Poly-ADP-ribosylation is known to influence the stability of transcripts encoding proinflammatory cytokines (*Ke et al., 2019*). PARP-dependent chromatin modification may counteract progression of viral infection. ADP-ribosylation also influences other innate immune mechanisms such as NF-kB expression, phagocytosis and macrophage polarization (*Kunze & Hottiger, 2019*). Here, co-occurrence of ART and LRR domains within the LRRC9 protein suggests that they may act together to support a version of host innate immunity. In this context, the LRR domains could perform signal sensors function, and the ART domain might play the effector role.

Overall, in the human proteome there are 234 proteins containing LRR repeats. The molecular functions significantly overrepresented among these proteins include peptide binding (11 proteins, FDR *p*-value 8E-6), ubiquitin-protein transferase activity (11 proteins), heparin binding (7 proteins) and G-protein-coupled peptide receptor activity (7 proteins). However, as many as 182 human LRR proteins are functionally uncharacterized, according to Panther-db. Human proteins containing LRR repeats are involved in processes such as neurogenesis (17 proteins, FDR 1E-5), ubiquitin-dependent protein catabolic process (11 proteins), positive regulation of innate immune response *via* toll-like receptor signaling pathway (10 proteins, FDR 8E-10).

The ART-like domain combined with LRR motifs has precedents. As many as 35 human LRR proteins possess enzymatic domains. Notable here are the NLRP immune sensors/effectors containing NACHT ATP-ase domains, totaling 17 human proteins (*Martinon & Tschopp, 2007*; *Pawlowski et al., 2001*). Other enzymes with LRR motifs include protein kinases (11 proteins), nucleases and peroxidases. Eight human LRR proteins also possess F-box motifs and are involved in ubiquitin ligase complexes.

A bioinformatics study analyzing non-LRR regions embedded within LRR proteins identified the LRRC9 family as containing such a "non-LRR island", noted a few conserved sequence motifs and homologs in a number of non-vertebrate eukaryotes but did not observe similarity to any characterized functional domains (*Matsushima et al., 2009*).

As no high-quality protein expression data is available for LRRC9, it is annotated as "existence validated on transcript level" in the neXtProt (*Zahn-Zabal et al., 2020*) database. According to PAXdb (*Wang et al., 2015*), LRRC9 protein expression is elevated in tonsil, esophagus and seminal vesicle tissues. This may suggest that LRRC9 protein expression is limited to specific circumstances, and is not easily captured in typical tissue proteomics experiments. However, recently, a study aimed at validating the "missing proteins" confirmed the existence of LRRC9 protein in human sperm by applying targeted mass spectrometry and antibody-based methods (*Carapito et al., 2017*). LRRC9 mRNA expression is highest within the brain as well as endocrine and male reproductive tissues, i.e., pituitary gland and testis, respectively (Protein Atlas database). The apparent discrepancy between mRNA and proteomics expression data for LRRC9 may be related to technical difficulties (e.g., too few unique peptides of sufficient length). According to the DeepLoc prediction, the subcellular localization of LRRC9 is cytoplasmic (likelihood 0.79).

For an uncharacterized protein, its physical interactors may shed light on its function. However, according to the BioGRID database, affinity capture-mass spectrometry analysis provided evidence that LRRC9 physically interacts with a single protein, ZFP36L2 (*Noguchi et al., 2018*). ZFP36L2 is an RNA-binding protein regulating cell cycle (*Galloway et al., 2016*) that contributes to pathogenesis of pancreatic ductal adenocarcinoma (PDAC). ZFP36L2 is responsible for an increase in cancer cell aggressiveness (*Yonemori et al., 2017*). A Kaplan–Meier survival analysis (Protein Atlas) shows that LRRC9 mRNA expression is positively correlated with survival among patients with pancreatic cancer ($p = 0.0034$), although the gene is not classified as prognostic. This may suggest an antagonistic relation between LRRC9 and ZFP36L2 activities in the context of ZFP36L2-dependent promotion

of PDAC progression. However, the LRRC9-ZFP36L2 interaction was only reported in a high-throughput experiment, and further investigation is required here.

LRRC9-ART should be classified as a separate ART-like domain family, not only because of the co-occurrence with LRR domains, but primarily because the LRRC9-ART similarity to known ARTs is very remote, and the conservation patterns in LRRC9-ART family are distant from those typical for either PARP or TASOR families.

Pseudoenzymes are characterized by the lack of enzymatic activity despite the common evolutionary origin and structural similarity to catalytically active homologues (*Jeffery, 2020*; *Murphy, Mace & Eyers, 2017*; *Ribeiro et al., 2019*). Pseudoenzymes do occur among PARP family members, including PARP13 that is distinguished by amino acid substitutions and structure changes preventing catalysis. Similarly to LRRC9-ART, PARP13 catalytic domain is characterized by lack of conserved His and Glu residues (*Karlberg et al., 2015*). It has been shown that substitution of catalytic His with Ala within the ART domain of PARP1 substantially decreases its catalytic activity (*Marsischky, Wilson & Collier, 1995*). The results were consistent with previous studies on diphtheria toxin, another member of H-Y-[EDQ] clade. Replacement of catalytic His reduced the toxin's activity by at least 70-fold (*Blanke et al., 1994*). Thus, the LRRC9-ART family, lacking most of the ART active site, can be hypothesized to be pseudoenzymes. However, one cannot exclude the possibility that this ART-like domain has retained or regained the ART catalytic function, or acquired a novel enzymatic activity, in both cases employing an atypical and/or migrated active site. Such a scenario has been recently reported for apparent pseudokinases, SelO and SidJ, that were shown to be AMPylases and polyglutamylases, respectively (*Black et al., 2019*; *Sreelatha et al., 2018*). It can also be envisaged that LRRC9 may require another protein factor to complement its active site, similarly to ADP-ribosyltransferases PARP1 and PARP2 that form complexes with HPF1 which contributes a "missing residue" to the otherwise "incomplete" ART active site (*Suskiewicz et al., 2020*). Another precedent for a "third party" protein required for ART activity is provided by the PARP9/DTX3L heterodimer whereas the otherwise inactive PARP9 needs the interaction partner to become active (*Yang et al., 2017*).

On the other side, in eukaryotic cells, poly-ADP-ribosylating proteins may perform some of their functions without utilizing the enzymatic activity. For example, PARP1 participates in NF-kB-dependent gene transcription *via* two different mechanisms but only one requires poly-ADP-ribosylation whereas the second seems not to involve PARP1 enzymatic function and depends on the structure of the protein (*Hassa et al., 2003*; *Weaver & Yang, 2013*). Both a pseudo-enzymatic character of LRRC9-ART, and migration of the active site are possible for this intriguing novel family. Experimental studies are needed to confirm one of those hypotheses. The unique, conserved domain architecture of LRRC9, suggests that this mysterious protein family could be involved in a defense mechanism, with some analogies to the innate immune system, coupling within a single molecule the detection of foreign objects (LRRs) and downstream signalling (ART domain).

ACKNOWLEDGEMENTS

We thank Drs Vincent Tagliabracci, Miles Black, Anna Muszewska and Marcin Grynberg for critical reading of the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Krzysztof Pawłowski was supported by the Polish National Agency for Scientific Exchange scholarship PPN/BEK/2018/1/00431 and by the Polish National Science Centre grants 2018/31/B/NZ2/00758 and 2019/33/B/NZ2/01409. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The Polish National Agency for Scientific Exchange scholarship PPN/BEK/2018/1/00431.
The Polish National Science Centre grants: 2018/31/B/NZ2/00758, 2019/33/B/NZ2/01409.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Zbigniew Wyżewski performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Marcin Gradowski and Marianna Krysińska performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Małgorzata Dudkiewicz and Krzysztof Pawłowski conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Raw data (including protein sequence alignments, CLANS analysis files, subsignificant alignment plots and PDB file for structure model) are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11051#supplemental-information.

REFERENCES

Adrain C. 2020. Pseudoenzymes: dead enzymes with a lively role in biology. *FEBS Journal* **287**:4102–4105 DOI 10.1111/febs.15535.

Aktories K, Lang AE, Schwan C, Mannherz HG. 2011. Actin as target for modification by bacterial protein toxins. *FEBS Journal* **278**:4526–4543 DOI 10.1111/j.1742-4658.2011.08113.x.

Akturk A, Wasilko DJ, Wu X, Liu Y, Zhang Y, Qiu J, Luo Z-Q, Reiter KH, Brzovic PS, Klevit RE, Mao Y. 2018. Mechanism of phosphoribosyl-ubiquitination mediated by a single Legionella effector. *Nature* **557**:729–733 DOI 10.1038/s41586-018-0147-6.

Aravind L, Zhang D, De Souza RF, Anand S, Iyer LM. 2015. The natural history of ADP-ribosyltransferases and the ADP-ribosylation system. *Current Topics in Microbiology and Immunology* **384**:3–32 DOI 10.1007/82_2014_414.

Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**:3387–3395 DOI 10.1093/bioinformatics/btx431.

Bai P. 2015. Biology of Poly(ADP-Ribose) polymerases: the factotums of cell maintenance. *Molecular Cell* **58**:947–958 DOI 10.1016/j.molcel.2015.01.034.

Barreiro E, Gea J. 2018. PARP-1 and PARP-2 activity in cancer-induced cachexia: potential therapeutic implications. *Biological Chemistry* **399**:179–186 DOI 10.1515/hsz-2017-0158.

Baysarowich J, Koteva K, Hughes DW, Ejim L, Griffiths E, Zhang K, Junop M, Wright GD. 2008. Rifamycin antibiotic resistance by ADP-ribosylation: structure and diversity of Arr. *Proceedings of the National Academy of Sciences of the United States of America* **105**:4886–4891 DOI 10.1073/pnas.0711939105.

Bhogaraju S, Kalayil S, Liu Y, Bonn F, Colby T, Matic I, Dikic I. 2016. Phosphoribosylation of ubiquitin promotes serine ubiquitination and impairs conventional ubiquitination. *Cell* **167**:1636–1649 DOI 10.1016/j.cell.2016.11.019.

Black M, Osinski A, Gradowski M, Servage K, Pawłowski K, Tomchick D, Tagliabracci V. 2019. Bacterial pseudokinase catalyzes protein polyglutamylation to inhibit the SidE all-in-one ubiquitin ligases. *Science* **364**:787–792 DOI 10.1126/science.aaw7446.

Blanke SR, Huang K, Wilson BA, Papini E, Covacci A, Collier RJ. 1994. Active-site mutations of the diphtheria toxin catalytic domain: role of histidine-21 in nicotinamide adenine dinucleotide binding and ADP-ribosylation of elongation factor 2. *Biochemistry* **33**:5155–5161 DOI 10.1021/bi00183a019.

Bluhm AP, Toledo RA, Mesquita FM, Pimenta MT, Fernandes FM, Ribela MT, Lazari MF. 2004. Molecular cloning, sequence analysis and expression of the snake follicle-stimulating hormone receptor. *General and Comparative Endocrinology* **137**:300–311 DOI 10.1016/j.ygcen.2004.03.014.

Boyer L, Doye A, Rolando M, Flatau G, Munro P, Gounon P, Clément R, Pulcini C, Popoff MR, Mettouchi A, Landraud L, Dussurget O, Lemichez E. 2006. Induction of transient macroapertures in endothelial cells through RhoA inhibition by Staphylococcus aureus factors. *Journal of Cell Biology* **173**:809–819 DOI 10.1083/jcb.200509009.

Burki F, Roger AJ, Brown MW, Simpson AGB. 2020. The new tree of eukaryotes. *Trends in Ecology & Evolution* **35**:43–55 DOI 10.1016/j.tree.2019.08.008.

Carapito C, Duek P, Macron C, Seffals M, Rondel K, Delalande F, Lindskog C, Freour T, Vandenbrouck Y, Lane L, Pineau C. 2017. Validating missing proteins in human sperm cells by targeted mass-spectrometry- and antibody-based methods. *Journal of Proteome Research* **16**:4340–4351 DOI 10.1021/acs.jproteome.7b00374.

Chiu CC, Chen HH, Chuang HL, Chung TC, Chen SD, Huang YT. 2009. Pseudomonas aeruginosa exotoxin A-induced hepatotoxicity: an animal model in rats. *Journal of Veterinary Medical Science* **71**:1–8 DOI 10.1292/jvms.71.1.

Choi JR, Shin KS, Choi CY, Kang SJ. 2016. PARP1 regulates the protein stability and proapoptotic function of HIPK2. *Cell Death & Disease* **7**:e2438 DOI 10.1038/cddis.2016.345.

Cohen MS, Chang xx. 2018. Insights into the biogenesis, function, and regulation of ADP-ribosylation. *Nature Chemical Biology* **14**:236–243 DOI 10.1038/nchembio.2568.

Corda D, Di Girolamo M. 2002. Mono-ADP-ribosylation: a tool for modulating immune response and cell signaling. *Science STKE [Signal Transduction Knowledge Environment]* **2002**:pe53 DOI 10.1126/stke.2002.163.pe53.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Research* **14**:1188–1190 DOI 10.1101/gr.849004.

Curtin NJ. 2005. PARP inhibitors for cancer therapy. *Expert Reviews in Molecular Medicine* **7**:1–20 DOI 10.1017/S146239940500904X.

Curtin NJ, Szabo C. 2013. Therapeutic applications of PARP inhibitors: anticancer therapy and beyond. *Molecular Aspects of Medicine* **34**:1217–1256 DOI 10.1016/j.mam.2013.01.006.

De Souza RF, Aravind L. 2012. Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions. *Molecular BioSystems* **8**:1661–1677 DOI 10.1039/c2mb05487.

De Vos M, Schreiber V, Dantzer F. 2012. The diverse roles and clinical relevance of PARPs in DNA damage repair: current state of the art. *Biochemical Pharmacology* **84**:137–146 DOI 10.1016/j.bcp.2012.03.018.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**:W465–W469 DOI 10.1093/nar/gkn180.

Douse CH, Tchasovnikarova IA, Timms RT, Protasio AV, Seczynska M, Prigozhin DM, Albecka A, Wagstaff J, Williamson JC, Freund SMV, Lehner PJ, Modis Y. 2020. TASOR is a pseudo-PARP that directs HUSH complex assembly and epigenetic transposon control. *bioRxiv*. 2020.2003.2009.974832 DOI 10.1101/2020.03.09.974832.

Dudkiewicz M, Lenart A, Pawłowski K. 2013. A novel predicted calcium-regulated kinase family implicated in neurological disorders. *PLOS ONE* **8**:e66427 DOI 10.1371/journal.pone.0066427.

Dudkiewicz M, Pawłowski K. 2019. A novel conserved family of Macro-like domains—-putative new players in ADP-ribosylation signalling. *PeerJ* **7**:e6863 DOI 10.7717/peerj.6863.

Dudkiewicz M, Szczepinska T, Grynberg M, Pawłowski K. 2012. A novel protein kinase-like domain in a selenoprotein, widespread in the tree of life. *PLOS ONE* **7**:e32138 DOI 10.1371/journal.pone.0032138.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**:113 DOI 10.1186/1471-2105-5-113.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* **47**:D427–D432 DOI 10.1093/nar/gky995.

Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N. 2004. Structural principles of leucine-rich repeat (LRR) proteins. *Proteins* **54**:394–403 DOI 10.1002/prot.10605.

Fatokun AA, Dawson VL, Dawson TM. 2014. Parthanatos: mitochondrial-linked mechanisms and therapeutic opportunities. *British Journal of Pharmacology* **171**:2000–2016 DOI 10.1111/bph.12416.

Fehr AR, Singh SA, Kerr CM, Mukai S, Higashi H, Aikawa M. 2020. The impact of PARPs and ADP-ribosylation on inflammation and host-pathogen interactions. *Genes and Development* **34**:341–359 DOI 10.1101/gad.334425.119.

Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**:3702–3704 DOI 10.1093/bioinformatics/bth444.

Galloway A, Saveliev A, Łukasiak S, Hodson DJ, Bolland D, Balmanno K, Ahlfors H, Monzón-Casanova E, Mannurita SC, Bell LS, Andrews S, Díaz-Muñoz MD, Cook SJ, Corcoran A, Turner M. 2016. RNA-binding proteins ZFP36L1 and ZFP36L2 promote cell quiescence. *Science* **352**:453–459 DOI 10.1126/science.aad5978.

Gamble MJ, Fisher RP. 2007. SET and PARP1 remove DEK from chromatin to permit access by the transcription machinery. *Nature Structural & Molecular Biology* **14**:548–555 DOI 10.1038/nsmb1248.

Greenwald SH, Pierce EA. 2019. Parthanatos as a cell death pathway underlying retinal disease. *Advances in Experimental Medicine and Biology* **1185**:323–327 DOI 10.1007/978-3-030-27378-1_53.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods in Molecular Biology* **537**:113–137 DOI 10.1007/978-1-59745-251-9_6.

Gunderson CC, Moore KN. 2015. Olaparib: an oral PARP-1 and PARP-2 inhibitor with promising activity in ovarian cancer. *Future Oncology* **11**:747–757 DOI 10.2217/fon.14.313.

Hassa PO, Buerki C, Lombardi C, Imhof R, Hottiger MO. 2003. Transcriptional coactivation of nuclear factor-kappaB-dependent gene expression by p300 is regulated by poly(ADP)-ribose polymerase-1. *Journal of Biological Chemistry* **278**:45145–45153 DOI 10.1074/jbc.M307957200.

Helft L, Reddy V, Chen X, Koller T, Federici L, Fernández-Recio J, Gupta R, Bent A. 2011. LRR conservation mapping to predict functional sites within protein leucine-rich repeat domains. *PLOS ONE* **6**:e21614 DOI 10.1371/journal.pone.0021614.

Hiranuma N, Park H, Baek M, Anishchanka I, Dauparas J, Baker D. 2020. Improved protein structure refinement guided by deep learning based accuracy estimation. *bioRxiv*.

Hoch NC, Polo LM. 2019. ADP-ribosylation: from molecular mechanisms to human disease. *Genetics and Molecular Biology* **43**:e20190075 DOI 10.1590/1678-4685-gmb-2019-0075.

Huang CC, Meng EC, Morris JH, Pettersen EF, Ferrin TE. 2014. Enhancing UCSF Chimera through web services. *Nucleic Acids Research* **42**:W478–W484 DOI 10.1093/nar/gku377.

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**:680–682 DOI 10.1093/bioinformatics/btq003.

Jankevicius G, Ariza A, Ahel M, Ahel I. 2016. The toxin-antitoxin system DarTG catalyzes reversible ADP-Ribosylation of DNA. *Molecular Cell* **64**:1109–1116 DOI 10.1016/j.molcel.2016.11.014.

Javle M, Curtin NJ. 2011. The role of PARP in DNA repair and its therapeutic exploitation. *British Journal of Cancer* **105**:1114–1122 DOI 10.1038/bjc.2011.382.

Jeffery CJ. 2020. Enzymes, pseudoenzymes, and moonlighting proteins: diversity of function in protein superfamilies. *FEBS Journal* DOI 10.1111/febs.15446.

Jwa M, Chang . 2012. PARP16 is a tail-anchored endoplasmic reticulum protein required for the PERK- and IRE1 α-mediated unfolded protein response. *Nature Cell Biology* **14**:1223–1230 DOI 10.1038/ncb2593.

Kagan JC, Roy CR. 2002. Legionella phagosomes intercept vesicular traffic from endoplasmic reticulum exit sites. *Nat Cell Biol* **4**:945–954 DOI 10.1038/ncb883.

Kalayil S, Bhogaraju S, Bonn F, Shin D, Liu Y, Gan N, Basquin J, Grumati P, Luo ZQ, Dikic I. 2018. Insights into catalysis and function of phosphoribosyl-linked serine ubiquitination. *Nature* **557**:734–738 DOI 10.1038/s41586-018-0145-8.

Kamboj A, Lu P, Cossoy MB, Stobart JL, Dolhun BA, Kauppinen TM, Murcia Gde, Anderson CM. 2013. Poly(ADP-ribose) polymerase 2 contributes to neuroinflammation and neurological dysfunction in mouse experimental autoimmune encephalomyelitis. *Journal of Neuroinflammation* **10**:821 DOI 10.1186/1742-2094-10-49.

Kamel D, Gray C, Walia JS, Kumar V. 2018. PARP inhibitor drugs in the treatment of breast, ovarian, prostate and pancreatic cancers: an update of clinical trials. *Current Drug Targets* **19**:21–37 DOI 10.2174/1389450118666170711151518.

Karlberg T, Klepsch M, Thorsell AG, Andersson CD, Linusson A, Schüler H. 2015. Structural basis for lack of ADP-ribosyltransferase activity in poly(ADP-ribose) polymerase-13/zinc finger antiviral protein. *Journal of Biological Chemistry* **290**:7336–7344 DOI 10.1074/jbc.M114.630160.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**:1160–1166 DOI 10.1093/bib/bbx108.

Ke Y, Wang C, Zhang J, Zhong X, Wang R, Zeng X, Ba X. 2019. The role of PARPs in inflammation-and metabolic-related diseases: molecular mechanisms and beyond. *Cell* **8**:1047 DOI 10.3390/cells8091047.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**:845–858 DOI 10.1038/nprot.2015.053.

Kim DS, Challa S, Jones A, Kraus WL. 2020. PARPs and ADP-ribosylation in RNA biology: from RNA expression and processing to protein translation and proteostasis. *Genes and Development* **34**:302–320 DOI 10.1101/gad.334433.119.

Kim IK, Stegeman RA, Brosey CA, Ellenberger T. 2015. A quantitative assay reveals ligand specificity of the DNA scaffold repair protein XRCC1 and efficient disassembly of complexes of XRCC1 and the poly(ADP-ribose) polymerase 1 by poly(ADP-ribose) glycohydrolase. *Journal of Biological Chemistry* **290**:3775–3783 DOI 10.1074/jbc.M114.624718.

Kim L, Kwon DH, Kim BH, Kim J, Park MR, Park ZY, Song HK. 2018. Structural and biochemical study of the Mono-ADP-Ribosyltransferase Domain of SdeA, a Ubiquitylating/Deubiquitylating Enzyme from Legionella pneumophila. *Journal of Molecular Biology* **430**:2843–2856 DOI 10.1016/j.jmb.2018.05.043.

Klockgether J, Tümmler B. 2017. Recent advances in understanding Pseudomonas aeruginosa as a pathogen. *F1000 Research* **6**:1261 DOI 10.12688/f1000research.10506.1.

Kobe B, Deisenhofer J. 1993. Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature* **366**:751–756 DOI 10.1038/366751a0.

Kobe B, Deisenhofer J. 1994. The leucine-rich repeat: a versatile binding motif. *Trends in Biochemical Sciences* **19**:415–421 DOI 10.1016/0968-0004(94)90090-6.

Kobe B, Deisenhofer J. 1995. Proteins with leucine-rich repeats. *Current Opinion in Structural Biology* **5**:409–416 DOI 10.1016/0959-440x(95)80105-7.

Komander D, Randow F. 2017. Strange new world: bacteria catalyze ubiquitylation via ADP ribosylation. *Cell Host & Microbe* **21**:127–129 DOI 10.1016/j.chom.2017.01.014.

Krishnakumar R, Kraus WL. 2010. PARP-1 regulates chromatin structure and transcription through a KDM5B-dependent pathway. *Molecular Cell* **39**:736–749 DOI 10.1016/j.molcel.2010.08.014.

Kummar S, Chen A, Parchment RE, Kinders RJ, Ji J, Tomaszewski JE, Doroshow JH. 2012. Advances in using PARP inhibitors to treat cancer. *BMC Medicine* **10**:25 DOI 10.1186/1741-7015-10-25.

Kunze FA, Hottiger MO. 2019. Regulating immunity via ADP-Ribosylation: therapeutic implications and beyond. *Trends in Immunology* **40**:159–173 DOI 10.1016/j.it.2018.12.006.

Lawaree E, Jankevicius G, Cooper C, Ahel I, Uphoff S, Tang CM. 2020. DNA ADP-ribosylation stalls replication and is reversed by RecF-mediated homologous recombination and nucleotide excision repair. *Cell Reports* **30**:1373–1384 DOI 10.1016/j.celrep.2020.01.014.

Letunic I, Bork . 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**:W242–W245 DOI 10.1093/nar/gkw290.

Leung A, Todorova T, Ando Y, Chang . 2012. Poly(ADP-ribose) regulates post-transcriptional gene regulation in the cytoplasm. *RNA Biology* **9**:542–548 DOI 10.4161/rna.19899.

Liang YC, Hsu CY, Yao YL, Yang WM. 2013. PARP-2 regulates cell cycle-related genes through histone deacetylation and methylation independently of poly(ADP-ribosyl)ation. *Biochemical and Biophysical Research Communications* **431**:58–64 DOI 10.1016/j.bbrc.2012.12.092.

Liu C, Yu X. 2015. ADP-ribosyltransferases and poly ADP-ribosylation. *Current Protein & Peptide Science* **16**:491–501 DOI 10.2174/1389203716666150504122435.

Livingstone CD, Barton GJ. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer Applications in the Biosciences* **9**:745–756.

Lyons B, Lugo MR, Carlin S, Lidster T, Merrill AR. 2018. Characterization of the catalytic signature of Scabin toxin, a DNA-targeting ADP-ribosyltransferase. *Biochemical Journal* **475**:225–245 DOI 10.1042/bcj20170818.

Malyuchenko NV, Kotova EY, Kulaeva OI, Kirpichnikov MP, Studitskiy VM. 2015. PARP1 inhibitors: antitumor drug design. *Acta Naturae* **7**:27–37 DOI 10.32607/20758251-2015-7-3-27-37.

Mangerich A, Bürkle A. 2012. Pleiotropic cellular functions of PARP1 in longevity and aging: genome maintenance meets inflammation. *Oxidative Medicine and Cellular Longevity* **2012**:321653 DOI 10.1155/2012/321653.

Maresso AW, Deng Q, Pereckas MS, Wakim BT, Barbieri JT. 2007. Pseudomonas aeruginosa ExoS ADP-ribosyltransferase inhibits ERM phosphorylation. *Cellular Microbiology* **9**:97–105 DOI 10.1111/j.1462-5822.2006.00770.x.

Marsischky GT, Wilson BA, Collier RJ. 1995. Role of glutamic acid 988 of human poly-ADP-ribose polymerase in polymer formation, Evidence for active site similarities to the ADP-ribosylating toxins. *Journal of Biological Chemistry* **270**:3247–3254 DOI 10.1074/jbc.270.7.3247.

Martinon F, Tschopp J. 2007. Inflammatory caspases and inflammasomes: master switches of inflammation. *Cell Death and Differentiation* **14**:10–22 DOI 10.1038/sj.cdd.4402038.

Matsushima N, Mikami T, Tanaka T, Miyashita H, Yamada K, Kuroki Y. 2009. Analyses of non-leucine-rich repeat (non-LRR) regions intervening between LRRs in proteins. *Biochimica et Biophysica Acta/General Subjects* **1790**:1217–1237 DOI 10.1016/j.bbagen.2009.06.014.

McCann KE. 2019. Advances in the use of PARP inhibitors for BRCA1/2-associated breast cancer: talazoparib. *Future Oncology* **15**:1707–1715 DOI 10.2217/fon-2018-0751.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**:D183–D189 DOI 10.1093/nar/gkw1138.

Munnur D, Ahel I. 2017. Reversible mono-ADP-ribosylation of DNA breaks. *FEBS Journal* **284**:4002–4016 DOI 10.1111/febs.14297.

Munro P, Benchetrit M, Nahori MA, Stefani C, Clément R, Michiels JF, Landraud L, Dussurget O, Lemichez E. 2010. The Staphylococcus aureus epidermal cell differentiation inhibitor toxin promotes formation of infection foci in a mouse model of bacteremia. *Infection and Immunity* **78**:3404–3411 DOI 10.1128/iai.00319-10.

Murphy JM, Mace PD, Eyers PA. 2017. Live and let die: insights into pseudoenzyme mechanisms from structure. *Current Opinion in Structural Biology* **47**:95–104 DOI 10.1016/j.sbi.2017.07.004.

Ng AC, Eisenberg JM, Heath RJ, Huett A, Robinson CM, Nau GJ, Xavier RJ. 2011. Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. *Proceedings of the National Academy of Sciences of the United States of America* **108(Suppl 1)**:4631–4638 DOI 10.1073/pnas.1000093107.

Ng A, Xavier RJ. 2011. Leucine-rich repeat (LRR) proteins: integrators of pattern recognition and signaling in immunity. *Autophagy* **7**:1082–1084 DOI 10.4161/auto.7.9.16464.

Noguchi A, Adachi S, Yokota N, Hatta T, Natsume T, Kawahara H. 2018. ZFP36L2 is a cell cycle-regulated CCCH protein necessary for DNA lesion-induced S-phase arrest. *Biology Open* **7**:bio031575 DOI 10.1242/bio.031575.

Offord V, Coffey TJ, Werling D. 2010. LRRfinder: a web application for the identification of leucine-rich repeats and an integrative Toll-like receptor database. *Developmental and Comparative Immunology* **34**:1035–1041 DOI 10.1016/j.dci.2010.05.004.

O'Sullivan J, Ferreira MTedim, Gagne JP, Sharma AK, Hendzel MJ, Masson JY, Poirier GG. 2019. Emerging roles of eraser enzymes in the dynamic control of protein ADP-ribosylation. *Nature Communications* **10**:1182 DOI 10.1038/s41467-019-08859-x.

Otto H, Reche PA, Bazan F, Dittmar K, Haag F, Koch-Nolte F. 2005. In silico characterization of the family of PARP-like poly(ADP-ribosyl)transferases (pARTs). *BMC Genomics* **6**:139 DOI 10.1186/1471-2164-6-139.

Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, Zhang F, Dolma S, Willems A, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M. 2019. The BioGRID interaction database: 2019 update. *Nucleic Acids Research* **47**:D529–D541 DOI 10.1093/nar/gky1079.

Pawłowski K, Lepisto M, Meinander N, Sivars U, Varga M, Wieslander E. 2006. Novel conserved hydrolase domain in the CLCA family of alleged calcium-activated chloride channels. *Proteins-Structure Function and Bioinformatics* **63**:424–439 DOI 10.1002/prot.20887.

Pawlowski K, Pio F, Chu Z, Reed JC, Godzik A. 2001. PAAD - a new protein domain associated with apoptosis, cancer and autoimmune diseases. *Trends in Biochemical Sciences* **26**:85–87 DOI 10.1016/S0968-0004(00)01729-1.

Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X. 2009. DOG 1.0: illustrator of protein domain structures. *Cell Research* **19**:271–273 DOI 10.1038/cr.2009.6.

Ribeiro AJM, Das S, Dawson N, Zaru R, Orchard S, Thornton JM, Orengo C, Zeqiraj E, Murphy JM, Eyers PA. 2019. Emerging concepts in pseudoenzyme classification, evolution, and signaling. *Science Signaling* **12**:eaat9797 DOI 10.1126/scisignal.aat9797.

Riccio AA, Cingolani G, Pascal JM. 2016. PARP-2 domain requirements for DNA damage-dependent activation and localization to sites of DNA damage. *Nucleic Acids Research* **44**:1691–1702 DOI 10.1093/nar/gkv1376.

Rosado MM, Bennici E, Novelli F, Pioli C. 2013. Beyond DNA repair, the immunological role of PARP-1 and its siblings. *Immunology* **139**:428–437 DOI 10.1111/imm.12099.

Rossi ML, Ghosh AK, Bohr VA. 2010. Roles of Werner syndrome protein in protection of genome integrity. *DNA Repair* **9**:331–344 DOI 10.1016/j.dnarep.2009.12.011.

Roy A, Zhang Y. 2012. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* **20**:987–997 DOI 10.1016/j.str.2012.03.009.

Saikatendu KS, Joseph JS, Subramanian V, Clayton T, Griffith M, Moy K, Velasquez J, Neuman BW, Buchmeier MJ, Stevens RC, Kuhn . 2005. Structural basis of severe acute respiratory syndrome coronavirus ADP-ribose-1-phosphate dephosphorylation by a conserved domain of nsP3. *Structure* **13**:1665–1675 DOI 10.1016/j.str.2005.07.022.

Samaras P, Schmidt T, Frejno M, Gessulat S, Reinecke M, Jarzab A, Zecha J, Mergner J, Giansanti P, Ehrlich HC, Aiche S, Rank J, Kienegger H, Krcmar H, Kuster B, Wilhelm M. 2020. ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Research* **48**:D1153–D1163 DOI 10.1093/nar/gkz974.

Simon NC, Aktories K, Barbieri JT. 2014. Novel bacterial ADP-ribosylating toxins: structure and function. *Nature Reviews. Microbiology* **12**:599–611 DOI 10.1038/nrmicro3310.

Sreelatha A, Yee SS, Lopez V, Park B, Pilch S, Servage KA, Zhang J, Jiou J, Kinch LN, Karasiewicz M, Łobocka M, Grishin NV, Orth K, Kucharczyk R, Pawłowski K, Tomchick DR, Tagliabracci VS. 2018. Protein AMPylation by an evolutionarily conserved pseudokinase. *Cell* **175**:809–821 DOI 10.1016/j.cell.2018.08.046.

Suskiewicz MJ, Zobel F, Ogden TEH, Fontana P, Ariza A, Yang JC, Zhu K, Bracken L, Hawthorne WJ, Ahel D, Neuhaus D, Ahel I. 2020. HPF1 completes the PARP active site for DNA damage-induced ADP-ribosylation. *Nature* **579**:598–602 DOI 10.1038/s41586-020-2013-6.

Szántó M, Brunyánszki A, Márton J, Vámosi G, Nagy L, Fodor T, Kiss B, Virág L, Gergely P, Bai P. 2014. Deletion of PARP-2 induces hepatic cholesterol accumulation and decrease in HDL levels. *Biochimica et Biophysica Acta/General Subjects* **1842**:594–602 DOI 10.1016/j.bbadis.2013.12.006.

Tchasovnikarova IA, Timms RT, Matheson NJ, Wals K, Antrobus R, Göttgens B, Dougan G, Dawson MA, Lehner PJ. 2015. GENE SILENCING, Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. *Science* **348**:1481–1485 DOI 10.1126/science.aaa7227.

Teloni F, Altmeyer M. 2016. Readers of poly(ADP-ribose): designed to be fit for purpose. *Nucleic Acids Research* **44**:993–1006 DOI 10.1093/nar/gkv1383.

Thul PJ, Lindskog C. 2018. The human protein atlas: a spatial map of the human proteome. *Protein Science* **27**:233–244 DOI 10.1002/pro.3307.

Trott O, Olson AJ. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **31**:455–461 DOI 10.1002/jcc.21334.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, Feilitzen Kvon, Forsberg M, Persson L, Johansson F, Zwahlen M, Heijne Gvon, Nielsen J, Ponten F. 2015. Proteomics. Tissue-based map of the human proteome. *Science* **347**:1260419 DOI 10.1126/science.1260419.

Volkamer A, Kuhn D, Grombacher T, Rippmann F, Rarey M. 2012. Combining global and local measures for structure-based druggability predictions. *Journal of Chemical Information and Modeling* **52**:360–372 DOI 10.1021/ci200454v.

Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, Von Mering C. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**:3163–3168 DOI 10.1002/pmic.201400441.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**:1189–1191 DOI 10.1093/bioinformatics/btp033.

Weaver AN, Yang ES. 2013. Beyond DNA repair: additional functions of PARP-1 in cancer. *Frontiers in Oncology* **3**:290 DOI 10.3389/fonc.2013.00290.

Webb B, Sali A. 2017. Protein structure modeling with MODELLER. *Methods in Molecular Biology* **1654**:39–54 DOI 10.1007/978-1-4939-7231-9_4.

Xu D, Jaroszewski L, Li Z, Godzik A. 2014. FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* **30**:660–667.

Xu D, Jaroszewski L, Li Z, Godzik A. 2015. AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. *Bioinformatics* **31**:2098–2105 DOI 10.1093/bioinformatics/btv092.

Yamamoto M, Yamasaki M, Tsukao Y, Tanaka K, Miyazaki Y, Makino T, Takahashi T, Kurokawa Y, Nakajima K, Takiguchi S, Mori M, Doki Y. 2017. Poly (ADP-ribose) polymerase-1 inhibition decreases proliferation through G2/M arrest in esophageal squamous cell carcinoma. *Oncology Letters* **14**:1581–1587 DOI 10.3892/ol.2017.6334.

Yang CS, Jividen K, Spencer A, Dworak N, Ni L, Oostdyk LT, Chatterjee M, Kuśmider B, Reon B, Parlak M, Gorbunova V, Abbas T, Jeffery E, Sherman NE, Paschal BM. 2017. Ubiquitin modification by the E3 Ligase/ADP-Ribosyltransferase Dtx3L/Parp9. *Molecular Cell* **66**:503–516 DOI 10.1016/j.molcel.2017.04.028.

Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proceedings*

of the National Academy of Sciences of the United States of America **117**:1496–1503 DOI 10.1073/pnas.1914677117.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12**:7–8 DOI 10.1038/nmeth.3213.

Yonemori K, Seki N, Kurahara H, Osako Y, Idichi T, Arai T, Koshizuka K, Kita Y, Maemura K, Natsugoe S. 2017. ZFP36L2 promotes cancer cell aggressiveness and is regulated by antitumor microRNA-375 in pancreatic ductal adenocarcinoma. *Cancer Science* **108**:124–135 DOI 10.1111/cas.13119.

Zahn-Zabal M, Michel PA, Gateau A, Nikitin F, Schaeffer M, Audot E, Gaudet P, Duek PD, Teixeira D, Rechde Laval V, Samarasinghe K, Bairoch A, Lane L. 2020. The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Research* **48**:D328–D334 DOI 10.1093/nar/gkz995.

Zhang C, Freddolino PL, Zhang Y. 2017. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Research* **45**:W291–W299 DOI 10.1093/nar/gkx366.

Zhang L, Chen F, Zhang X, Li Z, Zhao Y, Lohaus R, Chang X, Dong W, Ho SYW, Liu X, Song A, Chen J, Guo W, Wang Z, Zhuang Y, Wang H, Chen X, Hu J, Liu Y, Qin Y, Wang K, Dong S, Liu Y, Zhang S, Yu X, Wu Q, Wang L, Yan X, Jiao Y, Kong H, Zhou X, Yu C, Chen Y, Li F, Wang J, Chen W, Chen X, Jia Q, Zhang C, Jiang Y, Zhang W, Liu G, Fu J, Chen F, Ma H, Van de Peer Y, Tang H. 2020. The water lily genome and the early evolution of flowering plants. *Nature* **577**:79–84 DOI 10.1038/s41586-019-1852-5.

Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Soding J, Lupas AN, Alva V. 2017. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *Journal of Molecular Biology* DOI 10.1016/j.jmb.2017.12.007.

Warszawa, 29.05.2025 r.

dr Zbigniew Wyżewski
z.wyzewski@uksw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Wyżewski Z, Gradowski M, Krysińska M, Dudkiewicz M, Pawłowski K. 2021. A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution. PeerJ. 9:e11051** mój indywidualny udział w jej powstaniu polegał na przeprowadzeniu części eksperymentów dotyczących analizy sekwencji białka, napisaniu i zrecenzowaniu roboczych wersji manuskryptu oraz zatwierdzeniu jego ostatecznej wersji.

Zbigniew Wyżewski

Podpis

Warszawa, 1.09.2025 r.

dr Marcin Gradowski
marcin_gradowski@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Wyżewski Z, Gradowski M, Krysińska M, Dudkiewicz M, Pawłowski K. 2021. A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution. PeerJ. 9:e11051**. mój indywidualny udział w jej powstaniu polegał na przeprowadzeniu części eksperymentów dotyczących analizy sekwencji białka, przeanalizowaniu uzyskanych danych, przygotowaniu rysunków oraz zatwierdzeniu ostatecznej wersji projektu.

Podpis

Warszawa, 1.09.2025 r.

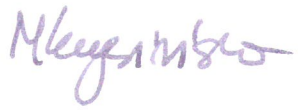
Marianna Krysińska
marianna_krysinska@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Wyżewski Z, Gradowski M, Krysińska M, Dudkiewicz M, Pawłowski K. 2021. A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution. PeerJ. 9:e11051.** mój indywidualny udział w jej powstaniu polegał na przeprowadzeniu części eksperymentów – w tym dotyczących analizy budowy domenowej białek LRRC9, analizy podobieństw strukturalnych, analiz filogenetycznych, zaprojektowaniu i wykonaniu rysunków i tabel do głównej części artykułu oraz zatwierdzeniu ostatecznej wersji manuskryptu.



Podpis

Warszawa, 29.07.2025 r.

dr hab. Małgorzata Dudkiewicz
małgorzata_dudkiewicz@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Wyżewski Z, Gradowski M, Krysińska M, Dudkiewicz M, Pawłowski K. 2021. A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution. PeerJ. 9:e11051** mój indywidualny udział w jej powstaniu polegał na zaprojektowaniu eksperymentów bioinformatycznych, wykonaniu badań dotyczących analiz sekwencyjnych, filogenetycznych i strukturalnych, analizie wyników, stworzeniu rysunków (Figure 1, Figure 3, Figuren 5), przygotowaniu tekstu artykułu w części dotyczącej dyskusji wyników, podsumowań i wniosków, dyskusji z recenzentami i zatwierdzeniu ostatecznej wersji artykułu.

Podpis

Dallas, 30.10.2025

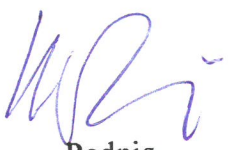
dr hab. Krzysztof Pawłowski
Krzysztof.Pawlowski@utsouthwestern.edu

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Wyżewski Z, Gradowski M, Krysińska M, Dudkiewicz M, Pawłowski K. 2021. A novel predicted ADP-ribosyltransferase-like family conserved in eukaryotic evolution. PeerJ. 9:e11051** mój indywidualny udział w jej powstaniu polegał na zaprojektowaniu eksperymentów, wykonaniu części badań dotyczących poszukiwania homologów białka LRRC9 i analizie uzyskanych danych, przygotowaniu rysunków dotyczących zmienności sekwencji, recenzowaniu i zatwierdzeniu ostatecznej wersji artykułu.



Podpis

scientific reports

Check for updates

OPEN

Pan-kinome of *Legionella* expanded by a bioinformatics survey

Marianna Krysińska[1], Bartosz Baranowski[2], Bartłomiej Deszcz[1], Krzysztof Pawłowski[1,3,4,5]✉ & Marcin Gradowski[1]✉

The pathogenic *Legionella* bacteria are notorious for delivering numerous effector proteins into the host cell with the aim of disturbing and hijacking cellular processes for their benefit. Despite intensive studies, many effectors remain uncharacterized. Motivated by the richness of *Legionella* effector repertoires and their oftentimes atypical biochemistry, also by several known atypical *Legionella* effector kinases and pseudokinases discovered recently, we undertook an in silico survey and exploration of the pan-kinome of the *Legionella* genus, i.e., the union of the kinomes of individual species. In this study, we discovered 13 novel (pseudo)kinase families (all are potential effectors) with the use of non-standard bioinformatic approaches. Together with 16 known families, we present a catalog of effector and non-effector protein kinase-like families within *Legionella*, available at http://bioinfo.sggw.edu.pl/kintaro/. We analyze and discuss the likely functional roles of the novel predicted kinases. Notably, some of the kinase families are also present in other bacterial taxa, including other pathogens, often phylogenetically very distant from *Legionella*. This work highlights Nature's ingeniousness in the pathogen–host arms race and offers a useful resource for the study of infection mechanisms.

The *Legionella* genus includes close to 70 species of mostly pathogenic Gram-negative bacteria[1,2]. The *Legionella* strains use several secretion systems to translocate effectors into the host cell[3,4]. Thus, these bacteria can modulate host cell signaling and metabolic processes to establish a favorable replicating environment within the host cell known as the *Legionella* Containing Vacuole (LCV). The best-known species of this genus is the human pathogen *Legionella pneumophila*. It is responsible for 80–90% of infection cases caused by all the *Legionella* species[4]. *L. pneumophila* and other *Legionella* species use up to 330 effectors[5]. Usually, *Legionella* bacteria live in natural water reservoirs although some of them are isolated from non-aquatic habitats[1,4]. In water, the bacterium infects a wide range of free-living amoeba which are the natural hosts. It can also survive in the artificial environment of human-made water systems. For *L. pneumophila*, the most frequent path of transmission to humans is through inhalation or microaspiration of water contaminated with the bacteria. Thus, the bacterium can reach human lungs and infect alveolar macrophages. This results in diseases such as lethal, nonspecific pneumonia (called Legionnaires' disease) or milder flu-like Pontiac fever[4,6]. Out of the at least 69 known *Legionella* species, about 25 are associated with human infections[1].

The *Legionella* effector proteins can affect diverse cellular processes such as cell cytoskeleton rearrangement, cell adhesion, signaling, transcription, apoptosis or metabolic processes[7]. Although a large proportion of these effectors are functionally uncharacterized, many were shown to be enzymes, e.g., kinases, proteases, phosphatases[7,8].

Many effectors do not act individually, rather, they functionally interact once inside the host cell. For instance, the SidM effector covalently adds an adenosine monophosphate (AMP) moiety to human Rab1 protein. Next, AMP can be removed by the SidD effector, thus antagonizing the SidM effect. Many such pairs of effectors, termed metaeffectors, have been described[9].

As protein kinases are among the basic enzymes that regulate most of the cellular processes, bacteria developed effector kinases which manipulate many processes in the cell[7,9]. Here, we focus on the Protein kinase-like superfamily (Pfam clan: CL0016) which combines protein families that share a common structure—Protein

[1]Department of Biochemistry and Microbiology, Warsaw University of Life Sciences — SGGW, Warsaw, Poland. [2]Laboratory of Plant Pathogenesis, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland. [3]Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [4]Department of Translational Medicine, Lund University, Lund, Sweden. [5]Howard Hughes Medical Institute, Dallas, TX, USA. ✉email: krzysztof.pawlowski@utsouthwestern.edu; marcin_gradowski@sggw.edu.pl

Kinase-Like fold (PKL)[10,11]. For example, *E. coli* NleH1/2 and *Salmonella* OspG effector kinases modulate the human host immune response by inhibition of the host NF-κB pathway[7]. Also, it was recently discovered that they target the microvillus protein Eps8 responsible for actin bundling. This causes a change in the structure of enterocytes and leads to diarrhea in children[12]. The recently discovered HopBF1 kinase from the plant pathogen *Pseudomonas syringae* is recognized by host HSP90 as a client. HSP90 is then phosphorylated by HopBF1 to completely inhibit the chaperone's ATPase activity. This dampens the plant's immune response[13].

Legionella has a considerable repertoire of characterized effector kinases, including eukaryotic-like protein kinases LegK1[7], LegK2[14], LegK3[7], LegK4[15], LegK7[16] as well as phosphatidylinositol (PI) kinases—LepB[17], AnkK[18] and MavQ[19].

LegK1 is considered to work similarly to NleH/OspG, by affecting the host NF-κB pathway. Thus, LegK1 activates the noncanonical NF-κB pathway through phosphorylation of NF-kappa-B p100 subunit, which prevents host cell apoptosis[7]. LegK2 targets the ARP2/3 complex to inhibit actin polymerization on the phagosome, thereby blocking phagosome/endosome fusion and helping to remodel phagosome into LCV[14]. LegK4 phosphorylates host Hsp70 to reduce the chaperone's ability to refold proteins which causes inhibition of cellular protein translation[15]. LegK7 functionally mimics host Hippo kinase by activating the MOB1A protein which supports bacterial growth[16]. MavQ, LepB and AnkK are kinases that phosphorylate Phosphatidylinositol (PI) or its various derivatives on the LCV, thus assuring its proper PI-based "decoration" and contributing to the evasion of the host cell degradation pathway[17–19]. Thus, *Legionella* uses a wide range of PKL proteins that hijack host signaling and metabolic pathways, which facilitates bacterial infection.

Besides effector kinases, *Legionella* has a large set of non-effector kinases (see Suppl. Table S1), including the ancient ADCK–UbiB2–ABC1 family (lpg2905 in *L. pneumophila*) involved in synthesis of ubiquinone (cofactor Q) in bacteria[20]. The well-known HipA kinase (lpg1934 in *L. pneumophila*, see also "Results and discussion" section) promotes multidrug tolerance by blocking translation, inhibition of growth, and induction of persistence[21]. Other kinases of small molecules phosphorylate antibiotics to block their actions[21,22].

Motivated by the richness of *Legionella* effector repertoires and their oftentimes atypical biochemistry, also by several atypical *Legionella* effector kinases and pseudokinases discovered by us and by others (MavQ[19], lpg2603[23], SidJ[24], AnkK[18], LepB[17]), we undertook an in silico survey and exploration of the pan-kinome of the *Legionella* genus.

In this study, we discovered 13 novel families (all are potential effector kinases; see Suppl. Table S4) with the use of non-standard bioinformatic approaches (Fig. 1, Suppl. Fig. S8). Together with 16 known families (representing 99 *Legionella* orthologous groups—LOGs[25]), we present a catalog of effector and non-effector *Legionella* PKL families, available at http://bioinfo.sggw.edu.pl/kintaro/. For the novel families, we focus on predicting their function, establish evolutionary history, and occurrence across the bacterial world.

Results and discussion

Charting kinases in the *Legionella* pan-proteome. The survey started from the *Legionella* pan-proteome with 16,416 orthologous groups of proteins from 41 species[25]. After clustering at 90% and 50% sequence identity thresholds[31] and splitting them into fragments[32] (see "Materials and methods" section), 21,616 sequences were analyzed by FFAS algorithm for distant similarity to kinases (Suppl. Table S2)[33]. Among the FFAS hits, 16 known protein kinase-like families were recognized by RPS-BLAST[34,35] and from the literature (Suppl. Table S1). Thirteen FFAS kinase-like hits were not automatically recognized as such and were validated by other remote sequence similarity search methods (HHpred/HHsearch) (Suppl. Table S2)[36], Phyre2 (Suppl. Table S2)[37], analysis of sequence logos[27] with secondary structure (Suppl. Fig. S8)[38] and de novo structure modeling using the RoseTTAFold[39] and AlphaFold2[40] methods supplemented with structural comparisons (FATCAT[41] and Dali[42] servers) (Suppl. Tables S3, S9). For most of the modeled structures of the novel kinases, significant similarity to known protein kinase structures was found (see prediction summary in Suppl. Table S3). For Lmor_1975, LLO_2159, and Lsai_0337, the similarity of structure models to known kinases was partial or weak (Suppl. Table S3).

The *Legionella* species differ greatly in numbers of kinase-like families, ranging from 8 to 43. This kind of diversity among effector and non-effector repertoires is believed to result from the adaptations to infecting different hosts (e.g., different amoeba species)[25,43]. For every species, effectors form the majority of the kinome (Fig. 2).

Interestingly, some of the novel families have many hundreds of homologs outside the *Legionellaceae* family, while some are restricted to *Legionellaceae* or even a subset thereof (Fig. 3). The two families with largest numbers of homologs (Lani_1194 and Lcin_0519) are discussed in detail in a later section. Among the 112 kinase families, there are predicted effectors and non-effector kinase families (Fig. 2, Suppl. Table S4)[44–46]. In almost every species analyzed, effector kinases constitute the majority of kinome, e.g., 17 out of 25 in *L. pneumophila*.

Almost all novel families have well-preserved key kinase residues (see Fig. 1, Suppl. Fig. S8). Only in Lani_1194 the catalytic lysine K72 (PKA) is replaced by R. In Lcin_0519 and Lani_1194, the equivalent of E91 cannot be identified by sequence and structure analysis. Only in two pseudokinase cases, Llan_0165 and Lspi_2187, the catalytic aspartate D166 (PKA) is not conserved, while N171 and D184 (PKA) are conserved in all novel families.

Sequence similarity analysis of novel kinase families. The CLANS graph analysis (Fig. 4) allows the investigation of sequence similarity relationships between the 13 novel *Legionella* kinase families and 49 known kinase families from all the domains of life (see "Materials and methods" section)[47]. This graph may indicate distant relationships between families, which are important for understanding their evolution and functionality. The CLANS graph represents quasi-distances between sequences, based on multiple pairwise alignments built by all-to-all the BLAST sequence comparisons. This approach is used consciously, because classical phyloge-

Figure 1. Most *Legionella* kinase-like families have conserved active site motifs. Sequence logos of active site motifs for selected families. Also, the "classic" kinases (ePK) shown. Residue numbering (top row) according to standard protein kinase A (PKA) nomenclature[26]. Asterisks denote the novel *Legionella* kinases. Double asterisks denote the novel families discussed in detail. DB: source database of kinase sequences used for the logos[27]. N indicates numbers of homologous sequences from BLAST search (E = 1e−4 threshold)[28–30]. In brackets—numbers of homologous sequences after CD-HIT clustering at the level of 99% sequence identity[31] (see "Materials and methods" section). For some families, it was not possible to identify the residue corresponding to E91 of PKA.

netics analysis would require an unambiguous multiple sequence alignment. Achieving such an alignment of diverse and very distant families is difficult due to the presence of family-specific regions. Even structure-based alignments suffer from this problem in diverse superfamilies. The analysis shows that the novel families generally do not cluster by connectivity and proximity with established, well-studied ones. An exception are four novel families clustering with the FAM20/CotH group (Lmor_1975; LLO_1015; LLO_2159; Lsai _0337), together with known *Legionella* kinases LepB and AnkK, which suggests they may be phosphorylating derivatives of phosphatidylinositol or other lipids. However, most novel families do not cluster with PKL families of known functions, e.g., protein kinases and lipid kinases. Phylogenetic trees built for three Protein Kinase-Like groups from *Legionella* and their selected eukaryotic counterparts using structure-based sequence alignments, do support the bacterial origin of the HipA-like *Legionella* kinases, and the phosphatidylinositol kinase-like proteins while supporting the likely eukaryotic origin of the known LegK1–4 kinases (see Suppl. Fig. S15).

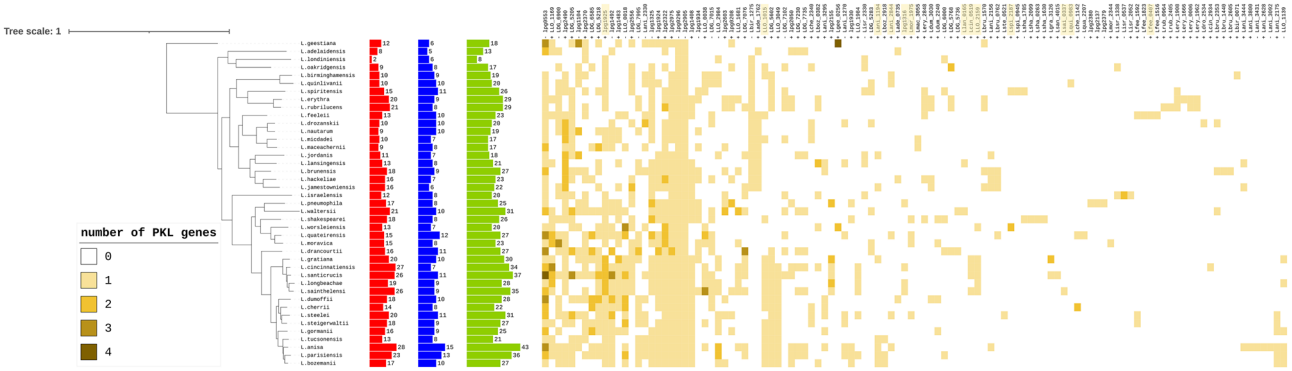


Figure 2. Distribution of 112 kinase LOGs among 41 *Legionella* species. Histograms on the left: counts of effector kinases shown in red, non-effectors in blue; percentage of effector kinases in a kinome shown in green. Clustered heatmap depicts the numbers of each LOG representative per species (range 0–4). Gene labels marked with plus (+) indicate effector families/LOGs. Novel families are marked by highlighted gene labels. Phylogenetic tree of *Legionella* species adapted from the publication by Burstein et al.[25,43].

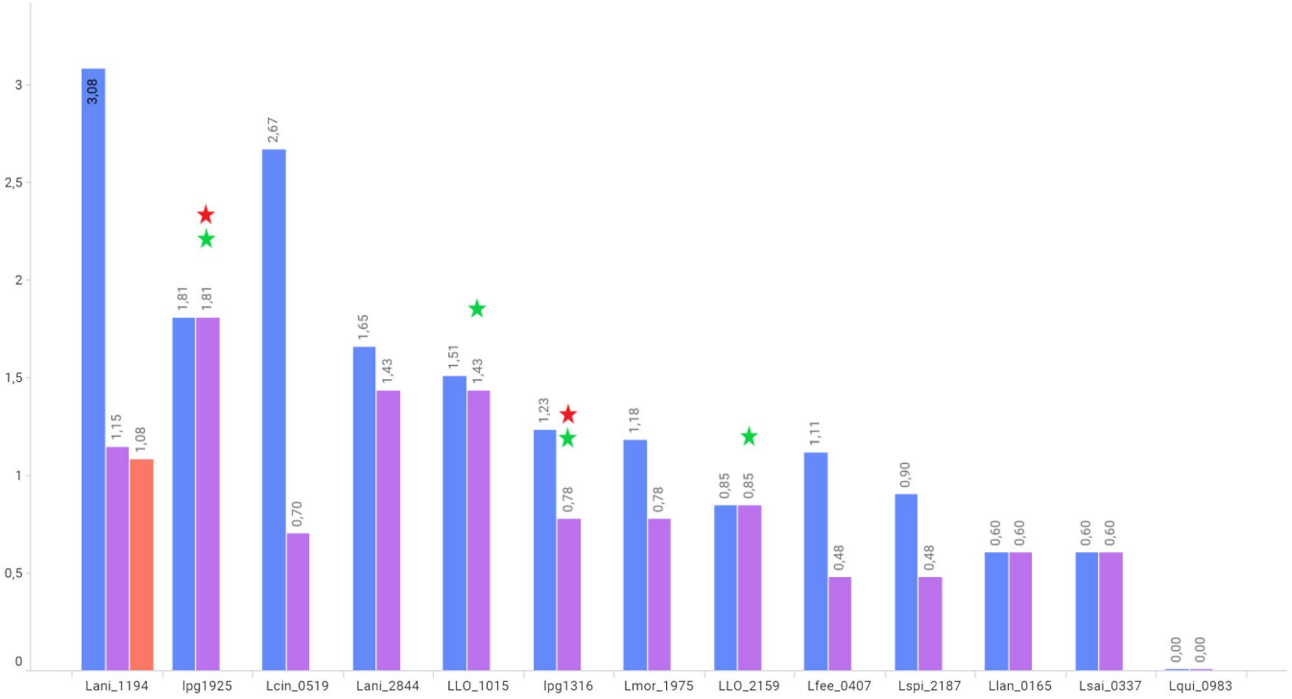


Figure 3. Numbers of species with homologs of novel *Legionella* kinase families. Numbers of species with homologs shown in logarithmic scale. The sequences were collected by BLAST search in the NR database (BLAST at E = 1e−4)[28–30]. Blue columns — *Bacteria*, purple — *Legionellaceae*, red — *Archaea*. Red asterisk — homologs present in *Legionella pneumophila subsp. pneumophila* str. Philadelphia 1, green asterisk — homologs in *Legionella longbeachae* NSW150.

Sequence similarity network suggests possible horizontal gene transfer events. A CLANS sequence similarity network including all kinases from 41 *Legionella* species and full kinomes of the hosts: human and amoebas *Dictyostelium discoideum* and *Acanthamoeba castellanii* can be used for a tentative overview of evolutionary relationships (Fig. 5, Suppl. 11)[47]. The CLANS graph should be treated as an inaccurate representation of the relationship network, where the complex multidimensional network of similarities is captured on a two-dimensional graph where similar protein kinases form clusters. The center of the graph is occupied by eukaryotic and eukaryotic-like kinases (ePK and ELK). Some *Legionella* kinases (e.g., LegK1–4) are found within and nearby this central cluster. This may indicate a horizontal gene transfer whereby eukaryotic host kinases could have been acquired by the bacteria. In contrast, some *Legionella* kinases are clearly separated from eukaryotic ones in the graph (e.g., Lani_1194, Lcin_0519, Lmor_1975, LepB, AnkK, MavQ, SidJ) which suggests bacterial origin and/or rapid evolution in the pathogen. Finally, others are clustered with atypical host kinases (e.g., ABC1, PI3_PI4_kinase, PIP5K) which may suggest "ancient" kinases present in bacteria and eukaryotes.

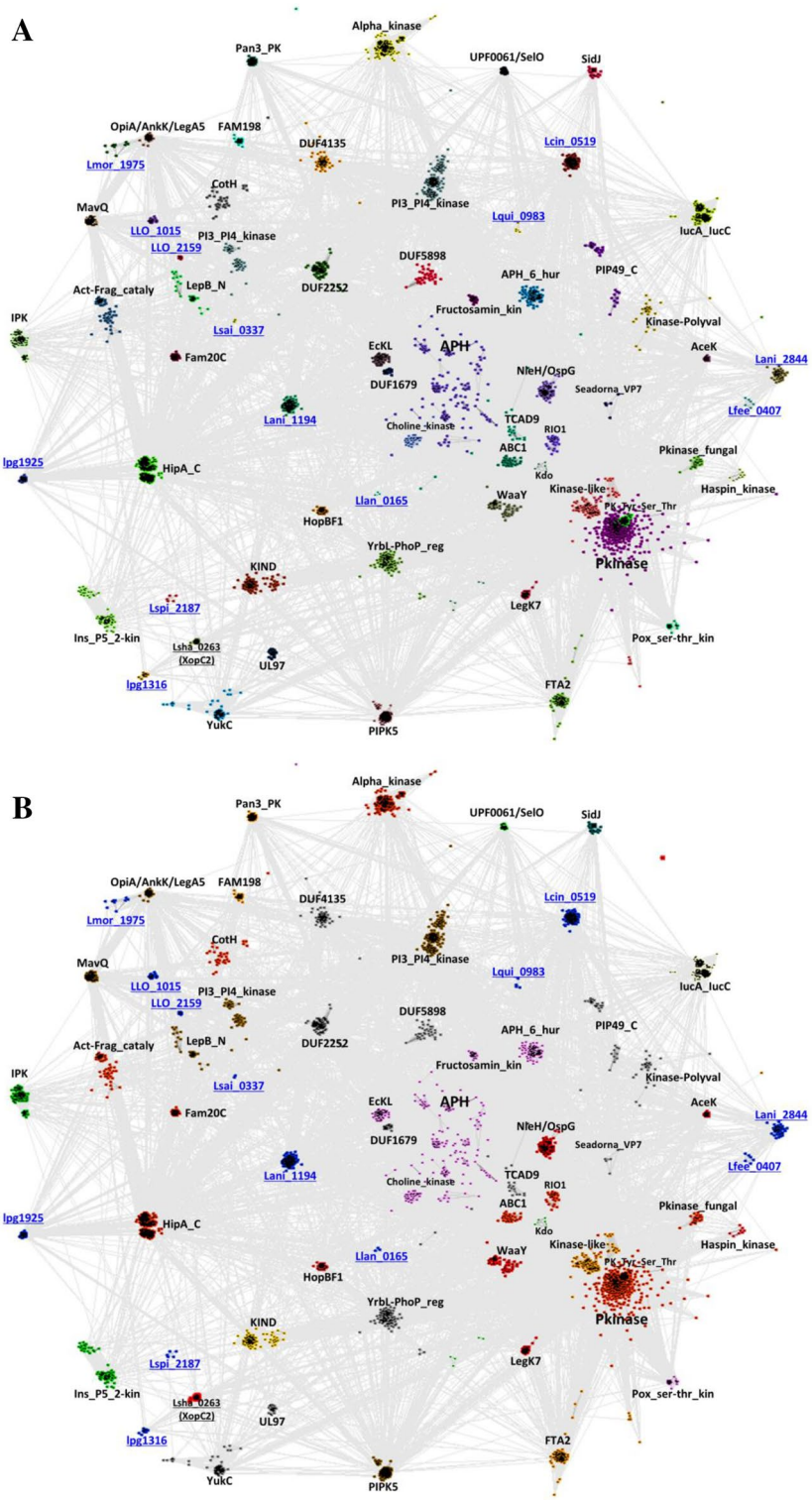


Figure 4. Sequence similarity graph for novel *Legionella* kinases and known PKL families from all domains of life. The CLANS graph includes representatives of all known PKL families (see "Methods" section). Graph edges represent protein sequence similarities detected by all-to-all BLAST comparisons up to the E-value of 1[47]. Pfam identifiers of selected families shown, for novel families, symbols of representative genes used. Novel families of *Legionella* (pseudo)kinases marked in blue underline. (**A**) Coloring by families. (**B**) Coloring by dominant function: red — protein phosphorylation, cyan — phospholipid phosphorylation, lime — lipopolysaccharide phosphorylation, pink — fructosamine phosphorylation, dark green — phosphorylation of inositol and derivatives, magenta — small molecule phosphorylation, brown — phosphorylation of phosphatidylinositol and derivatives, orange — pseudokinase (likely non-enzymatic functions), pale yellow — biosynthesis of small molecules, green-grey — glutamylation, light green — AMPylation (adenylylation), grey — unknown function or function predicted but unverified, blue — novel *Legionella* families.

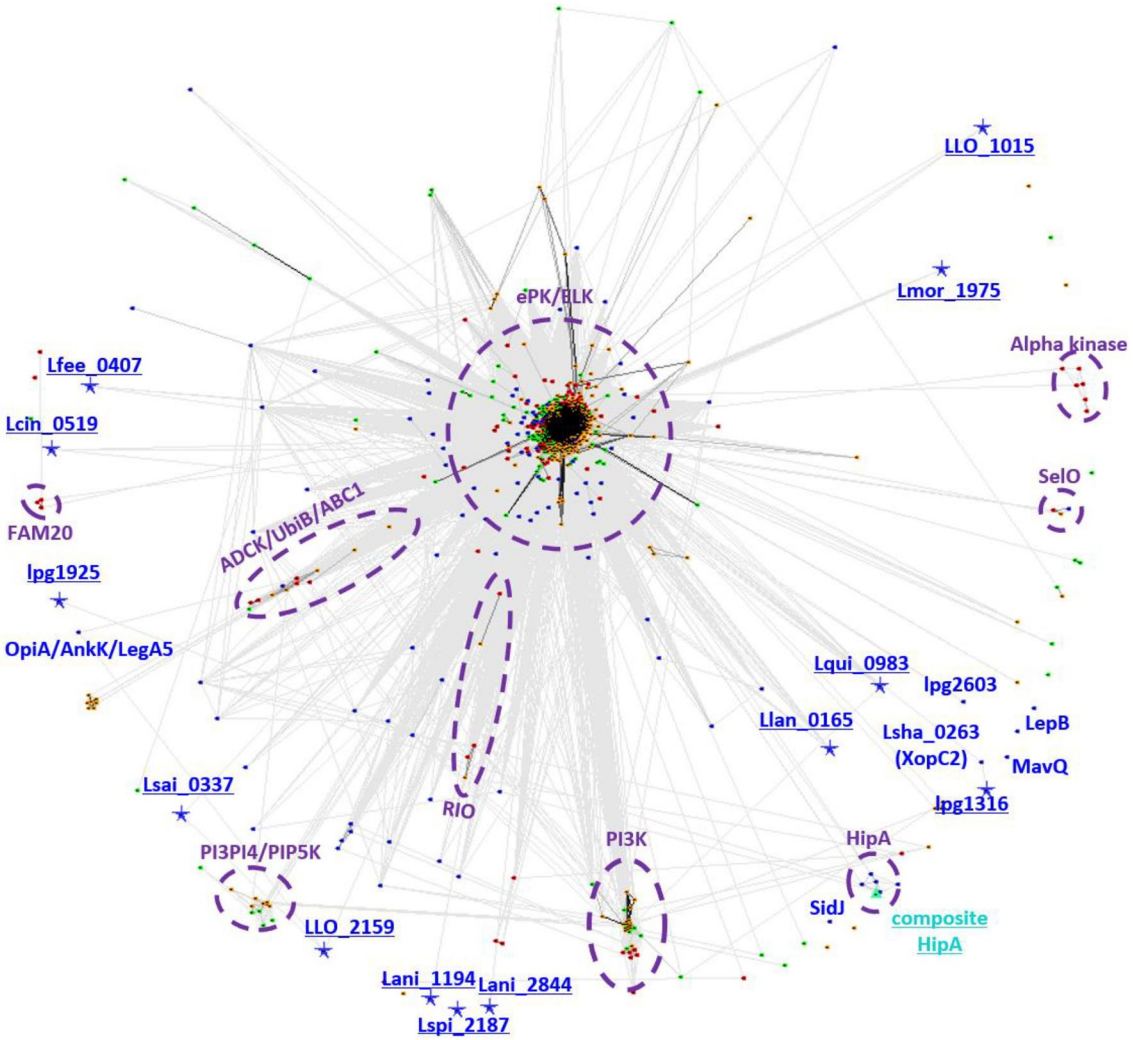


Figure 5. Sequence similarities: kinase-like proteins from *Legionella,* human and amoeba. The CLANS graph built as in Fig. 4 (up to BLAST E-value of 1[47]. Protein kinase-like proteins from the genus *Legionella* (blue), human (red) and amoebas — *Dictyostelium discoideum* (green) and *Acanthamoeba castellanii* (yellow). The novel kinase families are underlined and marked by asterisks. Ellipses mark selected protein families.

Some clusters are eukaryotic- or Metazoa-specific (e.g., Alpha kinases, FAM20), while others appear to be specific to bacteria (e.g., HipA). All the 13 novel families of *Legionella* kinases (see Fig. 5) are at the peripheries of the graph, indicating divergent evolution and arguing against host origin.

From among the 13 novel *Legionella* kinase families, those with most interesting functional implications are discussed in more detail below.

A kinase that may decorate bacterial secreted factors: Lani_1194. The putative effector kinase Lani_1194 is found in *Legionella anisa* and 15 other *Legionella* species (but not in *L. pneumophila*). *L. anisa* is the second most often isolated *Legionella* species in water samples, following *L. pneumophila*. This species is associated with cases of legionellosis. We can surmise that 6 out of 15*Legionella* species having this protein are human pathogens (*L. parisiensis, bozemanae, jamestowniensis, tucsonensis, jordanis* and *anisa*)[1,48,49].

In addition to *Legionella*, Lani_1194 homologs are found in 1201 species. The most numerous group here are soil bacteria of the order *Micromonosporales* (*Actinobacteria*), followed by *Flavobacteriales* (*Bacteroidetes*; bacteria of various environments) and soil bacteria, plant root symbionts—*Hyphomicrobiales* order of *Alphaproteobacteria*, e.g., the genera: *Rhizobium*, *Sinorhizobium*, *Mesorhizobium* and *Bradyrhizobium*. Majority of bacteria with Lani_1194 homologs appear to be non-pathogenic, although they are also found in some poorly studied strains of *Escherichia coli*, *Vibrio* and *Clostridium* whose pathogenicity is not yet determined. Also, Lani_1194 homologs are present in 12 species of *Archaea* (Fig. 6A). Among them are species from Gram-negative *Thermoproteota* (thermophilic or hyperthermophilic organisms)[50] and *Methanoculleus* genus (methanogenic *Archaea*)[51].

Lani_1194 and its homologs have a conserved kinase active site (see Fig. 1) albeit an arginine R38 is most likely the equivalent of the catalytic K72 of PKA. However, neither sequence analysis nor structure model allowed identification of the ion pair glutamate. The aspartate and asparagine residues corresponding to catalytic D166,

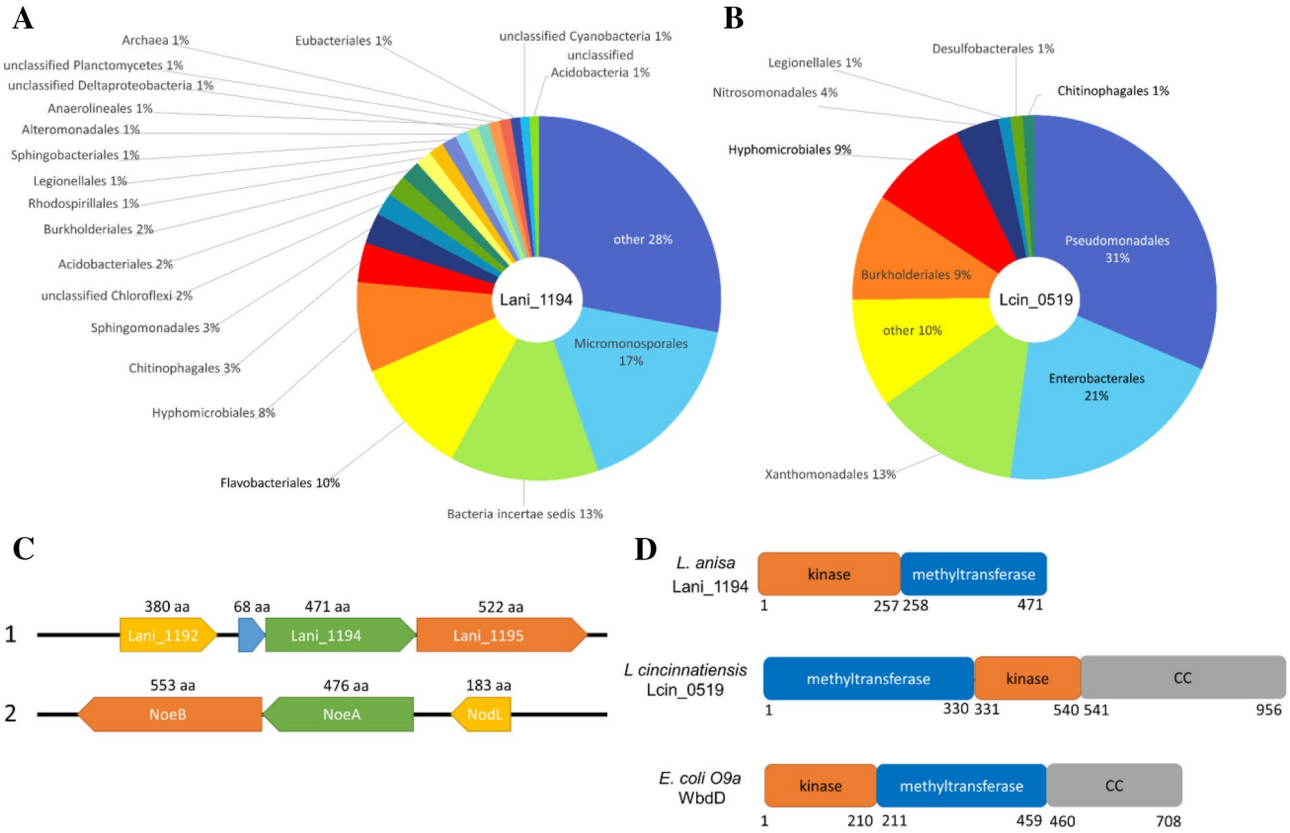


Figure 6. Lani_1194 and Lcin_0519—taxonomic spread, genomic neighborhoods, domain compositions. Organisms with homologs of: (**A**) Lani_1194 and (**B**) Lcin_0519 (found by BLAST search using the kinase domains as queries). Order level shown, or higher if not available. "Others" include taxa containing from 1 to 4 hits (organisms)[28,29]. (**C**) Genomic neighborhood of the protein Lani_1194 (1) and its homolog NoeA (2) in a nodulation-related operon from *Sinorhizobium meliloti*; lengths of encoded proteins shown. Coloring reflects homology. (**D**) Arrangement of structural domains of Lani_1194, Lcin_0519 and WbdD proteins. CC denotes the coiled-coil domain.

N171 and D184 of PKA are conserved albeit within atypical sequence motifs (see Fig. 1). Sequence conservation analysis and structure model allow us to delineate the extent of the kinase domain (see Fig. 7A, Suppl. Fig. S8), including a region remotely similar to the ATP-binding glycine-rich-loop.

The Lani_1194 protein also contains a second, easily identifiable domain, the class I S-adenosyl-methionine (SAM) dependent methyltransferase domain (SDM, MTase). Typically, SDMs transfer methyl groups from SAM to a wide range of acceptors, including small metabolites and biological macromolecules, DNA and proteins (e.g., histones) (Suppl. Table S5, Fig. 6D)[52]. According to the AlphaFold structure model, the two enzymatic domains form an extensive interface, with most contacts involving kinase C-lobe (see Fig. 7A).

The kinase—methyltransferase domain architecture is conserved: 94% of approx. two thousand Lani_1194 kinase domain homologs also have the methyltransferase domain.

Analysis of the genomic neighborhoods of Lani_1194 homologs indicated remarkable conservation of immediate genomic neighbors: Lani_1193 and Lani_1195 homologs occur in 65 and 38% of 816 analyzed genomic neighborhoods, respectively. Also, the closest genomic neighbors of Lani_1194 are most often located on the same DNA strand[53]. Such a conserved neighborhood may indicate an evolutionarily conserved functional unit (see Fig. 6C, Suppl. Table S6). Although Lani_1194 is functionally uncharacterized, its homolog and the homolog of its genomic neighbor Lani_1195, NoeA and NoeB (Suppl. Tables S5–S7), belong to an operon of *Sinorhizobium meliloti* which regulates the nodulation of particular *Medicago* plant species by chemically modifying nodulation factors (NFs), signaling molecules secreted by the bacteria to induce host plant to develop symbiosis-allowing root nodules. The biochemical "decoration" of NFs, specific for each bacterial strain and its host plant, occurs in the bacterial cytoplasm before NF secretion and is necessary for recognition of the bacterium as a potential symbiont[54].

The Lani_1195 and NoeB proteins are predicted to adopt the alkaline phosphatase fold (Suppl. Table S5). The Lani_1192 protein is annotated as O-antigen acetylase, its function is O-acetylation of LPS[55]. A similar function is performed by the NodL protein albeit with a different fold, from the NodL–NoeA–NoeB operon. NodL is also an acetyltransferase responsible for the O-acetylation of sulphated NFs.

NoeA protein, together with NodL and NodB, is possibly involved in the regulation of nodulation through modification of NF signaling molecules in *Rhizobium*[54]. In *Legionella*, the immediate genomic neighbors of the

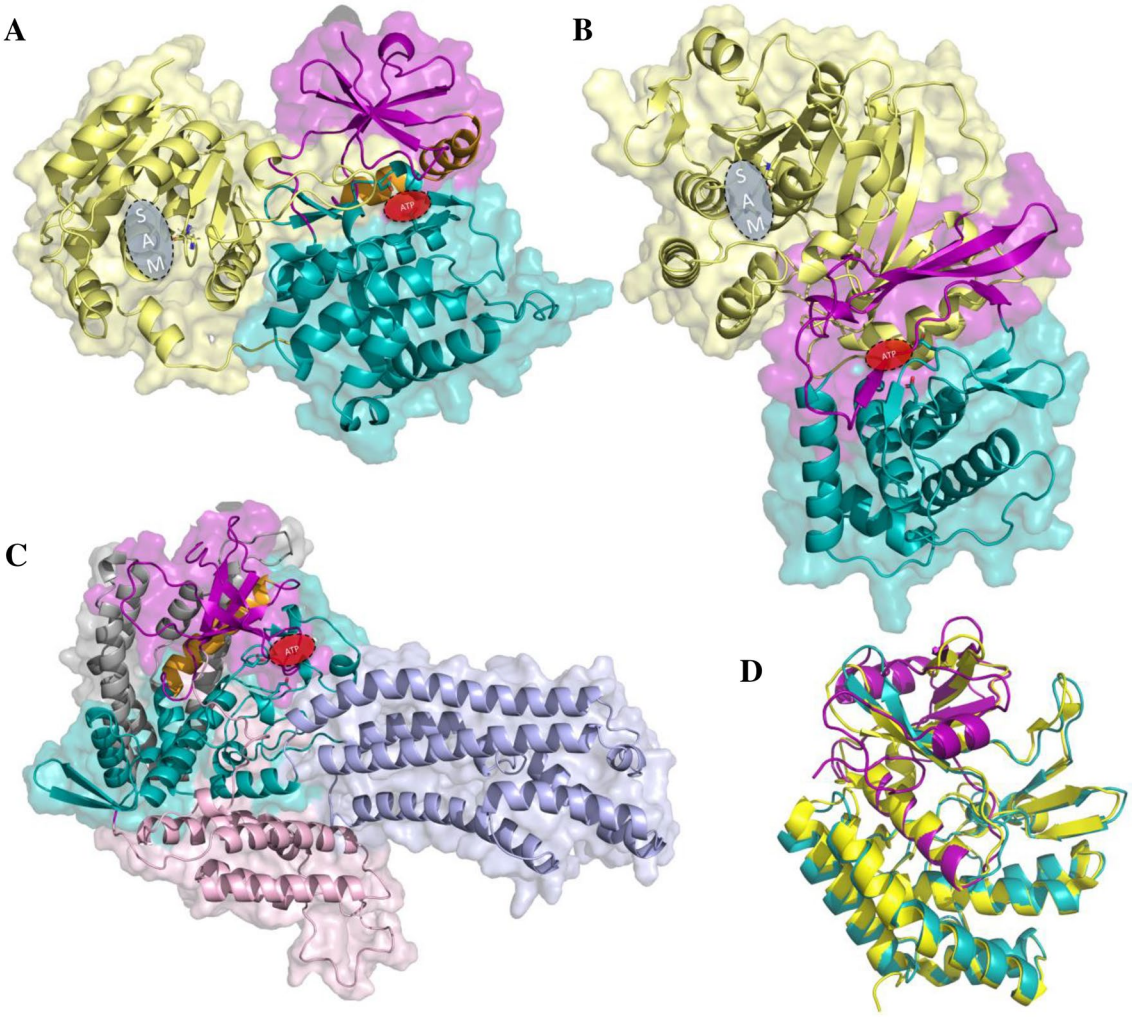


Figure 7. Structures of selected novel kinases. (**A**) Lani_1194, (**B**) Lcin_0519 and (**C**) Lmor_1975 structure models (AlphaFold2). (**D**). Structure comparison of composite HipA models (lpg2379—N-lobe and lpg2380—C-lobe) with lpg2370 structure (PDB:7VKB). Coloring in (**A–C**): Kinase N-lobes: purple, kinase C-lobes: teal, alpha-C helix in the kinase N-lobe: orange, methyltransferase domains: pale yellow. Additional domains in Lmor_1975: helical domain inserted between kinase N- and C-lobes: pink, helical bundle domain; light blue, C-terminal domain: gray. Coloring in (**D**): lpg2370: yellow, lpg2379: magenta, lpg2380: cyan. Residues corresponding to PKA active site D166 and D184 shown in stick representation. A predicted C-terminal coiled-coil region in Lcin_0519, and poorly predicted N-terminal helix in Lani_1194 omitted for clarity. Red ellipses mark the approximate active site region (ATP binding) of the kinase domain. Blue ellipse marks the approximate active site region (S-adenosylmethionine, SAM, binding) of the methyltransferase domain.

effector Lani_1194 (Lani_1193, Lani_1195) likely encode effector proteins (Suppl. Table S7). The roles of these nodulation gene homologs in *Legionella* infection are not clear but they might be decorating yet unknown signaling molecules secreted by the bacterium into the host cell or may act on the bacterial envelope.

A kinase that may decorate bacterial outer membrane lipopolysaccharides: Lcin_0519. The novel family of predicted effector kinases Lcin_0519 is found in the human pathogen *L. cincinnatiensis* and five other *Legionella* species[56,57].

Outside the *Legionella* genus, 871 homologs of Lcin_0519 in 333 species were found. The largest groups here are *Pseudomonadales* and *Enterobacteria*. In addition, *Xanthomonadales*, *Burholderiales* (mostly pathogens), *Hyphomicrobiales* (root symbiotic bacteria) and *Nitrosomonadales* (nitrification bacteria) are noticeable. Among the well-known organisms, it is found in some pathogenic human species, such as *Serratia marcescens*, *Klebsiella pneumoniae*, *Vibrio cholerae*, *Burkholderia cenocepacia*, and some known plant pathogens, such as *Pseudomonas syringae* or *Xanthomonas citri* (Fig. 6B).

The Lcin_0519 protein possesses the typical kinase catalytic residues (see Fig. 1, Suppl. Fig. S8), except the ion pair glutamate could not be identified. Indeed, AlphaFold structure model suggests that Lcin_0519 does not have an equivalent of the helix α-C present in most known protein kinases, and the β-sheet of the kinase N-lobe continues into the methyltransferase domain as its central β-sheet (see Fig. 7B).

Similarly to Lani_1194, the Lcin_0519 protein contains a second, easily detected domain, a methyltransferase (see Fig. 7B, Suppl. Table S5). According to the AlphaFold structure model, relative orientation of the two domains is different than in Lani_1194. The inter-domain interface in Lcin_0519 is even more extensive and involves both kinase lobes. This domain architecture is strictly conserved: 96% of proteins with Lcin_0519-like kinase domain also have the MTase domain (Fig. 6D), also common is a coiled-coil domain. Analysis of the co-occurrence of selected genes from close neighborhoods of Lcin_0519 homologs in 493 bacterial genomes (including 8 *Legionella* genomes) shows that 27% of the neighborhoods contain homologs of Lcin_0518 and Lcin_0520 while in 17% of neighborhoods there are also homologs of Lcin_0517[53] (Suppl. Table S6).

The Lcin_0518 protein is annotated as an ABC transporter of LPS O-antigen (Wzt), Lcin_0517—as an LPS transport system permease (Wzm) and Lcin_0520—as a glycosyltransferase (GTase). Together, these proteins in *Aquifex aeolicus* (Wzt, Wzm and GTase) secrete the complete O-antigen across the inner membrane for ligation to the LPS core[58].

Thus, also in *Legionella,* the Lcin_0519 kinase-MTase and its genomic neighbors can be predicted to be related to the modification of the bacterial outer membrane, e.g., Lcin_0519 might modify LPS through phosphorylation and methylation.

The protein domain composition of Lcin_0519 is reminiscent of a known enzyme, WbdD protein from *E. coli O9a*. WbdD has kinase, methyltransferase and CC domains (Fig. 6D). WbdD proteins are strain specific and regulate chain termination and length modifications of O-antigen[59]. However, WdbD and Lcin_0519 are clearly different, remotely related, kinase-MTase families. Although Lcin_0519 is annotated bioinformatically as an effector, it has not to our knowledge been studied experimentally. Thus, it can be speculated that Lcin_0519 may be not an effector, but indeed a "household" enzyme involved in the synthesis of LPS. In *Legionellas* it is known to be unique in comparison to most Gram-negative bacteria, highly variable between strains and species, and essential for infectivity[60,61]. Lcin_0519 may be therefore acting on the *Legionella* envelope and contributing to pathogenicity by adjusting envelope-host cell interactions to the requirements of the infection stage.

A kinase with a large internal insertion: Lmor_1975. Another unusual, predicted effector kinase was found in *L. moravica.* This protein, Lmor_1975, has homologs in only a few other closely related species (mainly *Legionella*, some potentially pathogenic to humans)[56].

Although sequence analysis (HHpred) detected Lmor_1975 similarity only to the C-lobe of LepB kinase, using sequence conservation and structure prediction we have identified equivalents of ion pair Lys and Glu in the N-lobe. Sequence analysis suggested, and AlphaFold structure model showed that Lmor_1975 has a large alpha-helical insertion between N-lobe and C-lobe, consisting of approx. 150 residues (see Fig. 7C, Suppl. Fig. S8). This is reminiscent of an insertion of approx. 80 amino acids found in atypical FAM69/DIPK kinases from Metazoans. The insertion in FAM69 contains an EF-hand calcium ion binding motif, located close to the ATP pocket between the N-lobe and the C-lobe and predicted to modulate kinase activity[62]. In Lmor_1975, the large insert and additional helical domains in the C-terminal region of the protein (see Fig. 7C) suggest a layer of regulation of kinase activity, possibly by interaction with intracellular structures or molecules.

The sequence similarity graph analysis located Lmor_1975 close to PI3K families: OpiA/AnkK[18] and MavQ[19], which suggests it may be a PI kinase (see Fig. 4).

The large helical insertion between N-lobe and C-lobe clearly obscures structural similarity to the PKL fold. The very weak similarity of Lmor_1975 to known kinases observed both in sequence and structure searches underscores the difficulty of recognizing distant homology in cases of large inserts within structural domains.

A likely "composite" HipA protein kinase formed from the products of lpg2378, lpg2379 and lpg2380 genes. Analyzing the "known" kinase effectors, we noticed a peculiar "composite" HipA kinase in *L. pneumophila.* HipA kinases play a very important role in stress response mechanisms of *E. coli* and many other Gram-negative bacteria by inducing a dormant state termed persistence. In *E. coli*, HipA is part of a toxin-antitoxin type system also including its genomic neighbor, the HipB antitoxin[63,64]. HipA phosphorylates glutamyl-tRNA synthetase, which results in inhibition of protein synthesis and growth arrest[64,65]. The activity of HipA is inhibited by binding to HipB and by HipB acting as a transcriptional autosuppressor of the hipBA operon[64].

In *L. pneumophila*, the putative "composite" kinase is encoded by two adjacent genes whose protein products together may form the complete HipA-type kinase domain (Fig. 8). Thus, lpg2379 encodes the kinase N-lobe and lpg2380 — the C-lobe. This likely indicates a gene fission phenomenon[66].

In the *L. pneumophila* genome near the lpg2379 and lpg2380 genes (8 kbp away) lies the lpg2370 gene which encodes a complete kinase domain of HipA type. The lpg2370 protein shares approx. 90% sequence identity with the lpg2379-lpg2380 pair, and consequently very high structural similarity (see Fig. 7D), which may indicate a recent duplication of an ancestral HipA-like gene and splitting of one the copies. The duplicated arrangement involving homologs of lpg2379-lpg2380 and lpg2370 genes is found only in 35 strains of *L. pneumophila* (e.g., Philadelphia-1, Burlington 1 (D-7841)).

Both lpg2370 and lpg2380 are predicted to be T4SS effectors (Suppl. Table S4).

Next to lpg2379-lpg2380 and to lpg2370 there are also genes encoding another element of the "classic" HipA protein, the N-terminal "HipA-coupled" domain (lpg2369 and lpg2378, respectively), responsible for dimerization during DNA binding[63] (see Fig. 8). Further, both HipA-like proteins are accompanied in the genome by homologs of the HipB antitoxin, lpg2368 and lpg2377, respectively. Interestingly, both HipB-like proteins are weakly predicted to be effectors. It remains to be tested if indeed *Legionella* delivers HipA-like kinase(s) to the host cell cytoplasm, and whether these effectors are accompanied by N-terminal subdomains and HipB suppressors.

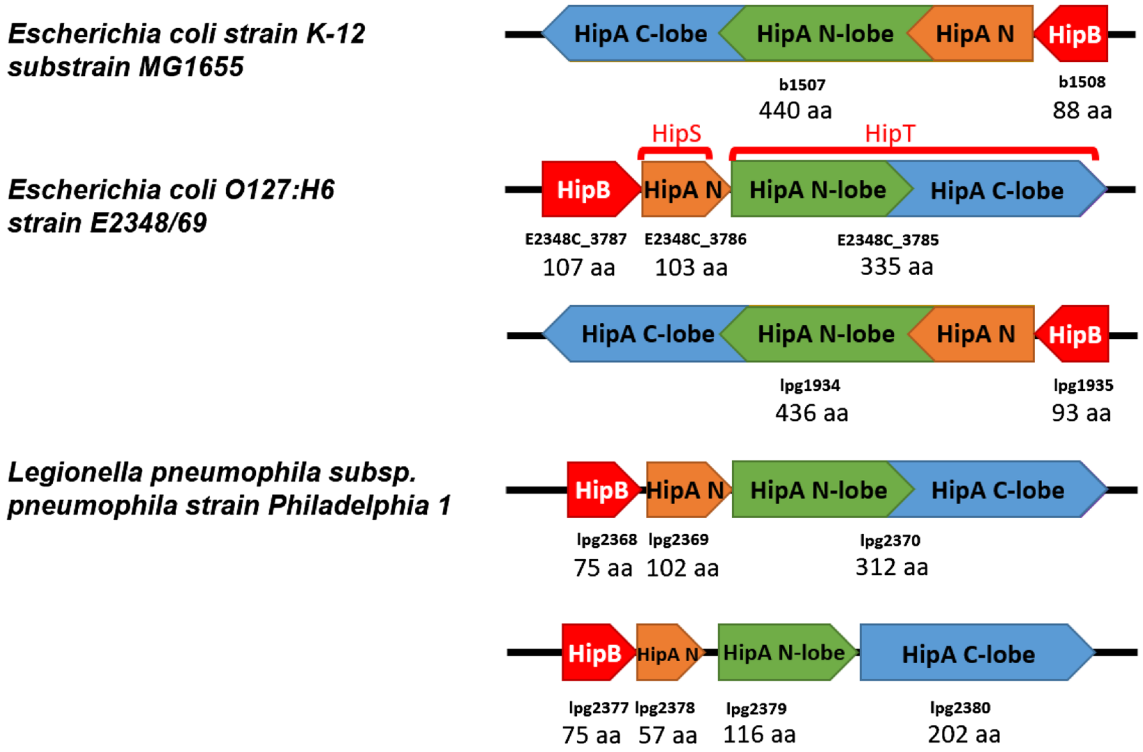


Figure 8. HipA-like modules in different bacteria. Gene loci names and protein lengths are shown below the gene diagrams. The colors represent HipB (red) and the subdomains of HipA (HipA-coupled N-terminal domain—orange, kinase N-lobe—green, kinase C-lobe—blue). The *L. pneumophila* protein lpg1934 (not shown) has the same gene/domain arrangement as *E. coli* K12 HipA.

Yet another *L. pneumophila* protein, lpg1934, appears to be a typical HipA (28% identity to *E. coli* HipA). It has a full kinase domain, an N-terminal "HipA-coupled" domain in one protein, and it's not recognized as an effector (Suppl. Table S4).

The lpg2379 and lpg2380 genes overlap by 40 nucleotides, a kind of overlap observed often in prokaryotic genomes. Also, lpg2378 and lpg2379 genes lie in the + 1 reading frame while lpg2380 lies in the + 2 reading frame. The fact that lpg2379 and lpg2380 genes lie in two different reading frames argues against separation of these genes being the result of gene misprediction or sequencing error[67].

In an analogy to our observation, a protein from the HipA family was recently discovered in *E. coli O127*, split into two proteins encoded by distinct genes: a kinase domain (HipT gene) and an N-terminal HipA-coupled domain (HipS gene). Recently, it has been shown that lpg2368–lpg2369–lpg2370 act as a HipBST toxin-antitoxin system similar to that in *E. coli*[68]. The lpg2379-lpg2380 pair is another case of an elaborate HipA module and an example of how gene fusion, fission and duplication shape and create new cellular signals[69,70]. The possibility that *Legionella* employs the purely bacterial HipA family to manipulate eukaryotic signaling is particularly interesting, given HipA have evolved in the context of bacterial intracellular signaling.

Further, the composite kinase may offer a yet unknown layer of kinase regulation by assembly of a functional enzyme from subunits from separate polypeptide chains.

Conclusions

In this bioinformatic analysis of 41 *Legionella* species, we cataloged 112 protein kinase-like Legionella Orthologous Groups (LOGs) within 29 families, of which 13 families are novel. We have discussed in detail sequence/structure features and proposed functional predictions for three novel families and a putative new composite HipA kinase. The novel PKL families identified by sequence searches were confirmed by artificial intelligence-based structure predictions.

Two novel families, Lani_1194 and Lcin_0519, were found to occur far beyond *Legionellas*. This introduces an intriguing prospect of related enzymatic machinery being used for different purposes in different biological scenarios, i.e., for nodulation-related signaling between rhizobial bacteria and plant hosts, and for rewiring intracellular signaling in amoebas and animals infected by *Legionellas*. Although literature evidence suggests most *Legionella* effectors act on host cell molecules or on each other, acting on bacterial own cell envelope can also be relevant for infection[60,61].

An inherent limitation of the present study is the fact that these functional predictions rely on literature data available for homologs. Nevertheless, this makes the novel kinase-like families even more attractive subjects for experimental studies. In a rather unlikely case the effector predictions for the novel families are wrong, these

families still may be attractive as targets of a therapeutic intervention, because even if not delivered to the host cell they are likely to perform roles important for the pathogen's survival.

Legionella kinomes are rich in effector kinases in addition to their sets of "household" kinases. This indicates their adaptation to different hosts—mostly *Protozoa*, but also higher eukaryotes. The most studied *Legionella* species — *L. pneumophila* has a set of genes for both infecting various *Amoebae* and macrophages in the human lung. Some of these kinases, such as LegK1-4, structurally and sequentially closely resemble eukaryotic kinases, perhaps having been "hijacked" by the way of gene transfer from eukaryotes and evolutionarily adapted. Others, while retaining the PKL fold, appear to be very distantly related to known kinases, which obscures their evolutionary origin, likely due to high evolutionary pressure.

Thus, we have created a catalog of *Legionella* (pseudo)kinases, available at http://bioinfo.sggw.edu.pl/kintaro/ thanks to the comprehensive analysis of the pan-proteome of 41 species of this genus. The discovery of these kinases may aid in developing new approaches to fight these pathogens. Moreover, these novel families are often found in other pathogens of animals and plants. Thus, the survey of *Legionella* pan-kinome presented herein offers starting points into studies of this pathogen's infection toolbox, but also a broader perspective on the ingeniousness of nature in diversifying, developing and repurposing the successful kinase-like superfamily.

Materials and methods

Search strategy. The general approach used in this work to search for novel kinase-like families was described recently[71]. Briefly, the screen for novel kinase-like proteins starts with a set of protein sequences (here *Legionella* proteins set) where redundancy is reduced and representative sequences are split into fragments. In the next steps algorithms for remote homology detection are used (FFAS, HHpred) for searching and validating similarity to kinases. Additionally, for candidate kinase-like proteins, three-dimensional structure models are built and compared with known kinase structures.

Sequence data. Protein sequence data for *Legionella* effectors was provided by the article by Burstein et al.[25].

Clustering. Due to the large size of the sequence data, the sequences were clustered by sequence identity using the CD-HIT algorithm[31]. Two clustering thresholds were used: 90% and 50% sequence identity. This reduces the load on the processor.

Splitting sequences. Sequences were split[32] into 300 aa length with overlap of 100 aa.

Remote homology detection. For distant similarity prediction to PKL families three methods were used, the profile-profile alignment and fold recognition algorithm — FFAS[33] (COG[72], Hsapiens[73], PDB[74], SCOP[75], Pfam[76] databases); homology detection and structure prediction—HHpred and similar HHsearch pipeline that uses hidden Markov model HMM-to-HMM comparison[36] (PDB[74], SCOP[75], Pfam[76] databases); and a similar method Phyre2, which additionally models 3D structure of query and compares it with 3D models library[37]. Standard parameters were used, however both significant hits and those not formally significant were taken into account.

Multiple sequence alignments and sequence logos. Novel families were collected using BLAST (NR, E-value = 1e−4)[28–30] and aligned using the MAFFT[77] algorithm with default settings. Next, the sequence logos were prepared using the WebLogo algorithm[27]. The WebLogo program generates the sequence logos based on the multiple sequence alignments. Here, the alignments are processed with an in-house script that removes the columns containing gaps in the "master" sequence.

Secondary structure prediction was performed by Jpred4[38].

Structure modeling and comparison. Novel PKL-like structures were modeled with use of RoseTTAFold[39] and AlphaFold2[40] (the best models have been selected). Comparisons of structures were performed using FATCAT[41] and Dali[42] servers.

Visual clustering of families (analysis of sequence similarity relations between families). To visualize clusters of protein kinase families, the cluster of sequences (CLANS) algorithm[47] was used with the BLOSUM62 scoring matrix and extraction of BLAST hits up to E-value of 1. The set of sequences was collected as follows:

Newly predicted protein kinase families collected by BLAST (NR database, E-value = 1e−4)[28–30] and clustered by CD-HIT at 50% sequence identity—Lani_1194. Lcin_0519, or at 99% sequence identity—Lani_2844, Lfee_0407, LLO_1015, LLO_2159, Lmor_1975, lpg1316, lpg1925, Lsai_0337, Lqui_0983, Llan_0165, Lspi_2187)[30].

Families of PKinase clan from the Pfam database: APH_6_hur (rp15 sequence set), APH (seed sequence set), Choline_kinase (seed), CotH (seed), DUF1679 (seed), DUF2252 (rp15; flipped N-lobe and C-lobe), DUF4135 (rp15), DUF5898 (rp15), EcKL (seed), Fam20C (seed), Frukosamin_kin (seed), FTA2 (rp35 sequence set), Haspin_kinase (seed), HipA_C (seed), Ins_P5_2-kin (seed), IPK (seed), IucA_IucC (seed), Kdo (seed), Kinase-PolyVal (rp35), Pan3_PK (seed), PI3_PI4_kinase (rp15, clustered at 40% sequence identity), PIP49_C (seed), PIP5K (seed), Pkinase_fungal (seed), Pkinase (rp15, because the rp15 set is very large, it was clustered at 25% identity level, and sequences longer than 300 residues were selected), PK_Tyr_Ser_Thr (seed),

Pox_ser-thr_kin (rp15), RIO1 (seed), Seadorna_VP7 (rp75), UL97 (rp15), WaaY (rp55), YrbL-PhoP_reg (rp35), YukC (rp35), families not yet included in PKinase clan, but having PKL fold (LepB_N (rp55 sequence set), FAM198 (rp15), SelO (seed))

Other proteins with predicted or known fold similar to PKL not included yet in Pfam database (collected by BLAST, NR, E-value = 1e−4) — OpiA/AnkK/LegA5, HopBF1, lpg1924/LegK7, MavQ, NleH–OspG, SidJ, Lsha_0263 (XopC2)[30]. All the families were manually curated and corrected (domains were extended when they appeared not to include full kinase-like structural domains). From Lmor_1975 and Lsha_0263, helical inserts were removed.

Substrates of secretion systems. Substrates of secretion systems were predicted with use of SignalP6.0[44], EffectiveDB (EffectiveT3, T4SEpre, EffectiveCCBD, EffectiveELD)[45] and BastionX[46]. All programs were used with default settings.

Coiled-coil domains. Coiled-coil domain was predicted by DeepCoil[78].

Transmembrane helices. Transmembrane helices were predicted with use of TMHMM[79].

Phylogenetic trees and species heatmap. Phylogenetic tree of *Legionella* strains was adapted from the article by Burstein et al.[25]. Heatmap of the number of PKL genes was clustered by hierarchical clustering (Manhattan method; single linkage) to see similar arrangements of genes in *Legionella* species.

For kinase-like family phylogenetic trees, multiple sequence alignments were done using the structure alignment program mTM-align[80]. Where no experimental structures were available (e.g., for the novel kinase families), structure models were built using AlphaFold. Alignment trimming was performed using ClipKit[81] and manually corrected. The phylogenetic tree was built using the MEGA program (default settings) using ML method and bootstrapping = 500[82]. For the PI3-PI4 kinase-like tree, all human representatives were used while amoeba sequences were clustered at 30% sequence identity threshold (cdhit). For the eukaryotic-like kinase tree, human sequences were clustered at 30%.Phylogenetic trees visualization was done in ITOL[43].

Potential horizontal gene transfer analysis. For this purpose, we use CLANS[47] analysis (parameters: BLOSUM62 scoring matrix; extraction BLAST HSP's up to E-values of 1) with kinomes of *Legionella* and its hosts (*Homo sapiens*, *Dictyostelium discoideum* and *Acanthamoeba castellanii*. CLANS analysis clusters similar sequences into groups.

Taxonomic distribution analysis of homologs was done using BLAST[28–30]. The numbers of bacterial and eukaryotic homologs of *Legionella* eukaryotic-like kinases were compared. In cases where the number of eukaryotic homologs of a *Legionella* ELK is significantly larger than the number of bacterial homologs, an eukaryote-to-bacteria horizontal gene transfer can be hypothesized.

Data availability

The following information was supplied regarding data availability: Raw data (including PDB files for protein structure models) are available in the Supplemental Files. Sets of aligned representative sequences of *Legionella* (pseudo)kinase families are available from the online database at http://bioinfo.sggw.edu.pl/kintaro/.

Received: 9 September 2022; Accepted: 9 December 2022
Published online: 16 December 2022

References

1. Chambers, S. T., Slow, S., Scott-Thomas, A. & Murdoch, D. R. Legionellosis caused by non-*Legionella pneumophila* species, with a focus on *Legionella longbeachae*. *Microorganisms* **9**, 291 (2021).
2. Gomez-Valero, L. *et al.* More than 18,000 effectors in the *Legionella* genus genome provide multiple, independent combinations for replication in human cells. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2265–2273 (2019).
3. Qin, T., Zhou, H., Ren, H. & Liu, W. Distribution of secretion systems in the genus Legionella and its correlation with pathogenicity. *Front. Microbiol.* **8**, 388 (2017).
4. Mondino, S. *et al.* Legionnaires' disease: State of the art knowledge of pathogenesis mechanisms of Legionella. *Annu. Rev. Pathol.* **15**, 439–466 (2020).
5. Black, M. H. *et al.* A Legionella effector ADP-ribosyltransferase inactivates glutamate dehydrogenase. *J. Biol. Chem.* **296**, 100301 (2021).
6. Khaledi, A., Bahrami, A., Nabizadeh, E., Amini, Y. & Esmaeili, D. Prevalence of Legionella species in water resources of Iran: A systematic review and meta-analysis. *Iran. J. Med. Sci.* **43**, 571–580 (2018).
7. Grishin, A. M., Beyrakhova, K. A. & Cygler, M. Structural insight into effector proteins of Gram-negative bacterial pathogens that modulate the phosphoproteome of their host. *Protein Sci.* **24**, 604–620 (2015).
8. Mattoo, S., Lee, Y. M. & Dixon, J. E. Interactions of bacterial effector proteins with host proteins. *Curr. Opin. Immunol.* **19**, 392–401 (2007).
9. Urbanus, M. L. *et al.* Diverse mechanisms of metaeffector activity in an intracellular bacterial pathogen. *Legionella pneumophila*. *Mol. Syst. Biol.* **12**, 893 (2016).
10. *Pfam: Clan: PKinase (CL0016)*. https://pfam.xfam.org/clan/CL0016. Accessed Jul. 04, 2021.
11. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
12. Pollock, G. L. *et al.* Targeting of microvillus protein Eps8 by the NleH effector kinases from enteropathogenic *E. coli*. *Proc. Natl. Acad. Sci.* **119**, e2204332119 (2022).
13. Lopez, V. A. *et al.* A bacterial effector mimics a host HSP90 client to undermine immunity. *Cell* **179**, 205–218 (2019).

14. Michard, C. *et al.* The Legionella kinase LegK2 targets the ARP2/3 complex to inhibit actin nucleation on phagosomes and allow bacterial evasion of the late endocytic pathway. *mBio* **6**, 15 (2015).
15. Moss, S. M. *et al.* A *Legionella pneumophila* kinase phosphorylates the Hsp70 chaperone family to inhibit eukaryotic protein synthesis. *Cell Host Microbe* **25**, 454–462 (2019).
16. Lee, P.-C. & Machner, M. P. The Legionella effector kinase LegK7 Hijacks the host hippo pathway to promote infection. *Cell Host Microbe* **24**, 429-438.e6 (2018).
17. Dong, N. *et al.* Modulation of membrane phosphoinositide dynamics by the phosphatidylinositide 4-kinase activity of the Legionella LepB effector. *Nat. Microbiol.* **2**, 16236–16236 (2016).
18. Ledvina, H. E. *et al.* A phosphatidylinositol 3-kinase effector alters phagosomal maturation to promote intracellular growth of francisella. *Cell Host Microbe* **24**, 285–295 (2018).
19. Hsieh, T.-S. *et al.* Dynamic remodeling of host membranes by self-organizing bacterial effectors. *Science*. https://doi.org/10.1126/science.aay8118 (2021).
20. Leonard, C. J., Aravind, L. & Koonin, E. V. Novel families of putative protein kinases in bacteria and archaea: Evolution of the "eukaryotic" protein kinase superfamily. *Genome Res.* **8**, 1038–1047 (1998).
21. Correia, F. F. *et al.* Kinase activity of overexpressed HipA is required for growth arrest and multidrug tolerance in *Escherichia coli*. *J. Bacteriol.* **188**, 8360–8367 (2006).
22. Fong, D. H., Lemke, C. T., Hwang, J., Xiong, B. & Berghuis, A. M. Structure of the antibiotic resistance factor spectinomycin phosphotransferase from *Legionella pneumophila*. *J. Biol. Chem.* **285**, 9545–9555 (2010).
23. Sreelatha, A. *et al.* A Legionella effector kinase is activated by host inositol hexakisphosphate. *J. Biol. Chem.* **295**, 6214–6224 (2020).
24. Black, M. H. *et al.* Bacterial pseudokinase catalyzes protein polyglutamylation to inhibit the SidE-family ubiquitin ligases. *Science* **364**, 787–792 (2019).
25. Burstein, D. *et al.* Uncovering the Legionella genus effector repertoire—Strength in diversity and numbers. *Nat. Genet.* **48**, 167–175 (2016).
26. Hanks, S. K. & Hunter, T. The eukaryotic protein kinase superfamily: Kinase (catalytic) domain structure and classification. *FASEB J.* **9**, 576–596 (1995).
27. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
29. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
30. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **47**, D23–D28 (2019).
31. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
32. Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *ERA*. https://doi.org/10.7939/R3FQ9QK0V (2000).
33. Xu, D., Jaroszewski, L., Li, Z. & Godzik, A. FFAS-3D: Improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* **30**, 660–667 (2014).
34. Lu, S. *et al.* CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).
35. Marchler-Bauer, A. *et al.* CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**, 281–283 (2002).
36. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **20**, 473 (2019).
37. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
38. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: A protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).
39. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a 3-track neural network. *Science* **373**, 871–876 (2021).
40. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
41. Li, Z., Jaroszewski, L., Iyer, M., Sedova, M. & Godzik, A. FATCAT 2.0: Towards a better understanding of the structural diversity of proteins. *Nucleic Acids Res.* **48**, W60–W64 (2020).
42. Holm, L. Dali server: Structural unification of protein families. *Nucleic Acids Res.* **50**, W210–W215 (2022).
43. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
44. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
45. Eichinger, V. *et al.* EffectiveDB—Updates and novel features for a better annotation of bacterial secreted proteins and type III, IV, VI secretion systems. *Nucleic Acids Res.* **44**, D669–D674 (2016).
46. Wang, J. *et al.* BastionHub: A universal platform for integrating and analyzing substrates secreted by Gram-negative bacteria. *Nucleic Acids Res.* **49**, D651–D659 (2021).
47. Frickey, T. & Lupas, A. CLANS: A java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).
48. Croze, A. *et al.* Intracellular behaviour of legionella non-pneumophila strains within three amoeba strains, including *Willaertia magna* C2c Maky. *Pathogens* **10**, 1350 (2021).
49. Herran, B., Grève, P., Berjeaud, J.-M., Bertaux, J. & Crépin, A. Legionella spp. All ears? The broad occurrence of quorum sensing elements outside *Legionella pneumophila*. *Genome Biol. Evol.* **13**, 032 (2021).
50. Ren, M. & Wang, J. Phylogenetic divergence and adaptation of nitrososphaeria across lake depths and freshwater ecosystems. *ISME J.* **16**, 1491–1501 (2022).
51. Maestrojuán, G. M., Boone, D. R., Xun, L., Mah, R. A. & Zhang, L. Transfer of *Methanogenium bourgense*, *Methanogenium marisnigri*, *Methanogenium olentangyi*, and *Methanogenium thermophilicum* to the Genus *Methanoculleus* gen. nov., Emendation of *Methanoculleus marisnigri* and methanogenium, and description of new strains of *Methanoculleus bourgense* and *Methanoculleus marisnigri*. *Int. J. Syst. Evol. Microbiol.* **40**, 117–122 (1990).
52. Martin, J. L. & McMillan, F. M. SAM (dependent) I AM: The S-adenosylmethionine-dependent methyltransferase fold. *Curr. Opin. Struct. Biol.* **12**, 783–793 (2002).
53. Chen, I.-M.A. *et al.* IMG/M vol 5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
54. Ardourel, M. *et al.* In *Rhizobium meliloti*, the operon associated with the nod box n5 comprises nodL, noeA and noeB, three host-range genes specifically required for the nodulation of particular Medicago species. *Mol. Microbiol.* **17**, 687–699 (1995).
55. Fox, K. L. *et al.* Novel lipopolysaccharide biosynthetic genes containing tetranucleotide repeats in *Haemophilus influenzae*, identification of a gene for adding O-acetyl groups. *Mol. Microbiol.* **58**, 207–216 (2005).
56. Ministry of Health. *The Prevention of Legionellosis in New Zealand: Guidelines for the Control of Legionella Bacteria* (Ministry of Health, 2011).

57. Jernigan, D. B. *et al.* Pulmonary infection due to *Legionella cincinnatiensis* in renal transplant recipients: Two cases and implications for laboratory diagnosis. *Clin. Infect. Dis.* **18**, 385–389 (1994).
58. Bi, Y. & Zimmer, J. Structure and ligand binding properties of the O antigen ABC transporter carbohydrate binding domain. *Structure* **28**, 252-258.e2 (2020).
59. Hagelueken, G. *et al.* Structure of WbdD: A bifunctional kinase and methyltransferase that regulates the chain length of the O antigen in *Escherichia coli* O9a: Crystal structure of *E. coli* WbdD. *Mol. Microbiol.* **86**, 730–742 (2012).
60. Kowalczyk, B., Chmiel, E. & Palusinska-Szysz, M. The role of lipids in legionella-host interaction. *Int. J. Mol. Sci.* **22**, 1487 (2021).
61. Shevchuk, O., Jäger, J. & Steinert, M. Virulence properties of the *Legionella pneumophila* cell envelope. *Front. Microbiol.* **2**, 74 (2011).
62. Dudkiewicz, M., Lenart, A. & Pawłowski, K. A novel predicted calcium-regulated kinase family implicated in neurological disorders. *PLoS ONE* **8**, e66427 (2013).
63. Gerdes, K., Bærentsen, R. & Brodersen, D. E. Phylogeny reveals novel HipA-homologous kinase families and toxin-antitoxin gene organizations. *mBio* **12**, e01058 (2021).
64. Vang Nielsen, S. *et al.* Serine-threonine kinases encoded by split hipA homologs inhibit tryptophanyl-Trna synthetase. *mBio* **10**, e01138 (2019).
65. Chowdhury, N., Kwan, B. W. & Wood, T. K. Persistence increases in the absence of the alarmone guanosine tetraphosphate by reducing cell growth. *Sci. Rep.* **6**, 1–9 (2016).
66. Wang, W., Yu, H. & Long, M. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. *Nat. Genet.* **36**, 523–527 (2004).
67. Pallejà, A., Harrington, E. D. & Bork, P. Large gene overlaps in prokaryotic genomes: Result of functional constraints or mispredictions? *BMC Genom.* **9**, 335 (2008).
68. Zhen, X. *et al.* Molecular mechanism of toxin neutralization in the HipBST toxin-antitoxin system of *Legionella pneumophila*. *Nat. Commun.* **13**, 4333 (2022).
69. Fani, R., Brilli, M., Fondi, M. & Lió, P. The role of gene fusions in the evolution of metabolic pathways: The histidine biosynthesis case. *BMC Evol. Biol.* **7**, S4 (2007).
70. Pasek, S., Risler, J.-L. & Brézellec, P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* **22**, 1418–1423 (2006).
71. Black, M. H., Gradowski, M., Pawłowski, K. & Tagliabracci, V. S. Chapter 19—Methods for discovering catalytic activities for pseudokinases. In *Methods in Enzymology* Vol. 667 (eds Jura, N. & Murphy, J. M.) 575–610 (Academic Press, 2022).
72. Galperin, M. Y. *et al.* COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2020).
73. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2020).
74. Goodsell, D. S. *et al.* RCSB protein data bank: Enabling biomedical research and drug discovery. *Protein Sci.* **29**, 52–65 (2020).
75. Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural classification of proteins—Extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–D309 (2014).
76. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
77. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
78. Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V. & Dunin-Horkawicz, S. DeepCoil—A fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics* **35**, 2790–2795 (2019).
79. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
80. Dong, R., Peng, Z., Zhang, Y. & Yang, J. mTM-align: An algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* **34**, 1719–1725 (2018).
81. Steenwyk, J. L., Iii, T. J. B., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
82. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).

Acknowledgements

The authors thank Drs Vincent Tagliabracci, Anna Muszewska, Marcin Grynberg and Victor Lopez for critical reading of the manuscript. They thank the Tagliabracci lab members for the many intellectually enriching projects on novel kinase families. K.P. was supported by Polish National Agency for Scientific Exchange scholarship 541 PPN/BEK/2018/1/00431 and Polish National Science Centre Grant 2019/33/B/NZ2/01409. M.G. was supported by the Polish National Science Centre Grant 2019/35/N/NZ2/02844. They thank Dr. Burstein for providing the phylogenetic tree of the 41 species of *Legionella*.

Author contributions

Performed the experiments and analyzed the data: M.G., M.K., B.B., B.D. and K.P. Prepared figures and/or tables: M.G., M.K., K.P. Wrote the original draft: M.G. Authored and reviewed drafts of the paper: M.G., M.K. and K.P. Conceived and designed the experiments: K.P. and M.G. All the authors approved the final draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-26109-x.

Correspondence and requests for materials should be addressed to K.P. or M.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Warszawa, 1.09.2025 r.

Marianna Krysińska
marianna_krysinska@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska, M., Baranowski, B., Deszcz, B., Pawłowski, K., & Gradowski, M. (2022). Pan-kinome of Legionella expanded by a bioinformatics survey.** ***Scientific reports*****,** ***12*****(1), 21782** mój indywidualny udział w jej powstaniu polegał na przeprowadzeniu eksperymentów dotyczących identyfikacji i charakteryzacji nowych rodzin, analizy podobieństw sekwencyjnych, analiz taksonomicznych i filogenetycznych oraz przeanalizowaniu uzyskanych danych, przygotowaniu grafik, wykresów i tabel do głównej części artykułu, opracowaniu i zrecenzowaniu roboczych wersji artykułu oraz zatwierdzeniu ostatecznej wersji manuskryptu.

Podpis

Warszawa, 28.07.2025 r.

dr Bartosz Baranowski
bartosz.piotr.baranowski@gmail.com

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Baranowski B, Deszcz B, Pawłowski K, & Gradowski M. (2022). Pan-kinome of *Legionella* expanded by a bioinformatics survey. *Scientific reports*, *12*(1), 21782** mój indywidualny udział w jej powstaniu polegał na wykonaniu części analiz bioinformatycznych dotyczących otoczeń genomicznych oraz zatwierdzeniu ostatecznej wersji manuskryptu.

Podpis

Warszawa, 29.05.2025 r.

Bartłomiej Deszcz
bartek_deszcz@wp.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Baranowski B, Deszcz B, Pawłowski K, & Gradowski M. (2022). Pan-kinome of *Legionella* expanded by a bioinformatics survey. *Scientific reports*, *12*(1), 21782** mój indywidualny udział w jej powstaniu polegał na wykonaniu części analiz bioinformatycznych dotyczących modelowania struktur białek oraz zatwierdzeniu ostatecznej wersji manuskryptu.

Podpis

Dallas, 30.10.2025

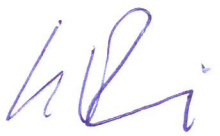
dr hab. Krzysztof Pawłowski
Krzysztof.Pawlowski@utsouthwestern.edu

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska, M., Baranowski, B., Deszcz, B., Pawłowski, K., & Gradowski, M. (2022). Pan-kinome of Legionella expanded by a bioinformatics survey.** ***Scientific reports*****,** ***12*****(1), 21782** mój indywidualny udział w jej powstaniu polegał na zaprojektowaniu eksperymentów, wykonaniu części badań dotyczących poszukiwania odległych homologów kinaz i analizie uzyskanych danych, recenzowaniu i zatwierdzeniu ostatecznej wersji artykułu.



Podpis

Warszawa, 1.09.2025 r.

dr Marcin Gradowski
marcin_gradowski@sggw.edu.pl

Rada Dyscypliny Nauki Biologiczne

Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy **Krysińska M, Baranowski B, Deszcz B, Pawłowski K, Gradowski M. 2022. Pan-kinome of *Legionella* expanded by a bioinformatics survey. Scientific Reports. 12: 21782.** mój indywidualny udział w jej powstaniu polegał na współkierowaniu projektem, planowaniu badań, , wykonaniu części analiz bioinformatycznych dotyczących analiz predykcji efektorów, podobieństwa do znanych rodzin, przewidywaniu funkcji białek kinaz w kontekście operonów, udziale w opracowywaniu wyników, udziale w pisaniu manuskryptu.

Podpis