Orientia tsutsugamushi: **analysis of the mobilome of a highly fragmented and repetitive genome reveals ongoing lateral gene transfer in an obligate intracellular bacterium.**

Suparat Giengkam[a], Chitrasak Kullapanich[a], Jantana Wongsantichon[a], Haley E. Adcox[b], Joseph J. Gillespie[c] and Jeanne Salje[a, d #]

Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand[a] ;

Department of Microbiology and Immunology, Virginia Commonwealth University Medical Center, School of Medicine, Richmond, Virginia, USA[b];

Department of Microbiology and Immunology, School of Medicine, University of Maryland Baltimore, MD 21201[c];

Department of Pathology, Department of Biochemistry, Cambridge Institute for Medical Research, University of Cambridge, UK;

Public Health Research Institute, Rutgers the State University of New Jersey, Newark, USA[e]

[#] Address correspondence to Jeanne Salje: jss53@cam.ac.uk

Abstract (250 words)

The rickettsial human pathogen *Orientia tsutsugamushi* (Ot) is an obligate intracellular Gram-negative bacterium with one of the most highly fragmented and repetitive genomes of any organism. Around 50% of its ~2.3 Mb genome is comprised of repetitive DNA that is derived from the highly proliferated Rickettsiales amplified genetic element (RAGE). RAGE is an integrative and conjugative element (ICE) that is present in a single Ot genome in up to 92 copies, most of which are partially or heavily degraded. In this report, we analysed RAGEs in eight fully sequenced Ot genomes and manually curated and reannotated all RAGE-associated genes, including those encoding DNA mobilisation proteins, P-type (*vir*) and F-type (*tra)* type IV secretion system (T4SS) components, Ankyrin repeat- and tetratricopeptide repeat-containing effectors, and other piggybacking cargo. Originally, the heavily degraded Ot RAGEs led to speculation that they are remnants of historical ICEs that are no longer active. Our analysis, however, identified two Ot genomes harbouring one or more intact RAGEs with complete F-T4SS genes essential for mediating ICE DNA transfer. As similar ICEs have been identified in unrelated rickettsial species, we assert that RAGEs play an ongoing role in lateral gene transfer within the Rickettsiales. Remarkably, we also identified in several Ot genomes remnants of prophages with no similarity to other rickettsial prophages. Together these findings indicate that, despite their obligate intracellular lifestyle and host range restricted to mites, rodents and humans, Ot genomes are highly dynamic and shaped through ongoing invasions by mobile genetic elements and viruses.

Keywords (3-10)

Orientia tsutsugamushi, Rickettsiales, obligate intracellular bacteria, intracellular pathogens, mobile genetic elements, comparative genomics, bacteriophage, integrative and conjugative elements, lateral gene transfer.

Importance

Obligate intracellular bacteria, or those only capable of growth inside other living cells, have limited opportunities for horizontal gene transfer with other microbes due to their isolated replicative niche. The human pathogen *Orientia tsutsugamushi* (Ot), an obligate intracellular bacterium causing scrub typhus, encodes an unusually high copy number of a ~40 gene mobile genetic element that typically facilitates genetic transfer across microbes. This proliferated element is heavily degraded in Ot and previously assumed to be inactive. Here, we conducted detailed analysis of this element in eight Ot strains and discovered two strains with at least one intact copy. This implies that the element is still capable of moving across Ot populations and suggests that the genome of this bacterium may be even more dynamic than previously appreciated. Our work raises questions about intracellular microbial evolution and sounds an alarm for gene-based efforts focused on diagnosing and combatting scrub typhus.

Introduction

Orientia tsutsugamushi (**Ot**) is an obligate intracellular Gram-negative bacterium that is a symbiont of trombiculid mites and causes the vector-borne human disease scrub typhus. Ot is a member of the alphaproteobacterial order Rickettsiales, which contains three well-studied families: Anaplasmataceae, Rickettsiaceae and Midichloriaceae[1,2], as well as four lesser-known families that have recently been described (Deianiraeaceae, Mitibacteraceae, Gamibacteraceae, and Athabascaceae)[3-5]. As a lineage within Rickettsiaceae, genus *Orientia* also includes "*Candidatus* Orientia chiloensis", which has recently been identified as an endemic species in Chile[6], and Candidatus *O. chuto*, which was isolated from a patient in Dubai[7]. There is extensive strain diversity within the Ot species, which can be found in rodents, mites and human patients across Southeast Asia. While strain diversity corresponds to differences in virulence in patients and in animal infection models, the molecular basis of these differences in virulence are not well understood. Ot strains are often classified according to serotype groupings, which are organised based on the human serological response to the highly antigenic surface protein TSA56. Major serotype groups are named after type strains and include Karp, Kato, Gilliam, Japanese-Gilliam, TA763, Saitama, Kuroki, Kawasaki and Shimokoshi.

At around 2-2.5 Mb, the genome of Ot is almost double the size of most Rickettsiales genomes and is one of the most fragmented and repetitive bacterial genomes reported to date[8,9]. With almost 50% of the genome comprised of repetitive DNA sequences, many experimental approaches are challenging: i.e. primer design, gene and genome sequencing, gene prediction and annotation, and comparative genomics. Complete genome sequences of two strains, Boryong and Ikeda, were published in 2008 using short read sequencing and bacterial artificial chromosome cloning[8,9]. An additional six strains (Karp, Kato, Gilliam, TA686, UT76, UT176) were fully sequenced in 2018 using long read PacBio technology[10]. A comparison of these eight genomes enabled the identification of 657 core genes and an open pangenome that is heavily characterized by gene duplication and pseudogenisation rather than the import of novel genes[10].

The Ot genome is dominated by an **integrative and conjugative element** (**ICE**), called **Rickettsiales amplified genetic element** (**RAGE**)[8,9,11], that has proliferated rampantly throughout the genome and is present in over 70 copies. RAGEs encode numerous repeated and pseudogenized genes, as well as single copy cargo genes that appear to be important for bacterial growth and pathogenesis. Whilst the high number of RAGE copies in the Ot genome remains unmatched, similar RAGEs have been described in several other *Rickettsia* species: *Rickettsia bellii* (single intact copy[12]) , *Rickettsia buchneri* (7 complete or near-complete genomic copies, and two plasmid encoded copies)[11], *Rickettsia massiliensis* (single intact copy[13]), *R. parkeri* str. Atlantic Rainforest (single intact copy[14]) , *R. felis* str. LSU-Lb (one plasmid encoded copy[15]) and *R. peacockii* (one partially degraded copy)[16]. In Ot, the ICE has not been controlled by the bacterial host and the RAGE has replicated to high levels in the genome[9,10,17]. The reasons for the differential fate of the RAGEs in Rickettsiaceae are unknown, but the small effective population size as well as the presence of population bottlenecks in obligate intracellular bacteria likely explain why it has had the ability to proliferate without strong negative selection in at least some rickettsial species.

Here, we present a thorough Ot phylogenomics analysis and re-annotation of the RAGEs and their cargo genes in eight strains: Gilliam, Boryong, UT76, UT176, Karp, Kato, Ikeda and TA686. We delineate the start and stop sites of all the intact and degraded RAGEs, allowing us to identify **inter-RAGE** (**IR**) regions with conserved clusters of genes. We further describe complete RAGEs in two Ot genomes (Kato and Gilliam). Finally, we annotate and classify the genes associated with DNA mobilisation and divergent **type IV secretion systems** (**T4SSs**), as well as the numerous multicopy cargo genes within the RAGEs, including those encoding **Ankyrin-repeat containing proteins** (**Anks**) and **tetratricopeptide repeat containing proteins** (**TPRs**), which are putative secreted effectors. This detailed disentangling of the superfluous RAGE-dominated mobilome from the core and accessory Ot genome is expected to enlighten research on Ot biology and overall genome evolution in obligate intracellular bacteria.

Results and discussion

RAGE and IR regions

The RAGE is an ICE that is present in Ot and certain other Rickettsiales genomes. The degree of amplification and degradation of RAGE in the Ot genome is so extensive - making up around 50% of the Ot genome in 71-93 distinct genomic regions – that the beginning and end sites of RAGEs cannot be easily identified by visual inspection. Here, we established objective criteria for the classification of genes into Ot RAGEs, and manually delineated each RAGE in the eight Ot strains in our study.

First, one or more copies of each of the following mobilisation genes must be present: integrase (*int)*, transposase (*tnp)*, F-type T4SS genes (*tra/trb)*, and relaxosome (*tra*). Second, one or more previously defined cargo or regulation genes[9] must be present: membrane proteins (reclassified here as Ot_RAGE_membrane protein, see below), DNA adenine methyltransferase (*dam*), DNA helicase, ATP-binding proteins (*mrp*), histidine kinases, SpoT-related proteins (synthetase and/or hydrolase domains), HNH endonuclease, peroxiredoxin, Anks, and TPRs. Third, the RAGE region begins with the first RAGE-associated gene and continues until a previously defined core gene[10] is reached. A run of core genes is classified as an IR element. Fourth, given the abundance of genes encoding **hypothetical proteins** (**HPs**) within RAGEs, those located between mobilisation/cargo RAGE genes are classified as being part of that RAGE. However, HP-encoding genes located between RAGE mobilisation/cargo genes and Ot core genes that cannot be resolved as being within RAGE or IR regions are classified as isolated HP-encoding genes. Fifth and final, single or multiple mobilisation or cargo RAGE genes are classified as isolated mobile genes or cargo genes, respectively. Some new cargo genes were identified by virtue of residing within RAGE regions in most or all genomes and these are discussed below.

In this way the entire genome of each Ot strain was classified into the following regions: RAGE region, IR region, isolated HP-encoding gene, isolated mobilisation gene, and isolated cargo gene (**Supplementary Dataset 1**). Using these criteria, we identified 71-93 RAGEs in the eight analysed genomes (**Fig. 1A, B**). The patterns of RAGE fragmentation and pseudogenization varied extensively between strains and it was not possible to map RAGEs between strains (**Fig. 1A**). This implies that RAGEs entered Ot strains one or more times as intact elements and subsequently underwent replication, pseudogenisation and recombination in independent trajectories.

Most RAGEs in Ot are degraded, both in terms of either completely lacking RAGE-associated genes or retaining genes that have been truncated or fragmented into predicted pseudogenes. A previous study found that the Ot strain Ikeda genome lacked any complete RAGES[9]. We assessed whether any of the strains in our analysis encoded complete RAGEs, defined as containing a full set of mobilisation genes and additional cargo genes as outlined in previous analyses[9]. All the strains in our analysis, with the notable single exception of Ikeda, encoded one or more complete set of RAGE genes (**Fig. 1B**). However, most of those RAGEs contained one or more mobilisation genes that were truncated. Accordingly, we carried out sequence alignments to define each RAGE gene as being full-length, truncated (containing one or more identifiable domains) or degraded (containing no identifiable complete domains). We then assessed whether any strains contained complete RAGEs with intact, full-length genes (**Fig. 1B**). We found that two strains, Gilliam and Kato, encoded complete RAGEs with full length mobilisation genes (**Fig. 1C**). This suggests that these strains may have obtained these elements recently and that they may be capable of mobilisation. ICEs normally have a preferred integration site, often within tRNA genes[11], as observed for *Rickettsia* species[18], Despite our discovery of complete RAGEs in these Ot genomes, no identifiable integration sites could be determined.

We previously used RNA sequencing analysis and comparative genomics to show that, despite the lack of synteny between Ot strains driven by the prolific RAGEs, small groups of proximate genes were transcribed at similar levels and maintained synteny across strains[19]. This demonstrates selection for gene order at the local level despite it being absent at a global level across Ot genomes. To further identify gene groups evolving under strong selective constraints relative to superfluous RAGEs, we analysed the IR regions, which harbour the majority of core Ot genes (**Fig. 1A**). Remarkably, this revealed 84 IR regions, ranging in length from 2 to 27 genes, most of which were conserved across all strains (**Fig. 1B**). Identification of these conserved IR gene groups illuminates highly conserved microsynteny that may encompass functionally linked genes sharing expression and/or regulatory programs.

Single copy cargo genes

The delineation of Ot genomes into RAGE and IR regions enabled us to better characterize RAGE cargo genes (**Fig. 2A, B**). In addition to the group of highly replicated multicopy cargo genes already described as RAGE components (discussed below), we identified numerous single copy genes previously overlooked for their occurrence within RAGEs (**Fig. 2A**). These include genes involved in fundamental processes of bacterial physiology and metabolism, e.g., tyrosine tRNA ligase (*tyrS),* RNA polymerase subunit omega (*rpoZ),* and the ClpP protease (*clpP)*, as well as genes encoding predicted secretory effectors likely involved in interactions with host cells, including phospholipase D (*pld)* and autotransporter proteins ScaA and ScaC (*scaA, scaC*) in all genomes, and ScaB (*scaB,* Boryong), ScaF (*scaF,* TA686) and ScaG (*scaG,* TA686) in individual strains. As many of these single copy genes have orthologs in other bacterial species that lack the RAGEs, it is likely that they were not introduced by mobile genetic elements. Rather, their current presence within RAGEs indicates they were probably incorporated into RAGE via recombination. However, a case-by-case basis may reveal certain conserved genes shuttling between Ot genomes via RAGE mobilisation. For instance, despite their conservation in all *Rickettsia* genomes, genes encoding secreted effectors and metabolite transporters were previously found piggybacking on RAGEs in the *R*. *buchneri* genome, illustrating the ability for RAGE to shuttle rickettsial genes important for the obligate intracellular lifestyle[11].

Highly abundant multi-copy cargo genes

Analysis of RAGE-associated cargo genes revealed 16 genes or (gene groups) present in numerous copies in all eight Ot genomes (**Fig. 2B**). Gene groups included membrane proteins, Dam DNA methyltransferases, DNA helicases, **multidrug resistance proteins** (**MRP**) and histidine kinases, SpoT hydrolase and synthetases, hypothetical/uncharacterized genes, mobile genetic elements (i.e., insertion sequences, transposases, integrases, and reverse transcriptases), Anks, TPRs, and *vir*- and *tra*-type T4SS genes. All of these, except *vir*-type T4SS genes, have been identified as RAGE cargo genes in previous studies[8,9,20,21]. For each gene group we assessed (i) whether all the genes annotated as belonging to this category were paralogs of the same gene or whether multiple distinct genes were present within one group, and (ii) whether some or all genes within a group were truncated and not able to form a full-length protein and, where a functional domain was known, whether this domain was present or not. Genes involved in DNA mobilisation, effector proteins, and T4SS genes are discussed in dedicated sections below, whilst other multi-copy cargo genes are discussed here.

Membrane proteins. The eight Ot genomes encode 21-41 RAGE associated genes annotated as membrane proteins (**Fig. 3A, Supplementary Dataset 2**). Analyses revealed that each Ot strain encodes exactly one copy of three genes encoding proteins with analogy to characterized membrane proteins: the YccA modulator of protease FtsH, vitamin transporter Vut1 and a gene similar to the rhamnose transporter RhaT. The remaining 18-38 genes encode paralogs of a gene we call Ot_RAGE_membrane protein, ranging in length from 90 to 663 bp. This protein lacks homology to any non-Ot genes and no known domain could be identified. Thus, the function of this gene in Ot is unknown.

DNA methyltransferases. Ot genomes encode 18 to 34 genes with similarity to DNA adenine methyltransferase (*dam*), all of which are located within RAGES (**Fig. 3B**). Sequence alignments demonstrate that 8-26 are full length proteins, defined as being equal in length to *E. coli dam* and encoding all seven known domains. The Ot genomes encode an additional 2-26 truncated *dam* genes where some domains are preserved, and fewer degraded copies with no identifiable domains. It is not known if these genes are functional, although previous studies[19,22] showed that they were not detected by proteomics analysis. They may have a specific role in methylation of RAGE during mobilisation and/or integration to protect from deleterious effects of single-stranded DNAse activity.

DNA helicases. DNA helicases unwind double stranded DNA and function in DNA and RNA metabolism, with general roles in DNA replication, repair and recombination. We found that all strains of Ot encode exactly two full length copies of the DNA helicase UvrD, which is involved in DNA repair (**Fig. 3C**). One copy is located within a RAGE whilst the second is located within IR82. The DnaB family of helicases, by contrast, is more numerous and degraded in Ot genomes (**Fig. 3C**). This helicase is involved in DNA replication and is present in 36-52 copies, with 1-23 being full length and mostly located within RAGEs. Each genome encodes one full length copy located at the interface of IR46 and IR47, which is likely the ancestral non-RAGE paralog involved in genome replication. Other copies are undoubtedly associated with RAGE mobilisation. Our previous proteomics analysis[19,22] detected expression of a DnaB gene, but due to sequence similarities between numerous paralogs, it was not possible to determine which specific gene(s) was expressed. Given its role in DNA replication, however, it is expected that at least one paralog would be expressed and functional.

MRPs and histidine kinases. We identified around 100 genes in the Ot genomes that were annotated as MRPs or histidine kinases (**Fig. 2B**, **3D**, **Supp. Fig. 1, Supplementary Dataset 3)**, or were found to contain histidine kinase domains (e.g. the sodium/proline symporter PutP). MRPs are members of the **ATP-binding cassette** (**ABC**) transporter protein family and were annotated as MRPs based on the presence of a **histidine kinase ATPase domain** (**HATPase**). Our analysis determined that two “MRP” proteins were distinct from all the other HATPase domain containing proteins in Ot: an MRP/NBP35 family ATP-binding protein and an ABC-membrane and AAA ATPase protein, both single copy in each genome. The former was located within an IR region (IR9) and the latter within RAGEs. Analysis of the remaining genes annotated as MRP/histidine kinases led us to identify several full-length orthologs of **two-component system** (**2CS**) histidine kinase genes. These include one histidine phosphotransferase gene (which does not contain an HATPase domain), one large hybrid sensor histidine kinase/response regulator gene and two 2CS histidine kinase genes (**Fig. 3D**). All were present in the same copy number and locations in each genome. Histidine phosphotransferase and the two 2CS sensor histidine kinase genes were consistently found in IR regions IR43, IR47, and IR63, respectively, whilst the hybrid sensor histidine kinase/response regulator gene was located within a RAGE. In our UT76 proteomics dataset[22], both the histidine phosphotransferase and the hybrid sensor histidine kinase/response regulator genes were detected, whilst the two sensor histidine kinase proteins were not (**Supplementary Dataset 1**). We also identified two copies of a histidine kinase domain containing sodium/pantothenate symporter, PanF, present in IR11 and IR55 regions in each genome, as well as a sodium/proline symporter, PutP, present in 4-8 copies and distributed into both RAGEs and IRs. Analysis of our previous proteomics dataset showed that PanF was expressed in strain UT76 as was one copy of PutP located in IR49[22] (**Supplementary Dataset 1**). The remaining MRP/histidine kinase genes were paralogs of one another, containing an HATPase domain and being present in 45-113 copies per genome with various degrees of truncation (**Supp. Fig. 1**). We classified these as degraded

HATPase domain containing proteins when an intact HATPase domain could no longer be detected due to the short length of the gene (**Fig. 3D, Supp. Fig. 1, Supplementary Dataset 1, 3**).

SpoT stringent response regulators. SpoT is a bifunctional synthetase/hydrolase that is essential for inducing and regulating the stringent response in *E. coli* and other bacteria through mediating intercellular levels of alarmone, or (p)ppGpp[23,24]. Ot genomes encode 36-78 genes with homology to SpoT. We identified exactly one full length SpoT gene, present in an IR region (IR46), encoding both synthetase and hydrolase domains in all Ot genomes (**Fig. 3E**). This gene was the only SpoT homolog found to be expressed in our previous proteomics analysis and was shown to be upregulated in extracellular Ot, consistent with a role in transitioning between different bacterial states[22]. We also identified exactly one gene in each genome that encodes only the SpoT synthetase domain in addition to a long C-terminal domain of unknown function. We then identified 15-46 SpoT genes that lacked the synthetase domain yet encoded the intact hydrolase domain as well as a further 16-44 SpoT genes that were truncated or degraded such that a functional hydrolase domain was no longer present.. Finally, we identified 1-4 genes in some genomes in which hydrolase domains are fused to other genes. Most rickettsial genomes harbour 6-12 SpoT genes, with some of the abovementioned architectures present (data not shown). Curiously, bifunctional (complete hydrolase and synthetase domains) genes are typical of most other Rickettsiales species, though not common in *Rickettsia* species and absent from notable human pathogens (e.g., *R*. *prowazekii*, *R*. *typhi*, *R*. *rickettsii*, and *R*. *conorii*)[11]. Still, the tendency for all Rickettsiales genomes to retain numerous single domain SpoT genes, even when RAGEs are absent, implies their function in some aspect of the stringent response. The presence of such drastic numbers and diverse architectures of SpoT genes in Ot genomes relative to other rickettsial species is intriguing and deserving of future investigation.

HPs. The Ot strains harbour 308 to 547 genes per genome that are annotated as hypothetical or uncharacterized, of which about half are located within RAGEs (**Fig. 2B** and **3F and Supplementary Dataset 4**). We determined whether all the RAGE hypothetical/uncharacterized genes were paralogs of a single RAGE gene or if they encoded multiple different genes. Sequence alignments for all the genes annotated as hypothetical or uncharacterized in the Karp genome were performed, which revealed that the genes clustered into 24 groups (**Fig. 3F**) with 18 of these encoding genes carrying known protein domains. Those without known domains were named RAGE_hypo_Gr1-7, with groups 1-6 encoding a domain of unknown function, and group 7 combining all remaining HP genes with no identifiable known domains. Three of these genes with known domains were present in exactly one copy in all genomes and encode: a phage portal protein, a zinc ribbon domain, and a rhodanese homology domain. Another single-copy gene found in all eight Ot genomes carries a domain of unknown function (DUF155). Furthermore, hypothetical genes containing a DnaA N-terminal domain were identified in 10-23 copies in all Ot genomes. In our prior study, one full length paralog, located in IR1 was expressed in UT76, whilst others were not detected[22] (**Supplementary Dataset 1**).

Other hypothetical/uncharacterized genes were distributed sporadically amongst the genomes. In order to get a sense of the distribution of the remaining RAGE-associated hypothetical genes that were not clustered into conserved groups, we analysed the remaining hypothetical genes in the RAGEs of the Karp genome only. There were no identifiable domains in any of these and the diversity was such that it was not possible to bin them into homologous groups. We annotated them all as belonging to a large and divergent 25th group (RAGE_hypo_Gr7). While little can be inferred about the function of these hundreds of genes, it is likely that at least some of these play important roles in the biology of Ot.

Putative effectors piggybacking on Ot RAGE

*Anks***.** The ankyrin repeat is one of the most common protein folds in nature, being widespread in eukaryotes and pervasive in many viruses and host-associated bacteria[25,26]. Ankyrin repeats are used to mediate a myriad of protein-protein interactions, and host-associated prokaryotes and viruses frequently express Anks to hijack or subvert host cell pathways that would be detrimental or beneficial to their survival[27]. Previously, several Ot Anks were shown to be secreted via the rickettsial **type I secretion system** (**T1SS**)[28]. Certain Ank effectors have been functionally characterized in strain Ikeda and shown to play important roles in host cell interactions[28-32]. However, a major challenge in comparing the host-pathogen cell biology of different Ot strains has been the difficulty assessing which Anks are most similar to those in other strains. This is important for determining the significance of Anks as species- versus strain-specific effectors underlying pathogenesis.

We defined a set of criteria for clustering Anks, with their subsequent characterisation within each Ot genome following the well described Ank repertoire of strain Ikeda [9]. We identified several new Ank groups in Ikeda, although some of these lack complete Ank repeats and are likely non-functional **Supplementary Dataset 5**). Our comparative analysis indicates Ot strains encode 47-66 Anks, with variability (67-94%) in the number of common vs. strain-specific proteins per genome (**Fig. 4A**). Ot Anks often harbour a single F-box domain, which are prominently known components of SCF (Skp1, Cullin1, F-box) ubiquitin ligase complexes but recent studies have described their participation in non-SCF protein-protein interactions involved in diverse eukaryotic functions?pathways?[33]. F-box-resembling PRANC (pox protein repeats of ankyrin-C-terminal) domains and coiled-coils were less frequently predicted.

Of the 54 orthologous groups of Ot Anks, all genomes were found to encode at least one copy of seven groups: Ank03, Ank08, Ank10, Ank11, Ank12, Ank20 and Ank24 (**Fig. 4B**). Ank03 is by far the most prominent Ank, being present in 4 to 32 copies in Ot genomes, with the other six families present in 1 to 4 copies (**Fig. 4B**). Curiously, while most Ot Anks are predominantly found within RAGEs, Ank20 is encoded in an IR region (IR84) in all analysed Ot genomes. Collectively, these seven Anks likely carry out essential functions in Ot biology. However, each strain likely utilizes unique Ank arsenals throughout its lifecycle given that some of the less conserved Anks have characterized functional roles; e.g, Ank01 and Ank06 of Ot str. Ikeda modulate NFkB transport to the nucleus[31]. As such, it is likely that there is functional redundancy between the Ank groups, with some of the 100 other Ank groups not found in Ikeda functioning similarly as Ank01 and Ank06 in genomes lacking these genes.

TPRs. The tetratricopeptide repeat is another protein motif that is commonly used in mediating inter-protein interactions, typically found in subunits of multi-protein complexes[34]. TPRs are widespread in Ot proteins (**Fig. 4C**), albeit with a lower number of copies per genome than Anks. Ot TPRs have been less characterised than the Ot Anks, with only one report demonstrating a role in inhibition of eukaryotic translation in Ot strain Boryong[35]. We compiled 21-48 TPRs per Ot genome and classified them into nine groups primarily based on within-protein location of tetratricopeptide repeats (**Fig. 4C, Supp. Fig. 2, Supplementary Dataset 5**). Whilst the positions were conserved within groups, the number of repeats was variable and indicated expansion and contraction of repeats, as well as processive gene degradation within each group (**Fig. 4D**). The prediction of SEC signal peptides in certain TPRs indicates at least some of these putative effectors may be secreted to the periplasm with possible translocation across the outer membrane, possibly via TolC as proposed for the RARP-1 effector of *R. typhi*[36]. Still, the lack of N-terminal secretion signals in most TPRs indicates other possible routes for TPR secretion that await characterisation.

Mobile genetic elements associated with RAGE

Integrases. ICEs, such as Ot RAGE, encode integrase genes to catalyse genomic integration, and conjugative genes (discussed in the next section) to catalyse horizontal gene transfer[37]. The Ot genomes analysed in this study encode 58-102 integrase genes

(**Fig. 5A**), of which only 3-13 per genome remain full length, consistent with progressive degradation of the Ot RAGE. Where present, the integrases are located at the start position of a RAGE (**Fig. 1C**). However, several integrase genes were located as isolated genes outside RAGE regions, reflecting the high mobility of these genes and the overall high recombination rates in Ot genomes.

Transposable elements. In addition to the OtRAGE, the Ot genome encodes two other types of transposable elements[9]: retrotransposons (group II introns) and DNA transposons. Whilst these are independent mobile genetic elements, they have been incorporated into the Ot RAGE regions, and it is likely that the different mobilizable elements impact each other's activity. Group II introns are self-splicing retrotransposons that catalyse their integration into genomes via an RNA intermediate, using an intron-encoded reverse transcriptase protein[38]. The Ot genomes encode 7-64 group II intron reverse transcriptase genes, although only UT76, Karp and TA686 encode full length genes, with the others being heavily degraded (**Fig. 5A**). All full-length reverse transcriptase genes were immediately followed by an HNH endonuclease gene likely required for catalysis.

Ot encodes several families of DNA transposons which have been previously classified in strain Ikeda[9]. This class of mobile elements is comprised of a transposase gene flanked by inverted repeat regions on either side, which together make up an **insertion sequence** (**IS**)[39]. Numerous families of IS have been identified in other bacteria. Given the large number of IS elements in each Ot genome and their highly degrative tendency, we selected two of the many frequently occurring IS genes, ISOt3 and ISOt5, to characterise in detail across all eight genomes (**Fig. 5A**). These were present in 0-120 (ISOt3) and 0-70 (ISOt5) full-length copies across the genomes. We also analysed the complete set of IS elements in one strain, Karp, and compared these with those previously predicted for str. Ikeda (**Fig. 5B**). An example of the analysis of one IS element in Karp, ISOt1, shows the distribution of full length and degraded copies typical of IS families in all Ot genomes (**Fig. 5C**). We identified the same set of IS elements that had previously been described in Ikeda[9] (**Fig. 5B, C**). We followed the classification and nomenclature established in Nakayaka et al[9], in which mISOt1, mISOt2 and mISOt4 denotes "miniature" versions of elements containing the same terminal inverted repeat sequences as ISOt1, ISOt2 and ISOt4 respectively. Within the nine IS classes found in Karp, most were heavily degraded with some, such as IS630 family transposase ISOt3, having no remaining full-length elements. We identified an additional seven groups of transposase genes in Karp that were not part of IS elements (**Fig. 5B**).

Bacteriophages are another source of horizontally transferred genetic material. These are thought to be rare in obligate intracellular bacteria due to the isolated lifestyle, although there are exceptions such as the WO prophage that is widespread in *Wolbachia* populations[40]. We searched for the presence of prophages in the Ot genomes using the online search tool PHASTER[41,42] and identified remnants of prophage genetic material in all strains except for Boryong (**Fig. 5D**, **E**; **Supplementary Dataset 7**). By contrast, no prophage regions were identified in *Rickettsia conorii, Rickettsia rickettsia, Rickettsia prowazekii, Anaplasma phagocytophilum* or *Anaplasma marginale,* although two sites were detected in both the genome of *Wolbachia* endosymbiont of *Drosophila melanogaster* and *Rickettsia bellii.* This suggests that prophages are not universally circulating in Rickettsiales populations, but are present in selected species such as Ot, wolbachiae and *R. bellii.* Whilst many of the Ot prophage genes identified by PHASTER include transposase and integrase genes, which may be of ICE origin rather than phage origin, phage-specific genes including capsid and envelope proteins were also found. In addition to the identification of potential prophage regions found by PHASTER, isolated phage-related genes, such as the phage portal protein previously annotated as a hypothetical protein (Fig. 3F), are also present in the Ot genomes. Sequence similarity searches indicated low similarity to a range of diverse phage sequences from free-living bacteria indicating that either the prophages came from numerous sources, or that the sequences within each strain were sufficiently degraded so they have lost identifiable homology to one another.

RAGE mobilisation genes

*F-T4SS***.** The RAGE encodes a conjugative T4SS highly similar to the F-T4SS of the archetypal F plasmid of *E. coli* (*tra/trb)*[8,9]. Previous comparisons of the RAGE T4SS with that of the *E. coli* F plasmid showed that it encodes 14 proteins predicted to form the T4SS scaffold, some of which are analogous to components within P-type T4SSs[18,43] (**Fig. 6A, B, Supplementary Dataset 8**). While syntenic to the *E*. *coli tra*/*trb* T4SS, the RAGE T4SS lacks genes involved in the regulation of conjugation, as well as other assembly factors and lytic transglycosylases (**Fig. 6C**). In this way, the RAGE T4SS is a streamlined version of the canonical F-T4SS. The Ot RAGE T4SS is also highly similar in gene order and composition to F-T4SSs characterized in the RAGEs of *R. buchneri*[11]*, R. bellii*[12]*, R. felis*[15] and *R. massiliae*[13]. As with these prior reports, we also did not identify a gene encoding a pilin protein (typically TraA in F-T4SSs) in RAGE T4SSs, though it may be that a pilus is synthesized using a different pilin gene since the RAGE-harbouring *R*. *bellii* forms large pili during host infection[12]. Experiments are needed to determine if the RAGE T4SS elaborates a pilus or functions pilus-less, as is noted for the P-T4SS of *Rickettsia* species[43], *Neorickettsia risticii*[44], and likely all Rickettsiales[45]. Another common peculiarity of these F-T4SSs is the split gene encoding TraK, the significance of which is unknown.

*Relaxosome***.** The relaxosome of the *E*. *coli tra*/*trb* F-T4SS encodes one multifunctional relaxase, TraI, which excises and binds single-stranded plasmid DNA[46]. In contrast, the Ot RAGE carries three genes *traI*, *traA$_{Ti}$* and *traD$_{Ti}$* predicted to comprise the relaxosome that mobilises RAGE (**Fig. 6C**). *E*. *coli* TraI harbours four distinct domains required for nicking, binding, and unwinding DNA. By contrast, Ot TraI lacks a domain for nicking DNA and shares very limited similarity to *E*. *coli* TraI. However, Ot TraA$_{Ti}$ carries a MobA-like domain that cleaves single- and double-stranded DNA at specific sites[47]. Curiously, all of the domains encompassed by Ot TraI and TraA$_{Ti}$ proteins are found in a single *Rickettsia* RAGE protein, named TraA$_{Ti}$-I, which is highly similar to both Ot TraI and TraA$_{Ti}$ but shares limited similarity to *E. coli* TraI. As their annotation indicates, RAGE $\mathrm{TraA_{Ti}}$ and $\mathrm{TraD_{Ti}}$ are similar to relaxosome proteins of plasmid Ti of *Agrobacterium tumefaciens*, TraA and TraD, which are required for T-DNA translocation into plant cells via the *vir* T4SS[48]. The significance of different relaxosome structures between Ot and *Rickettsia* RAGEs is unclear, although $traA_{Ti}$ and $traD_{Ti}$ genes are common on *Rickettsia* plasmids even when RAGE are absent[49]. It may be multiple RAGE types exist in the rickettsial mobilome and are defined by their cognate relaxosomes. The presence of transposases flanking relaxosome genes in all complete Ot and *Rickettsia* RAGEs may also signify that RAGEs evolve by recombining different relaxosome cassettes into the conjugation and cargo genes.

Proliferation of Ot RAGE mobilisation genes. Aside from shared synteny and mobilisation gene composition, *Rickettsia* and Ot RAGEs have common insertion points for antidote genes of toxin-antidote modules and transposases (**Fig. 6C**). However, certain *Rickettsia* RAGEs have cargo genes inserted at different sites within the mobilisation genes[11,15,18]. Furthermore, a recent study annotated a RAGE from the *Tisiphia* endosymbiont of *Cimex lectularius* that harbours unique mobilisation genes and cargo gene insertion sites[50]. This indicates that RAGEs are far more diverse and widespread across Rickettsiales than previously appreciated. Still, most *Rickettsia* genomes either lack RAGE entirely or show minimal evidence for RAGE insertion near a common genomic position, $\mathrm{tRNA^{Val\text{-}GAC}}$ [11]. This is in stark contrast to the proliferated nature of RAGEs in Ot genomes.

The scattershot distribution of RAGE in Ot genomes is particularly evinced by the mobilisation gene clusters that are present in numerous copies within the plethora of RAGEs. We find that over 50% of the RAGE T4SS and relaxosome genes are present as truncated pseudogenes, and that some of these clusters encode only a subset of the 18 possible Ot RAGE mobilisation genes (**Fig. 6D, E**). Given the high degree of pseudogenization, we sought to examine whether any strain encoded any RAGE mobilisation gene clusters containing a complete complement of full-length genes. We found that Karp, Kato, Gilliam and UT76 encoded at least one complete RAGE mobilisation gene set, whilst Ikeda, Boryong, TA686 and UT176 did not (**Fig. 6D, E**, **Supp. Fig. 3, Dataset 7**). There was a positive correlation between strains containing complete sets of RAGE mobilisation gene sets and the total number of full-length mobilisation genes (**Fig. 6D, E**). Moreover, the complete RAGE mobilisation gene clusters in Gilliam and Kato were located within complete RAGE regions (**Fig. 1C**, **Fig. 6E**, **Supp. Fig. 3**). Whilst several genomes lack complete RAGE mobilisation gene clusters, all except Boryong encode at least one full length copy of each RAGE mobilisation gene, albeit not in a contiguous cluster. Therefore, it is possible that all strains except Boryong could assemble a functional F-type T4SS competent to mediate transfer of RAGEs.

The impact of pervasive mobile genetic elements on the Ot genome

P-T4SS. Like other rickettsial species, Ot encodes a P-type T4SS related to the archetypal *vir* T4SS of the pTi plasmid of *A. tumefacians*. Relative to *vir*, this **Rickettsiales *vir* homolog *(rvh)*** T4SS has distinct features, including the scattered distribution of *rvh* gene clusters, duplication of *rvhB8*, *rvhB9* and *rvhB4,* 3-5 copies of *rvhB6*, and no gene encoding an equivalent to the VirB5 minor pilin subunit[51,52,45,51](**Fig. 7A,B**). These characteristics are nuanced: 1) RvhB4-II, RvhB8-II, and RvhB9-II carry atypical structural deviations from described VirB4, VirB8, and VirB9 family proteins, 2) RvhB6 proteins have large insertions flanking the VirB6-like membrane spanning region, and 3) a lack of a minor pilin subunit precludes formation of a T-pilus[52,53]. There is evidence that structurally different RvhB8-I and RvhB8-II proteins of *R*. typhi cannot dimerize[53], which led to the hypothesis that divergent duplications may autoregulate effector secretion[52]. The recent identification of *rvh* genes in all seven Rickettsiales families implies a highly important function[3,54]. Thus, we assessed the properties of the Ot *rvh* T4SS in the face of its rampant mobile genetic element-induced genome shuffling.

A single set of *rvh* genes is present in all the Ot genomes analysed here and has features resembling those described in other Rickettsiales species (**Fig. 6A-C**). Interestingly, Ot shares two key characteristics with the *rvh* T4SS of distantly-related Anaplasmataceae species as opposed to that of closely-related *Rickettsia* species. First, while Ot lacks genes for a VirB5 protein and therefore cannot form extracellular pili used for cellular attachment, it encodes 2-3 copies of *rvhB2*, the major pilus subunit. Given that Ot lacks **lipopolysaccharide** (**LPS**) on its surface, multiple RvhB2 proteins may act as divergent surface antigens in a similar fashion previously posited for Anaplasmataceae species, which collectively lack LPS biosynthesis genes and have multiple *rvhB2* genes throughout their genomes[45,55]. Second, Ot lacks any identifiable *rvhB1* gene, which is present only in *Rickettsia* spp. This gene encodes a lytic transglycosylase predicted to cleave **peptidoglycan** (**PGN**) to allow T4SS scaffold assembly[51]. Compared with *Rickettsia* spp., which synthesize a canonical PGN layer[56,57], the presence of a minimal cell wall in Ot and Anaplasmataceae species is consistent with the absence of *rvhB1* from these genomes[58,59]. These collective differences in *rvh* T4SS architecture present clear convergent evolution in Ot and Anaplasmataceae species in the context of shared cell wall morphology and probable responses to host cell immune pressures[60].

Our analysis shows that the identities of six *rvh* gene clusters are conserved across Ot strains, whilst the genomic positions of the clusters vary between strains (**Fig. 7D**). Clusters 1 (*rvhB7*, *rvhB8-I*, *rvhB9-II*, *rvhB10*, *rvhB11*, *rvhD4*), 2 (*rvhB6e*) and 4 (*rvhB4-II*) are located within RAGEs in all strains, with the other *rvh* genes consistently located in IR regions except for cluster 6 and clusters 3 and 6 in UT176 and TA686, respectively (**Fig. 7D, F**). Analysis of published datasets of proteomics and RNAseq data in Karp and UT76[19,22] show that RvhB2-3 and RvhB7 proteins are not detected under growth conditions used in those analyses, although transcription levels of *rvhB2-3* are high (**Fig. 7E**). All the other Rvh proteins are detected in UT76 and most are detected in Karp. The UT76 dataset compared peptide levels in two different bacteria populations: **intracellular bacteria** (**IB**) and

extracellular bacteria (**EB**)[22]. The EB/IB ratio of some multi-copy Rvh proteins differs between paralogs; e.g., RvhB2-1, which is present at a ratio of 0.76 compared with the ratio of 0.19 for RvhB2-2. This suggests expression of these proteins may be differentially regulated, potentially reflecting functional differences. Our collective analyses indicate that Ot Rvh genes can form a functional P-T4SS, despite the pervasive mobile element-induced gene shuffling in Ot genomes. While no Ot *rvh* transported effector has been described to date, it is highly likely that Ot utilizes the *rvh* T4SS during host cell infection, as secreted proteins that interact with the T4SS gatekeeper, RvhD4*,* have been described for *R. typhi*[61-63], *R. rickettsii*[62,64], *A. marginale*[65], *A. phagocytophilum*[66-69 70-72]and *Ehrlichia chaffeensis*[68,73 74-76].

Ot lacks defence mechanisms against invasive DNA

We show here that the Ot genome is exceptional in its abundance of invasive mobile genetic elements, including ICEs, transposases, group II introns and prophages. Bacteria have evolved a range of anti-viral mechanisms to minimise damage caused by mobile genetic elements[77-79]. We therefore sought to assess if Ot lacks these protective systems, possibly explaining the proliferation of mobile genetic elements. We used DefenseFinder to carry out a systematic search for all known anti-phage systems including restriction modification systems, CRISPr/Cas systems, and toxin-antidote defence modules[78,80]. We found that none of the Ot strains in our study had any identifiable defence systems. Whilst it is possible that this is due to sequence divergence, small size (e.g., certain toxin-antidote modules) or systems that have not yet been discovered, the software was able to detect three different restriction modification systems and the newly described Pyscar defence system[81] in the closely related free living alphaproteobacterial *Caulobacter crescentus.* In addition to lacking identifiable antiviral defence systems, Ot also has limited homologous recombination capability, a system that is frequently used in antiviral defence [78]. Whilst Ot encodes RecA and the alternative homologous recombination pathway RecFOR, it lacks the major repair complex RecBCD that can defend against some mobile genetic elements by degrading linear double stranded DNA[10]. Overall, Ot lacks identifiable mobile genetic elements defence systems likely explaining the proliferation of mobile DNA in these genomes.

Conclusions

The identification of complete RAGEs in two Ot strains raises the possibility that these ICEs are active at the population level. Evidence for this hypothesis awaits whole genome sequencing of large numbers of Ot isolates beyond the total of eight currently available. Whilst only two genomes encode complete RAGEs with full length genes, all encode all the genes required for RAGE mobilisation, albeit in dispersed locations across the genome. Future research is needed to determine whether such mosaic RAGEs can be mobilised or not.

The identification of potentially active RAGEs in Ot raises the question of how they can be transferred between Ot organisms during their lifecycle. Ot is an obligate intracellular bacterium and therefore bacterial cells have limited interactions with other bacteria of the same or different species. It is possible that different strains of Ot infect the same host cell in a mite or a rodent during co-infection by two species. Albeit rare, this could occur with sufficient frequency to enable horizontal gene transfer between species. Alternatively, it is possible that the extracellular form of Ot retains sufficient residual metabolic activity to support lateral DNA transfer in the cell-free extracellular state. Mites typically feed in a tight cluster, for example on the ear of an infested rodent, and the co-feeding pool may provide the environment for close encounters between Ot cells in an intracellular or extracellular state to mediate conjugation. Finally, while not detected in other environments or hosts (i.e. protists), there could be other opportunities for Ot strains to exchange DNA or acquire DNA from other intracellular species.

In conclusion, this study has led to the manual re-annotation of the genomes of eight strains of Ot, enabling the delineation of RAGE and IR regions. Open questions remain. Importantly, whilst intact RAGEs have been identified in two strains, the dynamics of the Ot RAGE are completely unknown. It is also unknown whether the current set of RAGEs within one genome results from one or multiple invasion events. The Ot RAGE encodes an F-T4SS, but it is not known if these are active, nor what they transport beyond the ICE itself. Progress towards answering these questions will enable further insights into the biology and pathogenicity of this important human pathogen.

Figure Legends

Figure 1. Ot RAGE and IR elements. A. An overview of the genomes of eight Ot strains with genes classified into RAGE and IR regions. Numbers at left refer to Ot strains listed in panel B. Grey arrows = RAGE regions; colored arrows = IR regions. The colors correspond to conserved IR regions between strains and demonstrate the lack of synteny between Ot genomes. **B.** Table summarizing RAGEs, IR regions and isolated mobile genes, cargo genes and hypothetical genes that could not be classified into RAGE or IR elements. Ot strains are listed accordingly to a previously estimated phylogeny[10], with numbers corresponding to full genome maps in panel A. **C.** Organization of genes in the four complete RAGEs found in our analysis. t = truncated (at least one identifiable domain present); d = degraded (no identifiable domains present). Detailed analysis of the reannotation and classification of all genes in the eight genomes are given in **Supplementary Dataset 1**.

Figure 2. Single and multi-copy cargo genes encoded on Ot RAGEs. A. Single or low-copy cargo genes encoded on Ot RAGE. Summary statistics show whether genes are present in single or multiple copies on RAGEs in different strains, and also in single or multiple copies in IRs. The exact number of copies is given for each gene. Blue text = number of copies in IR; red text = number of copies in RAGE. **B.** Frequency and distribution of high copy cargo genes (both full length and truncated/degraded) within RAGEs in eight strains of Ot. Numbers in brackets denote additional copies in IRs.

Fig. 3. Analysis of high copy cargo genes on RAGE elements in Ot. A-C. Frequency and distribution of RAGE cargo genes annotated as (**A**) membrane proteins, (**B**) Dam DNA methyltransferases, and (**C**) DNA helicases. DnaB is a replicative DNA helicase and UvrB is a repair DNA helicase. **D**. Frequency and distribution of RAGE cargo genes encoding MRP/histidine kinases, with examples of His kinase divergent architectures. **E**. Frequency and distribution of RAGE cargo genes encoding SpoT stringent response regulators, with

examples of divergent architectures. The bifunctional SpoT protein is compared to the canonical SpoT protein of *E. coli*. **E**. Frequency and distribution of RAGE cargo genes encoding HPs. DnaA_N, N-terminal domain of DnaA; RHOD, rhodanese homology domain; AHH, adenosyl homocysteine hydrolase; MagZ, nucleoside triphosphate pyrophosphohydrolase; na/nt, nucleic acid/nucleotide deaminase; BrkB-like, YihY/virulence factor BrkB family protein; PDu(A)C, copper chaperone; CdAMP_rec, cyclic diAMP receptor proteins; Rvt_1 (PF00078), reverse transcriptase Pfam PF00078; Rvt_N 19, domain of reverse transcriptase Rvt_N; DUF, domain of unknown function.

Figure 4. Putative effectors piggybacking on Ot RAGE. A. Frequency and distribution of Anks in Ot genomes. Anks are broken down into orthologous groups (OGs, present in two or more genomes) or singletons (unique to a genome). CC, coiled coil; PRANC (Pox proteins Repeats of ANkyrin, C-terminal), domain found at the C terminus of certain Pox virus proteins; F-box, motif of approximately 50 amino acids that functions in protein-protein interactions. **B**. (*top*) graphical view of Ank OG strain representation (2-8 genomes). Roughly 25% of Ank OGs are found in five or more strains, with variable levels of conservation in copy number per genome. (*bottom*) Architectures for Anks present in all Ot genomes, with proteins from Ot strain Karp. **C.** Frequency and distribution of TPRs in Ot genomes. Nine ortholog groups contain all the TPR s across eight Ot genomes. **D.** Examples of diverse TPR architectures for six proteins from Ot strain Karp.

Figure 5. A diversity of mobile genetic elements associates with Ot RAGEs. A. Frequency and distribution of Ot RAGE-associated genes encoding integrases, Group II intron-associated reverse transcriptases, and IS elements ISOt3 and ISOt5. **B**. Frequency and distribution of IS elements in Ikeda and Karp strains. **C**. Alignment showing classification of ISOt1 elements as full length or degraded. Full length copies of ISOt1 in Karp are shown by red dotted box. **D**. Overview of prophage elements in Ot genomes as identified by

PHASTER search tool . Int = integrase, Tnp = transposase, Env = envelope, Cap = capsid, Pro = protease. **E**. Overview of predicted phage region in TA686.

Figure 6. Characteristics of the F-type T4SS and relaxosome proteins encoded on Ot RAGEs. A. Composition of the Ot RAGE F-T4SS in relation to the *Agrobacterium tumefaciens vir* P-T4SS and the *Escherichia coli tra/trb* F-T4SS from the F operon. Analogues across divergent T4SSs are coloured similarly, with other colours as follows: dark gray, RAGE T4SS proteins found in F-T4SSs but not P-T4SSs; white, *E. coli* F-T4SS scaffold genes not present in RAGE T4SSs; light gray, other *E. coli* F operon genes not present in RAGE. For relaxosome proteins (olive green), domains were predicted with SMART [82]. **B**. Theoretical assembly of the RAGE T4SS in relation to data from other F- and P-type T4SSs. The uncertain synthesis of a pilus is depicted (see text for details). **C**. Comparison of the *E. coli* F operon to mobilisation genes of complete RAGEs from Ot str. Gilliam and *Rickettsia bellii* str. RML369-C. This *E. coli* strain, K-12 ER3466 (CP010442), has the F operon on a chromosomal segment flanked by transposases (yellow circles). Red shading and numbers indicate % aa identity across pairwise alignments. Dashed lines enclose the relaxosome genes, whose protein domains are described in panel A. INT, integrase; LRR, leucine rich repeat protein. **D**. Frequency and distribution of of full length and truncated *tra*/*trb* genes in Ot strains. Complete circles, genomes containing full sets of *tra*/*trb* genes within one or more RAGE; open circles, no complete *tra*/*trb* gene sets. Numbers in parentheses: number of complete RAGEs/number of complete RAGE genes containing truncated genes/incomplete RAGEs. Details of truncated genes and gene fusions are given in Supplementary Datasets 1 and 8. **E**. Genomic location of *tra*/*trb* gene clusters in Ot str. Gilliam. Triangles and highlighting depict complete RAGEs. Bracketed TraE and TraA_{TI} are commonly occurring pseudogenized duplications. Green circles, complete gene; small black circles, predicted pseudogene; Xs, gene absent with *tra*/*trb* gene cluster.

Figure 7. Synopsis of Ot P-type (*vir*-like) T4SS genes. A. Description of the general *rvh* T4SS characteristics, summarized from prior studies[43,45,51,52] . **B**. Theoretical assembly of the RAGE T4SS in relation to data from other P-type T4SSs. There is no synthesis of a T-pilus (see text for details) **C**. Comparison of genes encoding *vir* T4SS in *Agrobacterium tumefaciens,* the archetypal P-T4SS, and those encoding the *rvh* T4SS in *Rickettsia typhi* and Ot. **D**. Arrangement of *rvh* genes in Ot genomes. Red genes are located in RAGE regions whilst blue are located in IR regions. **E**. Previously published RNAseq and proteomics data showing relative expression levels of *rvh* genes in strains UT76 and Karp. These are taken from Atwal et. al 2022 (UT76) and Mika-Gospodorz et. al 2020 (Karp). UT76 data shows relative peptide counts in intracellular bacteria (IB) and extracellular bacteria (EB). Karp data shows presence or absence of detectable peptides from proteomics analysis (+/-) and relative RNA transcripts from RNAseq data (TPM/transcripts per million). **F**. Distribution of *vir* genes across Ot genomes showing lack of conservation of absolute position, despite similarities in gene groupings as shown in **Fig. 7D**.

Methods

Table Methods 1: Accession numbers of genomes used in this study

Strains	Genome accession numbers	Links
Boryong	AM494475.1	https://www.ncbi.nlm.nih.gov/nuccore/AM494475.1
	NC009488.1	https://www.ncbi.nlm.nih.gov/nuccore/NC_009488.1
UT76	LS398552.1	https://www.ncbi.nlm.nih.gov/nuccore/LS398552
UT176	LS398547.1	https://www.ncbi.nlm.nih.gov/nuccore/LS398547.1
Karp	LS398548.1	https://www.ncbi.nlm.nih.gov/nuccore/LS398548.1
Kato	LS398550.1	https://www.ncbi.nlm.nih.gov/nuccore/LS398550.1
Ikeda	AP008981.1	https://www.ncbi.nlm.nih.gov/nuccore/AP008981.1
	NC_010793.1	https://www.ncbi.nlm.nih.gov/nuccore/NC_010793.1
TA686	LS398549.1	https://www.ncbi.nlm.nih.gov/nuccore/LS398549.1
Gilliam	LS398551.1	https://www.ncbi.nlm.nih.gov/nuccore/LS398551.1

Identification of RAGE and inter-RAGE regions in the genome of Ot

The boundaries of RAGE and IRs were manually delineated in each genome using defined criteria. First, groups of genes whose relative position to one another was conserved across strains were identified manually by comparing the genomes of 8 Ot strains (Ikeda, Boryong, Karp, Kato, Gilliam, TA686, UT76 and UT176). This led to the identification and numbering of IR regions.

RAGE regions were subsequently identified using criteria largely drawn from *K Nakayama et al*, 2008[9].

The element was classified as a "complete RAGE gene" if the sequences encoded a full-length integrase gene at the left end (N-terminus), a full-length transposase gene, full-length set of conjugative transfer genes (tra genes: *TraA, TraB, TraC, TraD, TraE, TraF, TraG, TraH, TraI, TraK, TraL, TraN, TraU, TraV, TraW*), and nonconjugative genes (RAGE associated cargo genes) including one or all of the following: SpoT-related proteins (ppGpp hydrolase, (p)ppGpp synthetase, SpoT synthase, and SpoT hydrolase), DNA methyltransferase, DNA helicase, histidine kinases, ATP-binding proteins (mrp), HNH endonuclease, membrane proteins, ankyrin repeat proteins, and hypothetical proteins. The RAGE associated cargo genes in “complete RAGE gene” can be either full-length or truncated genes.

The element was classified as an “complete RAGE with truncated genes” if the sequence encoded the same gene set as above, but where one or more of the integrase, transposase, or Tra conjugative transfer genes were truncated.

The element was classified as an “incomplete RAGE” if the sequence encoded integrase or transposases, and at least one RAGE associated cargo gene.

The “isolated mobile gene” was defined as encoding one or more integrase or transposases without RAGE associated cargo genes.

The “isolated cargo gene” was defined as encoding one or more cargo genes without transposases, integrases, or Tra genes.

The “isolated hypothetical protein” was defined as encoding one or more hypothetical proteins at the boundary of conserved IRs or RAGEs.

The presence of a *dnaA* gene was used as an indicator gene for defining the end of a RAGE element (Fig. Methods 1). However, the criteria could not be applied for all RAGE elements when hypothetical proteins and transposases are located at the end of RAGE masking the original *dnaA* terminus. In the first case (Fig. Methods 1A), RAGE elements are located next to each other in the same direction. RAGE is terminated when integrase gene of the next RAGE is found. In the second case (Fig. Methods B), RAGE elements are located next to each other in opposite direction and two *dnaA* genes are located next to the each other. In this case the RAGE is terminated at the *dnaA* gene which belongs to RAGE on the left (forward direction) and RAGE on the right (reverse direction). In the third case (Fig. Methods C), RAGE elements are located next to each other in opposite direction. Two RAGEs were combined into one RAGE if a *dnaA* gene was not present in either RAGE.

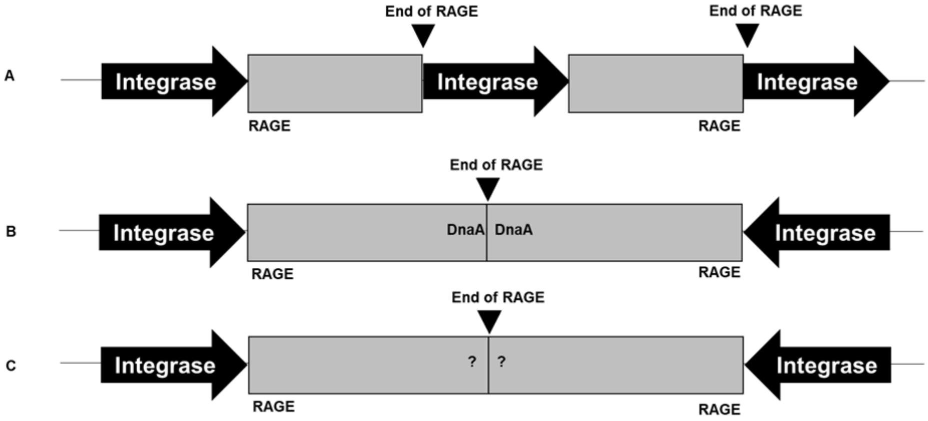


Figure Methods 1. Identification of RAGE termination positions.

Identification of RAGE associated cargo genes

The list of conserved nonconjugative genes or RAGE associated cargo genes were extracted based on *K Nakayama et al*, 2008[9]. The sequences were then inspected manually by observing the length of genes, conserved motifs/domains, and other elements such as signal peptides, transmembrane regions etc. A gene was defined as a “Full-length gene” if the sequence encoded a complete set of domains, a “Truncated gene” if the sequence encoded only a partial set of domains and a “Degraded gene” if no domain was identified on the sequence.

Analysis of multicopy cargo genes not associated with DNA mobilisation

Membrane proteins

Membrane proteins were manually extracted from the genome database and the SMART search engine[82] was used to identify membrane domains and other elements. The membrane protein was assigned as “OT_RAGE_membrane_protein” if no identifiable domain was identified. The membrane protein was assigned on new name if significant domain/main domains were found such as autotransporter proteins (Sca family), vut1-Putative vitamin uptake transporter, RhaT- Permease of the drug/metabolite transporter (DMT) superfamily, and Bax inhibitor-1 (BI-1)/YccA inhibitor of FtsH protease domains.

MRP/histidine kinases

All genes previously annotated as histidine kinase (HK) and multidrug resistance-associated proteins (MRPs) were extracted from the genome databases and analysed using SMART[82]. Both HK and mrp proteins contain an HATPase_c domain (Histidine kinase-type ATPases catalytic domain). This domain is found in several ATP-binding proteins, including: histidine kinase[83], DNA gyrase B[84], topoisomerases, and heat shock protein HSP90[85]. The new naming system of HK and mrp were classified based on the HATPase_c domain. HK or mrp proteins were renamed as “HATPase_c domain containing protein” if the search found an HATPase_c domain. HK or mrp proteins were renamed as “degraded HATpase_c” if no significant domains were found on the search. HK and mrp proteins in this study were classified as truncated gene because they only contained catalytic part (HATPase_c) and lacked other major domains such as sensor domain, HisKA(Histidine kinase A domain dimierization/His phospotransfer), and receiver Hpt (Histidine phosphotransfer)[86].

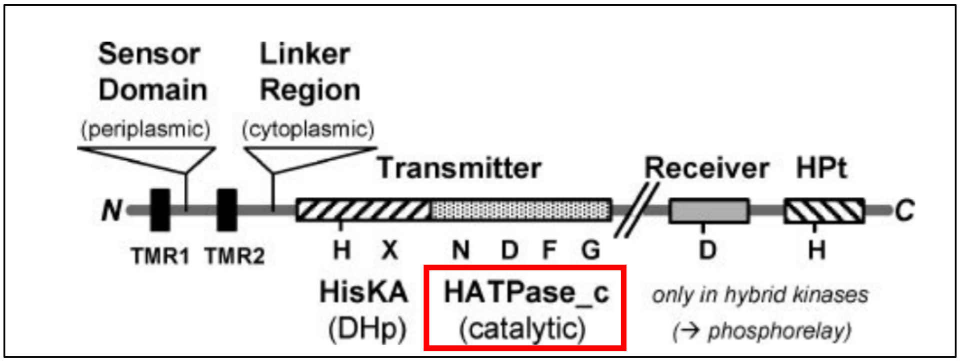


Fig. Methods 2. The overview of domain of histidine protein kinases. Figure is modified from ref[86].

In addition, the same HATPase_c domain was also found in symporter proteins. The sodium:proline symporter “PutP” was classified as full-length if the sequence contained symporter, HisKA, HATPase_c, and REC domains. PutP was was classified as “truncated PutP” or “degraded PutP” if the sequence lacked the symporter and REC domains. The symporter Sodium:pantothenate symporter "PanF" was classified as full-length if the sequence contained only the symporter region.

(p)ppGpp hydroplase/synthetases

We manually extracted all genes annotated as: SpoT, (p)pGpp, synthetase, and hydrolase from the genome databases. The sequences were then compared to their respective orthologs in *Escherichia coli* (ECO) and *Caulobacter crescentus* (CCS). Literature searches and GenomeNet motif search (pfam) were used to identify motifs and conserved regions in these sequences as shown (**Table Methods 2).**

Table Methods 2. Motifs conserved in SpoT. Underlined base(s) indicate important amino acid in the motif.

Domains	Conserved	Motifs	References
HD domain (Hydrolase)	region	AIDYAIHYHGXQTRESGDPYYYHPLHVALIIAQMKXDTVSVITALLHDTVEDTELTLSDIEREFGKEVAXLVDGVTKLXKLRFQSYHXQQAXNFRKLLLAISNDIRVLLVKLADRLHNMRTIESIKLLNKRIRIALETXEIYAPLAERIGA	Gemma C Atkinson et al., 2011
	H1	HXXXXR/KXXG/QXXYXXXP/Q/WXX	Justyna M Prusinska et al., 2019
	H2	I/VT/IAXLHD/N	Justyna M Prusinska et al., 2019
	H3	XLLXKLXXRXHNXXXX	Justyna M Prusinska et al., 2019
SYNTH domain (Synthetase)	region	G/ARXKXXYSIXXKMXXKXIXXXQLXDXXAXRXIXXXXXXXXXXCYXXLXXIHXXYXXXPXXXQDFIXXPKXNGYQSXHTXIXGPXXXXIEVQIRTXXMHXXXXXGXAAHWXYK	Gemma C Atkinson et al., 2011
	G	GRHK	Justyna M Prusinska et al., 2019
	YQS	NGYQSXHT	Fabio Lino Gratani et al., 2018
TGS domain (Thr-tRNA synthetase, GTPase and SpoT domain)	region	CFTPXGKLIALPKGATVVDFAYKXHSELGNKCIGAKISNKVVPLDTQLQNGDQVEIIT	Gemma C Atkinson et al., 2011
	H	DFAYXXHXXXG	Winter et al., 2018
Helical domain	region	TFAVTGKAQSEIRKFIRXQAYKKYIDLGKEILIQTLKKIQVANINVCIAKIAHXLNKKNVEEVFFXIGXELLSKKEIIKIIT	GenomeNet motif search (pfam)
CC domain (Concerve cysteine/RIS-Ribosome InterSubunit domain)	CC	CCYPLPGDLIIGLCT	Jain V et al., 2007; Gemma C Atkinson et al., 2011
ACT domain (Aspartokinase, Chorismate mutase and TyrA/RRM-RNA Recognition motif)	region	RNKIGSLASITTILENNNXNICNIKTTNXTQSTXQIIIDIEISTLEQLNKIXNILQSSXDIISVXR	Gemma C Atkinson et al., 2011

We employed a naming system based on the presence of domains/motifs in each CDS. genes containing all domains (HD, SYNTH, TGS, Helical, CC, ACT) were annotated as SpoT. Genes having only the HD domain were annotated as SpoT-hydrolase, genes encoding only SYNTH domain was annotated as SpoT-synthetase. Genes encoding the HD domain but lacking one or more of the conserved histidines was annotated as truncated hydrolase. Short fragments that could be aligned to Hydrolase but lack complete domains were annotated as degraded hydrolase. Genes containing HD domain merging with a part of HATPase were named as HATPase-SpoT-hydrolase and genes containing HD domain merging with a part of Mrp were named as Mrp-Hydrolase, respectively.

DNA methyltransferases

DNA methyltransferase (MTase) genes were extracted from the genome and analysed on SMART[82] for protein domain annotation. However, SMART does not provide the details of MTase motifs within the predicted domain. Therefore, multiple sequence alignment of DNA methyltransferase was further characterized for identification motifs using Geneious. We used the conserved amino acid residues in the Dam (DNA adenine methyltransferase) protein of *E. coli* (acc.no. P0AEE9) as a reference for identification motifs I-VII & motif X of MTase at C-terminal region[87]. The protein sequence of DNA methyltransferase containing motif I-VII & motif X was indicated as full-length gene. The protein sequence of DNA methyltransferase with incomplete motifs and unidentified motifs were indicated as truncated gene, and degraded gene, respectively.

Replicative DNA helicases

DNA helicase genes were filtered from the genome and their protein domains were characterized by SMART[82]. Multiple sequence alignments of the helicase genes was then carried out in order to identify motifs. We used the conserved amino acid residues in the DnaB protein of *E. coli* K12 (acc.no. NC000913.3) as a reference for the identification of motifs I-VII at C-terminal region [88]. The protein sequence of DnaB containing motif I-VII was indicated as full-length gene. The protein sequence of DnaB with incomplete motif and unidentified motifs was indicated as truncated gene and degraded gene, respectively.

Uncharacterised proteins

Between 308-464 genes annotated as hypothetical or uncharacterised were found in the eight genomes of *Orientia*. In this study, we only manually analysed uncharacterised proteins from Karp strain as a model to minimize the analysis time. Uncharacterised proteins from Karp were filtered from the genome and the protein domain was characterized by SMART[82]. Where clear groups of homologous genes within the set of Karp genes was found, these were classified into 25 defined groups. These were renamed according to known domains with which they had homology, or named Ot_RAGE_hypo_group 1-7. These 25 groups were then aligned to the other seven genomes in our dataset in order to determine the conservation of the groups of genes.

Some uncharacterised proteins were changed to new name, and no longer classified as hypothetical proteins, if they aligned to known genes such as DnaA, Phage portal protein, Lipase3 etc.

Analysis of multicopy genes involved in DNA mobilisation

Insertion sequence transposable elements

The presence of insertion sequence (IS) elements in Ot was investigated using the online search tool ISfinder[89] to match with attributes and nomenclatures previously submitted for Orientia-specific IS[9]. Each IS match was manually traced for completeness with flanking inverted repeats (IR) and direct repeats (DR) along respective genome sequences. Extensive analysis was performed with Karp strain to identify the complete set of IS elements and classified into classes. Only ISOt3 and ISOt5 were systematically analysed across all 8 different *Orientia* genomes.

Integrases

Genes annotated as integrase genes were extracted from the genomes and protein domains were screened by SMART[82]. Integrase in *Orientia* is a phage integrase which is classified into two major families: the tyrosine recombinases and the serine recombinases, based on mode of catalysis[90]. Then multiple sequence alignment of phage integrase domain was analysed for identification motifs using Geneious. We used the conserved amino acid residues in Bacteriophage P2-integrase (acc.no. AF063097.1) and Enterobacteria phage P2-integrase (acc.no. NC_009488.1) as references for the identification of three domains; arm-type binding motifs at N-terminal region, core-type binding (CB), and catalysis at C-terminal region. The His-X-X-Arg motifs and second conserved Arginine on catalytic domain were also included in the alignment[90,91]. The protein sequence of phage integrase containing arm-type binding, core-type binding, and catalysis motifs was indicated as full-length gene. The protein sequence of phage integrase with incomplete motif and unidentified motif was annotated as truncated gene and degraded gene, respectively.

Reverse transcriptases

Reverse transcriptase (*rvt*) genes were filtered from the genome and protein domains were characterized by SMART[82]. Multiple sequence alignment of *rvt* was then analysed for identification motifs. We used the conserved amino acid residues of group II intron reverse transcriptase/maturase (LtrA) in *E. coli* (acc.no. WP_096836589.1), and *Lactococcus lactis* (acc.no. NZ_CP059048.1) as a reference for the identification of three domains; reverse transcriptases (RVT_N) at N-terminal site, reverse transcriptases (RT), and Group II intron, maturase-specific domain (GIIM) at C-terminal site[92-94]. The protein sequence of reverse transcriptase containing RVT_N, RT, and GIIM was indicated as full-length gene. The protein sequence of reverse transcriptase with incomplete motif and unidentified motif was indicated as truncated gene and degraded gene, respectively.

Transposases

Genes annotated as transposase genes were first extracted from the genome. Then, protein domains and motifs were further characterized by SMART[82] and Geneious, respectively. Transposase gene in *Orientia* belong to restriction endonuclease-like proteins or PD-(D/E)XK nucleases and DD[E/D]- transposase, which generally contain the catalytic domain, and transposon-binding domain[95,96]. Some transposases additionally contain C-terminal or

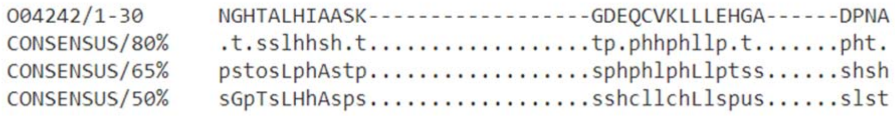
N-terminal domains[97]. In this study, we used the conserved amino acid residues of in *E. coli* (acc.no. NC 002695.2) as a reference for identification restriction endonuclease-like motifs (I-IV). Three conserved active sites in motif II and III, one of which is aspartic acid (D), one is either glutamic (E) or aspartic acid (D) and/or the last one is lysine (K), were identified for characterization of PD-(D/E)XK nucleases and DD[E/D]- transposase motifs [95,96]. The protein sequence of transposase containing motifs (I-IV) and PD-(D/E)XK signature residues was indicated as full-length gene. The protein sequence of transposase with incomplete motif and unidentified motif was indicated as truncated gene and degraded gene, respectively.

Prophage genes

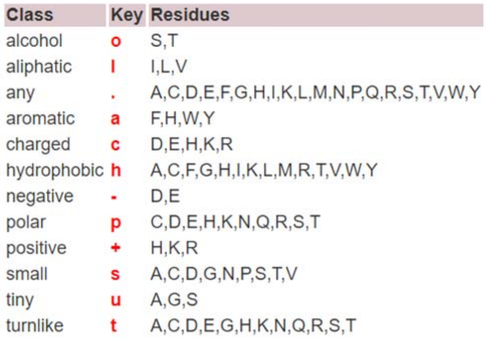
Potential prophage sequences within 8 *Orientia* genomes were identified using PHASTER (PHAge Search Tool Enhanced Released)[41] where specific phage related proteins such as ‘coat’, ‘fiber’, ‘head’, ‘plate’, ‘tail’, 'integrase', 'terminase', 'transposase', 'portal', 'protease' or 'lysin' within bacterial genomes were recognized using a sequence identity search.

Analysis of Ankyrin repeat containing proteins

To identify Ankyrin repeat (AR) proteins (Anks) from previously annotated records, all Ank sequences were extracted and analyzed using SMART[82] to predict ARs and other domains including coiled-coil, F-box, and PRANC. SMART[82] defines an Ank as a 33-residue motif. Ank commonly involves in protein-protein interaction. The core of the repeat seems to be a helix-loop-helix structure. SMART’s consensus for an ankyrin repeat is shown in Fig. Methods 3. It is important to note that the protein structure or functionality of any Ank was not characterized in this study. Therefore, any Ank that contains only one or two repeats may be non-functional.



```
O04242/1-30      NGHTALHIAASK------------------GDEQCVKLLLEHGA------DPNA
CONSENSUS/80%    .t.sslhhsh.t..................tp.phhphllp.t.......pht.
CONSENSUS/65%    pstosLphAstp..................sphphlphLlptss......shsh
CONSENSUS/50%    sGpTsLHhAsps..................sshcllchLlspus......slst
```



Class	Key	Residues
alcohol	o	S,T
aliphatic	l	I,L,V
any	.	A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y
aromatic	a	F,H,W,Y
charged	c	D,E,H,K,R
hydrophobic	h	A,C,F,G,H,I,K,L,M,R,T,V,W,Y
negative	-	D,E
polar	p	C,D,E,H,K,N,Q,R,S,T
positive	+	H,K,R
small	s	A,C,D,G,N,P,S,T,V
tiny	u	A,G,S
turnlike	t	A,C,D,E,G,H,K,N,Q,R,S,T

Fig .Methods 3 Consensus term for Ank domain by SMART.

Identification of homologous Ank repeats across Ot strains were manually inspected using Geneious. The criteria for identification of a homolog Ank is based on sequence similarity and repeated units. Individual Ank sequences of the other 7 strains were blasted against Ot strain Ikeda. Then, the sequence that presented the highest identity (>80-90%) was chosen to verify the similarity of Ankyrin repeats and other domains. The sequences were given the name based on the published Ikeda Anks if the overall sequence's identity to Ikeda Anks was more than 80% and presented a similar set of ARs. The sequences were given a new name if the overall sequence identity to Ikeda Anks was less than 80% and presented a different set of ARs, this included extra repeated units or missing repeated units.

Unidentified Anks or hypothetical proteins were manually searched and inspected using Geneious. To search unidentified Anks in Ikeda and other 7 strains, each repeat unit of an individual published Ikeda Anks were imported into "Find motifs" tool, and the maximum mismatches were set up to 10. The closely matched sequences were further identified the repeated units and other domains by SMART. Then, the sequences of unidentified Ank were blasted against the published Anks or blasted within the strain to check whether it was different from identified Anks or not. The newly identified Anks were given a name by continued ranking after the published Ikeda Anks, starting with Ank21, Ank22, Ank23, etc.

Analysis of Tetratricopeptide repeat containing proteins

To identify TPR proteins from the previous annotated records of Orientia 8 strains, all TPR sequences were extracted and analyzed using SMART[82]. The program predicted the location of TPR motifs and other domains including signal peptide and transmembrane region. All identified TPR proteins were grouped based on the similarity of the location of TPR motifs. Each group consists of one long gene, the master gene, and multiple shorter duplication remnants. For unidentified TPR were manually searched using Geneious. All TPR proteins were renamed based on the group number followed the number of TPR repeats.

Analysis of P-type IV secretion systems (*rvh)*

Literature search and blast search (NCBI and KEGG) were performed to identify the presence of each Rvh subunit (RvhB1 to VirB11 and RvhD4) in Ot. The amino acid sequences of each Vir subunit present in Ot Boryong strain (OTS) were compared to their respective orthologs in *Rickettsia bellii* (RBE) and *Agrobacterium tumefaciens* (ATU) for their percent of amino acids identity, length of amino acid sequences, and presence of motifs.

The presence of motifs was used as the major criteria to identify Vir subunits as indicated in Table Methods 3:

Table Methods 3. Vir proteins and their motifs. Underlined base(s) indicate important amino acid in the motif. Red letter indicates variability in amino acid sequence.

Subunits	Motifs	References
VirB2-1	TM region 1 (GXXXXXXXXXXXXXXXIXXXXG), TM region 2 (AII/VI/VXXXA/SXX)	Krogh Anders et al., 2001, Lai Erh-Min and Clarence I. Kado, 2000,
VirB2-2	TM region 1 (GXXXXXXXXXXXXXXXIXXXXG), TM region	

	2 (AII/VIXXXA/SXX)	Gillespie Joseph J et al., 2009
VirB2-3	TM region 1 (GXXXXXXXXXXXXXXXIXXXXG), TM region 2 (AII/VIXXXSXX)	
VirB3	L-TRP-GV motif (LXXXXTRPXXXXGV)	Cao TB and Saier MH 2001, Gillespie Joseph J et al., 2009
VirB4-1	Walker A (GPXGXGKT), Motif C (FDKDRGXE), Walker B (RIXXXXDGXXXXXXXDE), Motif D (LXXXRKXN), Motif E (IXATQ)	Gillespie Joseph J et al., 2009, Gillespie Joseph J et al., 2016
VirB4-2	Walker A (GXXXXGK/R), Walker B (SLXXXXXXXIXXXDX)	
VirB6-1	Variable TM region (VXAFXXLYXXXXGXXILLX), Conserved cytoplasmic loop (PXXXXXXXFXXTXXXXXXW)	Judd Paul K et al., 2005, Lawley TD et al., 2003, Gillespie Joseph J et al., 2016
VirB6-2	Variable TM region (XXAALXLYXXFFXXXXXXX), Conserved cytoplasmic loop (PXXXXXXXFXXTXXXXXXW)	
VirB6-3	Variable TM region (IXAXLXLYXMXXGXXFXLG), Conserved cytoplasmic loop (PXXXXXXXFXXTXXXXXXW)	
VirB6-4	Variable TM region (XXXXXXLYXTXXGXXFXLG), Conserved cytoplasmic loop (PXXXXXXXFXXT/IXXXXXXW)	
VirB6-5	Variable TM region (VXXXLXLXXXFXGXXFLIG), Conserved cytoplasmic loop (PXXXXXXXFXXTXXXXXXW)	
VirB7	Conserved cys between position 15-35; Possibility of being a small lipoprotein	Gillespie Joseph J et al., 2010
VirB8-1	Homodimerization domain I (YXXXREXY), VirB4 interaction region (XXXYK)	Bailey Susan et al., 2006, Terradot Laurent et al., 2005, Gillespie Joseph J et al., 2009
VirB8-2	Homodimerization domain I (YXXXREXY), NPxG motif (NPXG), VirB4 interaction region (XXXYR)	
VirB9-1	DXR-YXP motif (DXRXXXXXYXP), Beta 1 (NXXYXX), Beta 2-3 - OM extrusion region (PXXXXDXXXXTXXXF - PXXXXXG/DDXXXE), VirB7 interaction region (RXGXXXXCXXN)	Gillespie Joseph J et al., 2009
VirB9-2	DXR-YXP motif (DXRXXXXXYXP)	
VirB10	OM pore gating (DXLGXXGXXGXV), Beta 6a (VLXSAX), Beta 7a (XTXXXNQG)	Banta Lois M et al., 2011, Gillespie Joseph J et al., 2016
VirB11	Linker A (IRXXSXXXXXL), Beta 1 (XXEXXXNXPG), Beta 5 (LPXXXRXQXXXPP), ATPase region/Beta 7 - Alpha E (GXTXXXKTT), Beta 8 (ERXIXXED), Alpha F - Beta 10 (LXXXXLRXRPDRIXXXE), Beta 11 (GHPGSIXTXH)	Jorge Ripoll-Rozada & Ignacio Arechaga, 2013
VirD4	DNA binding motif A (APTXXGKGXGXVIPXXXXXXXSVXXXDXK), DNA binding motif B (XFLLDEFXXLGKXXX)	L. Leloup et al, 2002, Renu B. Kumar and Anath Das, 2002

1001

VirB1 and VirB5 could not be identified in *Ot*. However, VirB7 was annotated based on gene positioning, conserved cysteine(s), and its important role of being a small lipoprotein in T4SS of *Rickettsia* species.

These *rvh* genes identified in Boryong were then set as a reference strain for Ot. Each *vir* gene was then blasted (nucleotide blast using Geneious Prime) to identify the presence of each *rvh* across the eight strains of Ot (Boryong, Ikeda, Karp, Kato, Gilliam, TA686, UT76, UT176). The amino acid sequences of each Rvh subunit from the eight strains of *Ot* were then aligned (Multiple alignment - Clustal Omega) to verify their motifs. Even though some of the gene copies appear to be a truncation or pseudogene due to loss of some motif(s) like that in RvhB4-II, RvhB8-II, and RvhB9-II, they are well characterized in literature. So, these names were kept the same in our annotation.

Analysis of F-type IV secretion systems (RAGE T4SS)

Literature search and blast search (NCBI and KEGG) were performed to identify the presence of each Tra subunit (F-type T4SS: TraA to TraN, TraU to TraW, TrbC, TrbE and P-type T4SS: TraA$_{\text{Ti}}$, TraD$_{\text{Ti}}$,) in Ot. The amino acid sequences of each Tra subunit present in Ot Ikeda strain (OTT) were then compared to their respective orthologs in *Rickettsia bellii* (RBE) and *Escherichia coli* (ECZ) for F-type T4SS or *Agrobacterium tumefaciens* (ATU) for P-type T4SS. The presence of motifs was used as the major criteria to identify Tra subunits as indicated in Table Methods 4:

Table Methods 4. Motifs conserved in Tra and Trb proteins. Underlined base(s) indicate important amino acid in the motif. Red letter indicates variability in amino acid sequence.

Subunits	Motifs	References
TraE	Anchor region (LVKYNKXLLXXTXXL/IAXXXX), Predicted conserved region-1 (SXXXXXXXYLXXXA), Predicted conserved region-2 (KXXXXXSXFFXXXXXV), Predicted conserved region-3 (VXIXGXXXXWXXXXKXXXXXK/RXYXLXXK) - GenomeNet Motif search (TraE region)	Frost Laura S et al., 1994, Kelley Lawrence A et al., 2015, Bragagnolo Nicholas et al., 2020

TraB	Coiled-coil domain (IXXXXQ/KXXXXL/FXXXXKXXXX), Predicted conserved region (GXXSSERAXXR) - GenomeNet Motif search (TrbI-like region), OM pore gating (GXXGXXGXV), Alpha-3 (GXXXGXXXA/VXXXLXDXXIKR/QAXXXXP)	Gilmour Matthew W, Banta Lois M et al., 2011, Gillespie Joseph J et al., 2016
TraV	Conserved cysteine region (CXXXXXXXXXXXXF/L/VXCXXXXXXXC), Predicted conserved region (LXXLF/LXXXXXXG/CE) - GenomeNet Motif search (TraV region)	Harris R L et al., 2001, Harris R L et al., 2002
TraC	Predicted conserved region-1 (YXXYXXE/KXXLFXNXXXXGFXLXXXP), Predicted conserved region-2 (YXXLXXQXXXXXFXLXXXXD) - GenomeNet Motif search (TraC region), Walker A (GXXGXGKX), Motif C (V/AXDXGXXXK), Walker B (RXXXXLXXIDEXW), Motif D-E (RXXXGXXXXXTQ)	Lawley T D et al., 2003, Gillespie Joseph J et al., 2009, Gillespie Joseph J et al., 2016
TraW	TrbC interaction region (EXXXLXVIMXXLXXXXXXGXXXXXXXXF), Predicted conserved region-1 (NP/SLXXXXXXXXLXXIXGDDXXQVXWXK), Predicted conserved region-2 (FDQXXXLXXXXXIXXXPA) - InterPro Motif search (TraW region)	Shala-Lawrence Agnesa et al., 2018
TraU	Signaling domain (AXXXCXG), Hydrophobic-2 (CMVXLG/W), Hydrophobic-3 (YWLXIXX), Hydrophobic-4 (FXNXXAXXACXAD), Hydrophilic-1 (KXXXRXQM), Hydrophilic-2 (W/LRKRXC/Y)	Moore Deanna et al., 1990, Frost Laura S et al., 1994
TrbC	Signaling domain (MXIRVMXLXXLLXVNN), Predicted conserved region-1 (FVSFSXXXXXLK), Predicted conserved region-2 (GXXXXXRG/RXXNNXXXXT) - GenomeNet Motif search (TrbC region), TraW interaction region (IDP/SXLFXXYXXXXVP/LXXVX)	Maneewannakul S et al., 1991, Shala-Lawrence Agnesa et al., 2018
TraN	Conserved cysteine region-1 (SCXEGXX), Conserved cysteine region-2 (SXCXXXE), Conserved cysteine region-3 (IGXXC), Conserved cysteine region-4-5 (CXXXKXXYCXFXSK/RLAXXXQ/H), Conserved cysteine region-6 (CRG/DXTVXE/KLQXXXF), Predicted conserved region-1 (ECXE), Predicted conserved region-2 (CXLXXXXC), Predicted conserved region-3 (CLXXXXXYXC), Predicted conserved region-4 (CXKXXXXXXN/HCC) - GenomeNet Motif search (TraN region)	Klimke William A et al, 2005
TraF	Predicted conserved region (G/XXXWYNX) - GenomeNet motif search (TraF region), C-X-X-C motif (CXXC), Beta 10 (VPXXXL/SX), Alpha 7 (ISXD/N/EXXXXXXL)	Elton Trevor C et al., 2005
TraH	TrbI interaction region-1 (TXXGXXQXQ/LAAGYYXXGXLXXRT), TrbI interaction region-2 (NIXXXAX), Predicted conserved region-1 (CXXIXXYLXSFSXIXG/V/REXL), Predicted conserved region-2 (FLSSIGXXXXXXXYXXXXXISG), Predicted conserved region-3 (LXQXXXEXXXXXR) - GenomeNet motif search (TraH)	Aruthyunov Denis et al., 2010, Lawley T D et al., 2003
TraG	Membrane spanning region-1 (M/WXWXXXXXXIXXXLXXX), Membrane spanning region-2 (QSVXXXL/VXXXXXXXVFPMXXLXXXXXIXKXWIXXIIWVXSWPVXF), Membrane spanning region-3 (XXA/ST/MXXXLAXXXPXLSWXVXK/NXXXXXXXXLXXXFSXXXV), Cleavage site (ASXXGX), Predicted conserved region-1 (SXXXXLSXXL), Predicted	Firth Neville and Ron Skurray, 1992, Audette Gerald F et al., 2006

	conserved region-2 (KQXXEQXXXXXXYXXQXS) - GenomeNet motif search (TraG N-terminal region)	
TraD	Transmembrane region-1 (MXXQXXXNXXXIGLXXXXXWXXXXXYQ), Transmembrane region-2 (FLXXSXXXEXXXXFXIX), Walker A (GTXGXGKXX), Walker B (XXWFXXDELP)	Frost Laura S et al., 1994, Lessl Monika et al., 1992
TraI	Helicase region-5 (XHGYAXTXXXXQ/KXAXXXXXXXVLXXXXXXXXXX), Predicted conserved region-1 (IXEGXEXXXXLXXXXIXGXIIXXXXI/VXXXXNXXP/LXXG/S), Predicted conserved region-2 (AVXNXVXXXXAXXVXE/DXKXXXXXXXXKXXFNXVLKXXGL) - GenomeNet motif search (Toprim region)	Farrand Stephen K et al., 1996
$TraA_{Ti}$	Nickase region-1 (AIXFXXXXXXXRS/IXGXXSCXK/NXXYXXXXXXXXXXXXXXXXXXXXXXXXXVXH), Nickase region-2 (NEVER/QXXXXXXNSXXXXXIVIA/VLP/Q), Nickase region-3 (NXHXH/NXXXXXRXXXXXG), Helicase region-1 (XXGXAGXGKXXXXXXA/V), Helicase region-2 (XXV/IXDE/KAGMV/A), Helicase region-3 (XXXXXLXGDXXQL/RXXXEXGXXFXXXXXXXXXXXL), Helicase region-5 (XHGYAXTXXKXQ/HGAXXXXXXV/ILXXXXXXXXX), Helicase region-6 (YV/TXMT/IRH/YXXXXXLY)	Farrand Stephen K et al., 1996, Alt-Morbe J et al., 1996
$TraD_{Ti}$	Predicted conserved region-1 (RKXXXR/QXXXXXG/AXXV/LXXAXL), Predicted conserved region-2 (IGXXXFXXXXXN) - GenomeNet motif search (TraD region)	Farrand Stephen K et al., 1996

Note that $TraA_{Ti}$ found in *Rickettsia* is fragmented into $TraA_{Ti}$ and TraI in *Ot*. The longest *tra* and *trb* genes identified in Ikeda were set as references and were then blasted (nucleotide blast using Geneious Prime) to identify their presence across the 8 strains of Ot (Boryong, Ikeda, Karp, Kato, Gilliam, TA686, UT76, UT176). The amino acid sequences of each Tra subunit from the 8 strains of Ot were then aligned (Multiple alignment - Clustal Omega) to verify their motifs. Those amino acid sequences with difference greater than 10% from full length gene in Ikeda or missing motif(s) are considered pseudogene (truncation).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All data generated by this work is available within the manuscript and supporting information.

Author Contributions

Project design and supervision (JS); data analysis and figure preparation (SG, CK, JW, HA, JS, JJG); original manuscript writing (JS); manuscript revisions (SG, CK, JW, JJG, JS).

Acknowledgements and Funding

We thank all the members of the lab in Cambridge and Bangkok for discussions and support. We are also grateful to Liz Batty and Nick Day in MORU for discussions, encouragement, and support. This work was funded by a Royal Society Dorothy Hodgkin Fellowship DH140154 (JS); Swiss National Science Foundation IZSTZ0_193925 (JS, CK); Wellcome Trust Senior Research Fellowship 224277/Z/21/Z (JS), grants from the National Institutes of Health, USA, R21 AI156762 and R21 AI166832 (JJG) and American Heart Association grant 20PRE35210610 (HA).

References

1 Giannotti, D., Boscaro, V., Husnik, F., Vannini, C. & Keeling, P. J. The "Other" Rickettsiales: an Overview of the Family "Candidatus Midichloriaceae". *Appl Environ Microbiol* **88**, e0243221, doi:10.1128/aem.02432-21 (2022).

2 Salje, J. Cells within cells: Rickettsiales and the obligate intracellular bacterial lifestyle. *Nat Rev Microbiol*, doi:10.1038/s41579-020-00507-2 (2021).

3 Schon, M. E., Martijn, J., Vosseberg, J., Kostlbacher, S. & Ettema, T. J. G. The evolutionary origin of host association in the Rickettsiales. *Nat Microbiol* **7**, 1189-1199, doi:10.1038/s41564-022-01169-x (2022).

4 Castelli, M. *et al.* Deianiraea, an extracellular bacterium associated with the ciliate Paramecium, suggests an alternative scenario for the evolution of Rickettsiales. *ISME J* **13**, 2280-2294, doi:10.1038/s41396-019-0433-9 (2019).

5 Ettema, T. The evolutionary origin of host association in an ancient bacterial clade. doi:10.7554/eLife.19469.001 (2022).

6 Weitzel, T. *et al.* Endemic Scrub Typhus in South America. *N Engl J Med* **375**, 954-961, doi:10.1056/NEJMoa1603657 (2016).

7 Izzard, L. *et al.* Isolation of a novel Orientia species (O. chuto sp. nov.) from a patient infected in Dubai. *J Clin Microbiol* **48**, 4404-4409, doi:10.1128/JCM.01526-10 (2010).

8 Cho, N. H. *et al.* The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. *Proc Natl Acad Sci U S A* **104**, 7981-7986, doi:10.1073/pnas.0611553104 (2007).

9 Nakayama, K. *et al.* The Whole-genome sequencing of the obligate intracellular bacterium Orientia tsutsugamushi revealed massive gene amplification during reductive genome evolution. *DNA Res* **15**, 185-199, doi:10.1093/dnares/dsn011 (2008).

10 Batty EM, C. S., Blacksell SB, Richards A, Paris D, Bowden R, Chan C, Lachumanan R, Day N, Donnelly P, Chen SL, Salje J. Long-read whole genome sequencing and comparative analysis of six strains of the human pathogen Orientia tsutsugamushi. *Plos Negl Trop Dis* (2018).

11 Gillespie, J. *et al.* A Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. *J Bacteriol* **194**, 376-394, doi:10.1128/JB.06244-11 (2012).

12 Ogata, H. *et al.* Genome sequence of Rickettsia bellii illuminates the role of amoebae in gene exchanges between intracellular pathogens. *PLoS Genet* **2**, e76, doi:10.1371/journal.pgen.0020076 (2006).

13 Blanc, G. *et al.* Lateral gene transfer between obligate intracellular bacteria: evidence from the Rickettsia massiliae genome. *Genome Res* **17**, 1657-1664, doi:10.1101/gr.6742107 (2007).

14 Londono, A. F. *et al.* Whole-Genome Sequence of Rickettsia parkeri Strain Atlantic Rainforest, Isolated from a Colombian Tick. *Microbiol Resour Announc* **8**, doi:10.1128/MRA.00684-19 (2019).

15 Gillespie, J. J. *et al.* Genomic diversification in strains of Rickettsia felis Isolated from different arthropods. *Genome Biol Evol* **7**, 35-56, doi:10.1093/gbe/evu262 (2014).

16 Felsheim, R., Kurtti, T. & Munderloh, U. Genome sequence of the endosymbiont Rickettsia peacockii and comparison with virulent Rickettsia rickettsii: identification of virulence factors. *PLoS One* **4**, e8361, doi:10.1371/journal.pone.0008361 (2009).

17 Cho, N., Kim, H., Lee, J. & Kim, I. The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host– cell interaction genes. *PNAS* **104**, 7981-7986 (2007).

18 Hagen, R., Verhoeve, V., Gillespie, J. & Driscoll, T. in *Genome Biol Evol* Vol. 10(12) 3218-3229 (2018).

19 Mika-Gospodorz, B. *et al.* Dual RNA-seq of Orientia tsutsugamushi informs on host-pathogen interactions for this neglected intracellular human pathogen. *Nat Commun* **11**, 3363, doi:10.1038/s41467-020-17094-8 (2020).

20 Hagen, R., Verhoeve, V., Gillespie, J. & Driscoll, T. Conjugative transposons and their cargo genes vary across natural populations of Rickettsia buchneri infecting the tick Ixodes scapularis. *Genome Biol Evol* (2018).

21 Nakayama, K. *et al.* Genome comparison and phylogenetic analysis of Orientia tsutsugamushi strains. *DNA Res* **17**, 281-291, doi:10.1093/dnares/dsq018 (2010).

22 Atwal, S. *et al.* The obligate intracellular bacterium Orientia tsutsugamushi differentiates into a developmentally distinct extracellular state. *Nat Commun* **13**, 3603, doi:10.1038/s41467-022-31176-9 (2022).

23 Atkinson, G., Tenson, T. & Hauryliuk, V. The RelA/SpoT homolog (RSH) superfamily: distribution and functional evolution of ppGpp synthetases and hydrolases across the tree of life. *PLoS One* **6**, e23479, doi:10.1371/journal.pone.0023479 (2011).

24 Hauryliuk, V., Atkinson, G., Murakami, K., Tenson, T. & Gerdes, K. Recent functional insights into the role of (p)ppGpp in bacterial physiology. *Nat Rev Microbiol* **13**, 298-309, doi:10.1038/nrmicro3448 (2015).

25 Mosavi, L., Minor, D. & Peng, Z. Consensus-derived structural determinants of the ankyrin repeat motif. *Proc Natl Acad Sci U S A* **99**, 16029-16034, doi:10.1073/pnas.252537899 (2002).

26 Jernigan, K. K. & Bordenstein, S. R. Ankyrin domains across the Tree of Life. *PeerJ* **2**, e264, doi:10.7717/peerj.264 (2014).

27 Frank, A. C. Molecular host mimicry and manipulation in bacterial symbionts. *FEMS Microbiol Lett* **366**, doi:10.1093/femsle/fnz038 (2019).

28 VieBrock, L. *et al.* Orientia tsutsugamushi ankyrin repeat-containing protein family members are Type 1 secretion system substrates that traffic to the host cell endoplasmic reticulum. *Front Cell Infect Microbiol* **4**, 186, doi:10.3389/fcimb.2014.00186 (2014).

29 Beyer, A. *et al.* Orientia tsutsugamushi Strain Ikeda Ankyrin Repeat-Containing Proteins Recruit SCF1 Ubiquitin Ligase Machinery via Poxvirus-Like F-box Motifs. *J Bacteriol*, doi:10.1128/JB.00276-15 (2015).

30 Beyer, A. *et al.* Orientia tsutsugamushi Ank9 is a multifunctional effector that utilizes a novel GRIP-like Golgi localization domain for Golgi-to-endoplasmic reticulum trafficking and interacts with host COPB2. *Cell Microbiol*, doi:10.1111/cmi.12727 (2017).

31 Evans, S., Rodino, K., Adcox, H. & Carlyon, J. Orientia tsutsugamushi uses two Ank effectors to modulate NF-κB p65 nuclear transport and inhibit NF-κB transcriptional activation. *PLoS Pathog* **14**, e1007023, doi:10.1371/journal.ppat.1007023 (2018).

32 Adcox, H. E. *et al.* Orientia tsutsugamushi Nucleomodulin Ank13 Exploits the RaDAR Nuclear Import Pathway To Modulate Host Cell Transcription. *mBio* **12**, e0181621, doi:10.1128/mBio.01816-21 (2021).

33 Kipreos, E. T. & Pagano, M. The F-box protein family. *Genome Biol* **1**, REVIEWS3002, doi:10.1186/gb-2000-1-5-reviews3002 (2000).

34 Perez-Riba, A. & Itzhaki, L. S. The tetratricopeptide-repeat motif is a versatile platform that enables diverse modes of molecular recognition. *Curr Opin Struct Biol* **54**, 43-49, doi:10.1016/j.sbi.2018.12.004 (2019).

35 Bang, S. *et al.* Inhibition of eukaryotic translation by tetratricopeptide-repeat proteins of Orientia tsutsugamushi. *J Microbiol* **54**, 136-144, doi:10.1007/s12275-016-5599-5 (2016).

36 Kaur, S. *et al.* TolC-dependent secretion of an ankyrin repeat-containing protein of Rickettsia typhi. *J Bacteriol* **194**, 4920-4932, doi:10.1128/JB.00793-12 (2012).

37 Johnson, C. & Grossman, A. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annu Rev Genet* **49**, 577-601, doi:10.1146/annurev-genet-112414-055018 (2015).

38 Belfort, M. & Lambowitz, A. M. Group II Intron RNPs and Reverse Transcriptases: From Retroelements to Research Tools. *Cold Spring Harb Perspect Biol* **11**, doi:10.1101/cshperspect.a032375 (2019).

39 Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev* **38**, 865-891, doi:10.1111/1574-6976.12067 (2014).

40 Wang, G. *et al.* Bacteriophage WO Can Mediate Horizontal Gene Transfer in Endosymbiotic Wolbachia Genomes. *Front Microbiol* **7**, 1867, doi:10.3389/fmicb.2016.01867 (2016).

41 Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* **44**, W16-21, doi:10.1093/nar/gkw387 (2016).

42 Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: a fast phage search tool. *Nucleic Acids Res* **39**, W347-352, doi:10.1093/nar/gkr485 (2011).

43 Gillespie, J. *et al.* Secretome of obligate intracellular Rickettsia. *FEMS Microbiol Rev*, doi:10.1111/1574-6976.12084 (2014).

44 Lin, M., Zhang, C., Gibson, K. & Rikihisa, Y. Analysis of complete genome sequence of Neorickettsia risticii: causative agent of Potomac horse fever. *Nucleic Acids Res* **37**, 6076-6091, doi:10.1093/nar/gkp642 (2009).

45 Gillespie, J. *et al.* Phylogenomics reveals a diverse Rickettsiales type IV secretion system. *Infect Immun* **78**, 1809-1823, doi:10.1128/IAI.01384-09 (2010).

46 Matson, S. W. & Ragonese, H. The F-plasmid TraI protein contains three functional domains required for conjugative DNA strand transfer. *J Bacteriol* **187**, 697-706, doi:10.1128/JB.187.2.697-706.2005 (2005).

47 Scherzinger, E., Lurz, R., Otto, S. & Dobrinski, B. In vitro cleavage of double- and single-stranded DNA by plasmid RSF1010-encoded mobilization proteins. *Nucleic Acids Res* **20**, 41-48, doi:10.1093/nar/20.1.41 (1992).

48 Escobar, M. A. & Dandekar, A. M. Agrobacterium tumefaciens as an agent of disease. *Trends Plant Sci* **8**, 380-386, doi:10.1016/S1360-1385(03)00162-6 (2003).

49 Weinert, L. A., Welch, J. J. & Jiggins, F. M. Conjugation genes are common throughout the genus Rickettsia and are transmitted horizontally. *Proc Biol Sci* **276**, 3619-3627, doi:10.1098/rspb.2009.0875 (2009).

50 Verhoeve, V. I., Lehman, S. S., Driscoll, T. P., Beckmann, J. F. & Gillespie, J. J. Metagenome diversity illuminates origins of pathogen effectors. *bioRxiv*, doi:10.1101/2023.02.26.530123 (2023).

51 Gillespie, J. J. *et al.* An anomalous type IV secretion system in Rickettsia is evolutionarily conserved. *PLoS One* **4**, e4833, doi:10.1371/journal.pone.0004833 (2009).

52 Gillespie, J. J. *et al.* The Rickettsia type IV secretion system: unrealized complexity mired by gene family expansion. *Pathog Dis* **74**, doi:10.1093/femspd/ftw058 (2016).

53 Gillespie, J. *et al.* Structural Insight into How Bacteria Prevent Interference between Multiple Divergent Type IV Secretion Systems. *MBio* **6**, e01867-01815, doi:10.1128/mBio.01867-15 (2015).

54 Verhoeve, V. I. & Gillespie, J. J. Origin of rickettsial host dependency unravelled. *Nat Microbiol* **7**, 1110-1111, doi:10.1038/s41564-022-01187-9 (2022).

55 Sutten, E. L. *et al.* Anaplasma marginale type IV secretion system proteins VirB2, VirB7, VirB11, and VirD4 are immunogenic components of a protective bacterial membrane vaccine. *Infect Immun* **78**, 1314-1325, doi:10.1128/IAI.01207-09 (2010).

56 Oyler, B. *et al.* Rickettsia typhi peptidoglycan mapping with data-dependent tandem mass spectrometry. *BioRxiv* (2023).

57 Figueroa-Cuilan, W. M. *et al.* Quantitative analysis of morphogenesis and growth dynamics in an obligate intracellular bacterium. *Mol Biol Cell*, mbcE23010023, doi:10.1091/mbc.E23-01-0023 (2023).

58 Atwal, S. *et al.* Evidence for a peptidoglycan-like structure in Orientia tsutsugamushi. *Mol Microbiol* **105**, 440-452, doi:10.1111/mmi.13709 (2017).

59 Atwal, S. *et al.* Discovery of a Diverse Set of Bacteria That Build Their Cell Walls without the Canonical Peptidoglycan Polymerase aPBP. *mBio* **12**, e0134221, doi:10.1128/mBio.01342-21 (2021).

60 Gillespie, J. J. & Salje, J. Orientia and Rickettsia: different flowers from the same garden. *Curr Opin Microbiol* **74**, 102318, doi:10.1016/j.mib.2023.102318 (2023).

61 Rennoll-Bankert, K. *et al.* RalF-Mediated Activation of Arf6 Controls Rickettsia typhi Invasion by Co-Opting Phosphoinositol Metabolism. *Infect Immun* **84**, 3496-3506, doi:10.1128/IAI.00638-16 (2016).

62 Lehman, S. *et al.* The Rickettsial Ankyrin Repeat Protein 2 Is a Type IV Secreted Effector That Associates with the Endoplasmic Reticulum. *mBio* **9**, e00975-00918, doi:10.1128/mBio.00975-18 (2018).

63 Voss, O. *et al.* Risk1, a Phosphatidylinositol 3-Kinase Effector, Promotes Rickettsia typhi Intracellular Survival. *mBio* **11**, doi:10.1128/mBio.00820-20 (2020).

64 Aistleitner, K., Clark, T., Dooley, C. & Hackstadt, T. Selective fragmentation of the trans-Golgi apparatus by Rickettsia rickettsii. *PLoS Pathog* **16**, e1008582, doi:10.1371/journal.ppat.1008582 (2020).

65 Lockwood, S. *et al.* Identification of Anaplasma marginale type IV secretion system effector proteins. *PLoS One* **6**, e27724, doi:10.1371/journal.pone.0027724 (2011).

66 Lin, M., den Dulk-Ras, A., Hooykaas, P. & Rikihisa, Y. Anaplasma phagocytophilum AnkA secreted by type IV secretion system is tyrosine phosphorylated by Abl-1 to facilitate infection. *Cell Microbiol* **9**, 2644-2657, doi:10.1111/j.1462-5822.2007.00985.x (2007).

67 Niu, H., Kozjak-Pavlovic, V., Rudel, T. & Rikihisa, Y. Anaplasma phagocytophilum Ats-1 is imported into host cell mitochondria and interferes with apoptosis induction. *PLoS Pathog* **6**, e1000774, doi:10.1371/journal.ppat.1000774 (2010).

68 Rikihisa, Y. & Lin, M. Anaplasma phagocytophilum and Ehrlichia chaffeensis type IV secretion and Ank proteins. *Curr Opin Microbiol* **13**, 59-66, doi:10.1016/j.mib.2009.12.008 (2010).

69 Rikihisa, Y., Lin, M. & Niu, H. Type IV secretion in the obligatory intracellular bacterium Anaplasma phagocytophilum. *Cell Microbiol* **12**, 1213-1221, doi:10.1111/j.1462-5822.2010.01500.x (2010).

70 Park, J. M. *et al.* An Anaplasma phagocytophilum T4SS effector, AteA, is essential for tick infection. *bioRxiv*, doi:10.1101/2023.02.06.527355 (2023).

71 Zhu, J. *et al.* Development of TEM-1 beta-lactamase based protein translocation assay for identification of Anaplasma phagocytophilum type IV secretion system effector proteins. *Sci Rep* **9**, 4235, doi:10.1038/s41598-019-40682-8 (2019).

72 Kim, Y., Wang, J., Clemens, E. G., Grab, D. J. & Dumler, J. S. Anaplasma phagocytophilum Ankyrin A Protein (AnkA) Enters the Nucleus Using an Importin-beta-, RanGTP-Dependent Mechanism. *Front Cell Infect Microbiol* **12**, 828605, doi:10.3389/fcimb.2022.828605 (2022).

73 Liu, H., Bao, W., Lin, M., Niu, H. & Rikihisa, Y. Ehrlichia type IV secretion effector ECH0825 is translocated to mitochondria and curbs ROS and apoptosis by upregulating host MnSOD. *Cell Microbiol* **14**, 1037-1050, doi:10.1111/j.1462-5822.2012.01775.x (2012).

74 Rikihisa, Y. The "Biological Weapons" of Ehrlichia chaffeensis: Novel Molecules and Mechanisms to Subjugate Host Cells. *Front Cell Infect Microbiol* **11**, 830180, doi:10.3389/fcimb.2021.830180 (2021).

75 Yan, Q. *et al.* Iron robbery by intracellular pathogen via bacterial effector-induced ferritinophagy. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2026598118 (2021).

76 Yan, Q. *et al.* Ehrlichia type IV secretion system effector Etf-2 binds to active RAB5 and delays endosome maturation. *Proc Natl Acad Sci U S A* **115**, E8977-E8986, doi:10.1073/pnas.1806904115 (2018).

77 Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* **13**, 2561, doi:10.1038/s41467-022-30269-9 (2022).

78 Rocha, E. P. C. & Bikard, D. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLoS Biol* **20**, e3001514, doi:10.1371/journal.pbio.3001514 (2022).

79 Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, doi:10.1126/science.aar4120 (2018).

80 Abby, S. S., Neron, B., Menager, H., Touchon, M. & Rocha, E. P. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One* **9**, e110726, doi:10.1371/journal.pone.0110726 (2014).

81 Tal, N. *et al.* Cyclic CMP and cyclic UMP mediate bacterial immunity against phages. *Cell* **184**, 5728-5739 e5716, doi:10.1016/j.cell.2021.09.031 (2021).

82 Letunic, I., Khedkar, S. & Bork, P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Research* **49**, D458-D460, doi:10.1093/nar/gkaa937 (2021).

83 Song, Y., Peisach, D., Pioszak, A. A., Xu, Z. & Ninfa, A. J. Crystal structure of the C-terminal domain of the two-component system transmitter protein nitrogen

regulator II (NRII; NtrB), regulator of nitrogen assimilation in Escherichia coli. *Biochemistry* **43**, 6670-6678 (2004).

84 Bellon, S. *et al.* Crystal structures of Escherichia coli topoisomerase IV ParE subunit (24 and 43 kilodaltons): a single residue dictates differences in novobiocin potency against topoisomerase IV and DNA gyrase. *Antimicrobial agents and chemotherapy* **48**, 1856-1864 (2004).

85 Wright, L. *et al.* Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chemistry & biology* **11**, 775-785 (2004).

86 Mascher, T., Helmann, J. D. & Unden, G. Stimulus perception in bacterial signal-transducing histidine kinases. *Microbiology and molecular biology reviews* **70**, 910-938 (2006).

87 Hagemann, M. *et al.* Identification of the DNA methyltransferases establishing the methylome of the cyanobacterium Synechocystis sp. PCC 6803. *DNA Research* **25**, 343-352 (2018).

88 Leipe, D. D., Aravind, L., Grishin, N. V. & Koonin, E. V. The bacterial replicative helicase DnaB evolved from a RecA duplication. *Genome research* **10**, 5-16 (2000).

89 Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**, D32-36, doi:10.1093/nar/gkj014 (2006).

90 Groth, A. C. & Calos, M. P. Phage integrases: biology and applications. *Journal of molecular biology* **335**, 667-678 (2004).

91 Abremski, K. E. & Hoess, R. H. Evidence for a second conserved arginine residue in the integrase family of recombination proteins. *Protein Engineering, Design and Selection* **5**, 87-91 (1992).

92 Gladyshev, E. A. & Arkhipova, I. R. A widespread class of reverse transcriptase-related cellular genes. *Proceedings of the National Academy of Sciences* **108**, 20311-20316 (2011).

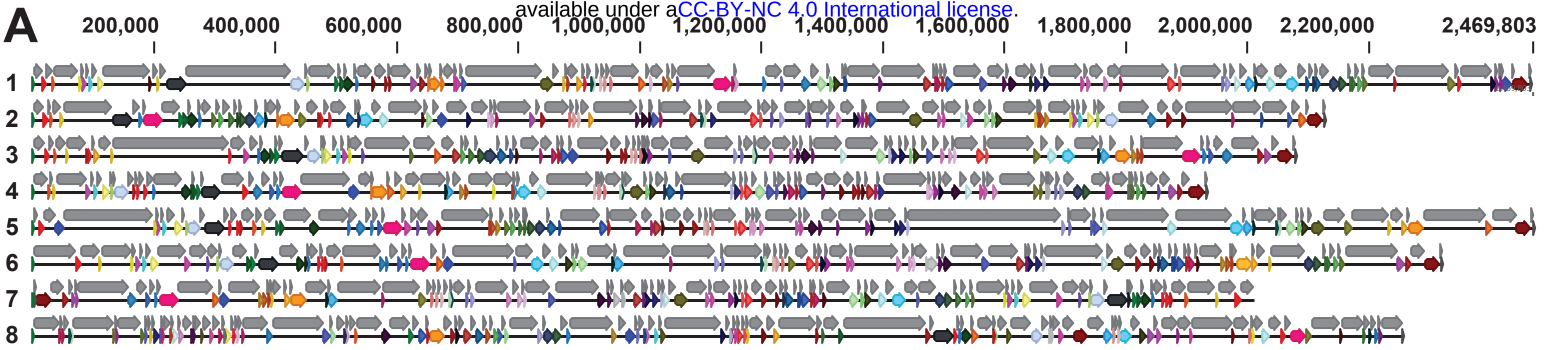
93 Blocker, F. J. *et al.* Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *Rna* **11**, 14-28 (2005).

94 Zhao, J. & Lambowitz, A. M. A bacterial group II intron-encoded reverse transcriptase localizes to cellular poles. *Proceedings of the National Academy of Sciences* **102**, 16133-16140 (2005).

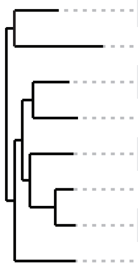
95 Nesmelova, I. V. & Hackett, P. B. DDE transposases: Structural similarity and diversity. *Advanced drug delivery reviews* **62**, 1187-1195 (2010).

96 Knizewski, L., Kinch, L. N., Grishin, N. V., Rychlewski, L. & Ginalski, K. Realm of PD-(D/E) XK nuclease superfamily revisited: detection of novel families with modified transitive meta profile searches. *BMC structural biology* **7**, 1-9 (2007).

97 Davies, D. R., Goryshin, I. Y., Reznikoff, W. S. & Rayment, I. Three-dimensional structure of the Tn 5 synaptic complex transposition intermediate. *Science* **289**, 77-85 (2000).



B



Ot strain	Total RAGE	Complete RAGE w/o truncated/ split genes	Complete RAGE w/ truncated/ split genes	Incomplete RAGE	No. IR regions	No. isolated mobilisation genes	No. isolated cargo genes	No. isolated HP-encoding genes
1 Gilliam	91	3	7	81	84	52	10	49
2 Boryong	84	0	1	83	84	25	17	65
3 UT76	71	0	2	69	84	31	11	52
4 UT176	72	0	3	69	84	42	9	40
5 Karp	77	0	13	64	84	25	13	39
6 Kato	81	1	10	70	84	33	12	54
7 Ikeda	76	0	0	76	84	36	16	57
8 TA686	93	0	7	86	84	33	12	46

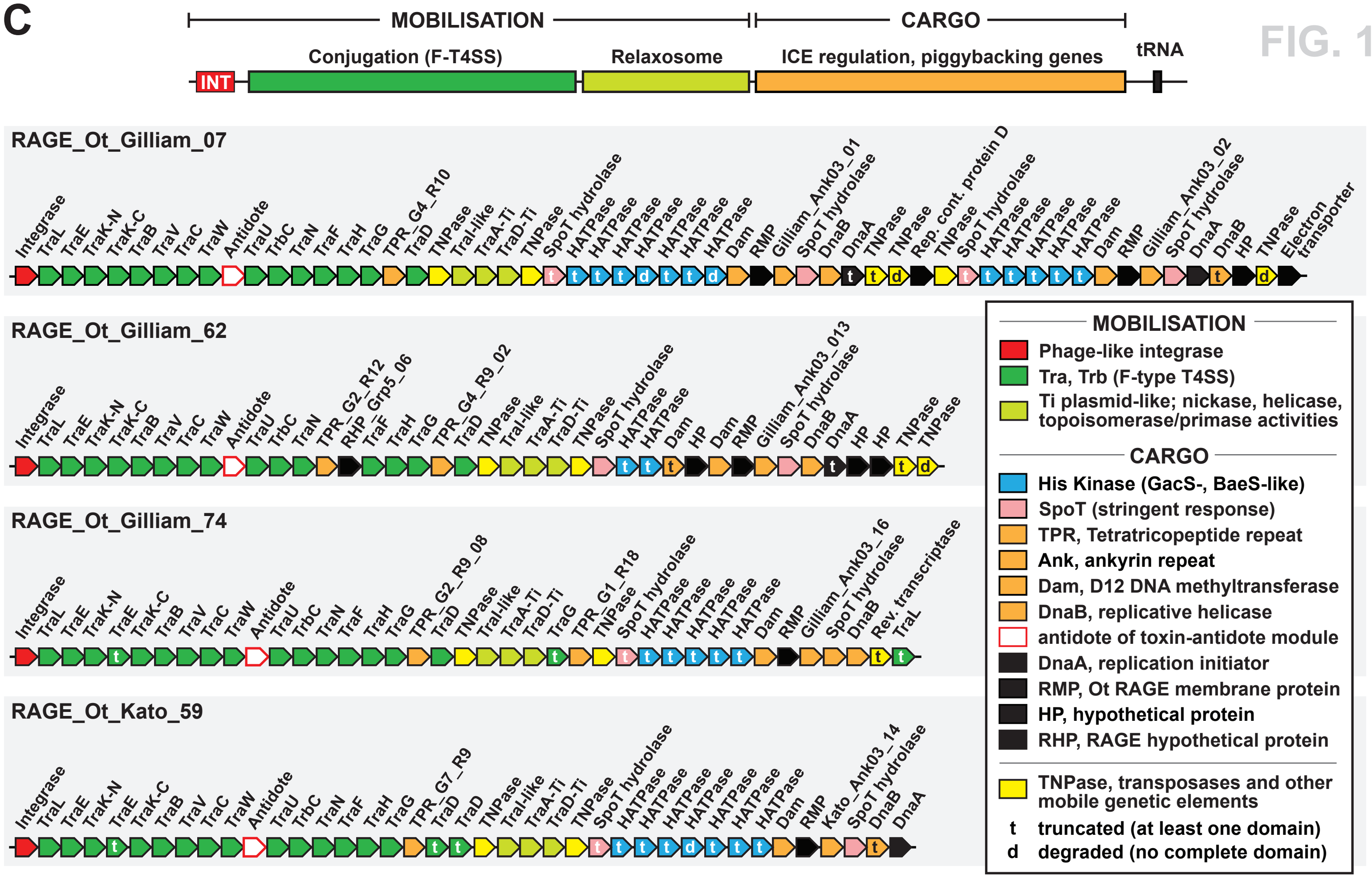


FIG. 1

A

🟡 single copy in RAGE 🟢 single copy in RAGE and inter-RAGE 🔵 Inter-RAGE
🔴 multi-copy in RAGE 🟣 multi-copy in RAGE and inter-RAGE x absent

RAGE-associated genes	Description	*Gilliam*	*Boryong*	*UT76*	*UT176*	*Karp*	*Kato*	*Ikeda*	*TA686*
18 kDa heat shock protein	ACD_sHsps-like: alpha-crystallin-type small heat shock protein	x	x	x	x	🟡	x	x	x
MidA family	SAM-dependent methyltransferase	🟢 2	🟢 2	🟢 2	🟢 2	🟢 2	🟢 2	🟢 2	🟢 2
TimA	Tim44/TimA family of putative adaptor protein (transport)	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1
glpE	Thiosulfate sulfurtransferase GlpE	x	x	x	x	🟡	x	x	x
cspC	Cold shock protein, CspA family	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
gcvT	Aminomethyltransferase (glycine degradation)	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
MdlB	ABC transporter ATP-binding protein (intracellular bacteria hallmark)	🟢 2	🔵 2	🟣 2,2	🔵 2	🟢 2	🟣 2,2	🟢 2	🟢 2
lipB	Octanoyltransferase (octanoyl-ACP to lipoate-dependent enzymes)	🟡	🟡	🔴 2	x	🟡	🟡	🟡	🟡
DDE_Tnp_1_3	Transposase DDE domain or inactive derivative	x	🟡	🟡	x	🟡	🟡	🟡	🔴 5
uvrA	Excinuclease ABC subunit A	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
pgsA	CDP-diacylglycerol-glycerol-3-P 3-phosphatidyltransferase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
tyrS	Tyrosine--tRNA ligase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
lacA	Ribose-5-phosphate isomerase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
dapD	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
ScaA	ScaA autotransporter protein	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
MdtL	Multidrug resistance protein MdtL	x	x	x	x	🟡	x	x	x
pld	Phospholipase D family protein	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
uvrD	DNA helicase II	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1	🟢 1
rpoZ	DNA-directed RNA polymerase subunit omega	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
bamB	Outer membrane protein assembly factor BamB	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
ssb	Single-stranded DNA-binding protein	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
PPDK	Pyruvate, phosphate dikinase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
map	Type I methionyl aminopeptidase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
ctaA	Heme A synthase (Cytochrome oxidase assembly protein)	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
clpP	ATP-dependent Clp protease proteolytic subunit	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
HipB	Transcriptional regulator, contains XRE-family HTH domain	🟣 2,3	🟢 4	🟢 4	🟢 4	🟢 4	🟢 4	🟢 4	🟢 4
ScaC	ScaC autotransporter (T5SS)	🟡	🟡	🟡	🟡	🟡	🟡	x	🟡
bcr/cflA	Bcr/CflA family drug resistance efflux transporter	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🔴 2
Sco1	Cytochrome oxidase Cu insertion factor	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
rne/rng	Ribonuclease E/G	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
trpS	Tryptophan--tRNA ligase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
SLC5-6-like_sbd (NT): HK (CT)	Sodium solute symporter fused to sensor HK/response regulator	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
rppH	RNA pyrophosphohydrolase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
ndk	Nucleoside-diphosphate kinase	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
cyaY	Iron donor protein CyaY	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
Dam	Site-specific DNA-adenine methylase	🟡	🟡	x	x	x	x	x	x
rSAM_XyeB	Radical SAM/SPASM domain peptide maturase, XyeB family	🟡	🟡	x	x	x	x	x	x
YifB	Predicted ATPase with chaperone activity	🟢 3	🟢 3	🔵 4	🔵 4	🔵 4	🔵 4	🔵 4	🟢 3
prmA	Ribosomal protein L11 methyltransferase	x	x	🟡	x	x	🟡	🟡	x
Abhydrolase_6	Alpha/beta hydrolase family (MhpC-like)	x	x	x	x	x	🟡	🟡	x
prfB|prfB	Peptide chain release factor RF-2	x	x	x	x	x	🟡	x	x
gpgP	Glucosyl-3-phosphoglycerate phosphatase	x	x	x	x	x	🟡	x	x
HPD	4-hydroxyphenylpyruvate dioxygenase (tyrosine degradation)	x	x	x	x	x	x	x	🟡
sfp	4'-phosphopantetheinyl transferase sfp	x	x	x	x	x	x	x	🟡
glyA	Serine hydroxymethyltransferase	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🟡
acoA	Acetoin:2,6-dichlorophenolindophenol oxidoreductase subunit alpha	x	x	x	x	x	x	x	🟡
HNHc	Endonuclease signature (viral, prokaryotic, and eukaryotic proteins)	x	x	x	x	x	x	x	🟡
fol_rel_CADD	Putative folate metabolism protein, CADD family	x	x	x	x	x	x	x	🟡
xseB	Exodeoxyribonuclease VII small subunit	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🟢 1
ProP	Proline/glycine betaine transporter	🔵 6	🔵 6	🔵 6	🔵 6	🟢 6	🔵 6	🔵 6	🟢 5
ltrA	Group II intron-encoded protein LtrA	x	x	x	x	x	x	x	🟡
lysS	lysine--tRNA ligase	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🟢 1
FolP	Dihydropteroate synthase	x	x	x	x	x	x	x	🟡
aroK	Shikimate kinase 1	x	x	x	x	x	x	x	🟡
menG	2-phytyl-1,4-naphtoquinone methyltransferase	x	x	x	x	x	x	x	🟡
dnaQ	DNA polymerase III subunit epsilon	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🔵 1	🟢 1
gpmA	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	x	x	x	🟡	x	x	x	x
pthXo1	pthXo1 CDS, TAL effector protein PthXo1	x	x	x	🟡	x	x	x	x
secA-like	preprotein translocase SecA subunit-like protein	🔴 5	🟡	🔴 3	🟡	🔴 8	🔴 3	🔴 3	🔴 8
SenC	Required for optimal cytochrome c oxidase activity	x	x	x	🟡	x	x	x	x
mta	HTH-type transcriptional activator mta	x	x	x	🟡	x	x	x	x
prfB	peptide chain release factor RF-2	x	x	x	x	x	🟡	🟡	x

B

MOBILISATION
CARGO

		Gilliam	*Boryong*	*UT76*	*UT176*	*Karp*	*Kato*	*Ikeda*	*TA686*
MOBILISATION	Phage-like integrases (RAGE insertion)	95	77	99	58	89	102	83	98
	Conjugation and relaxosome proteins (*tra* genes)	647	664	607	421	686	621	512	572
CARGO (high copy): highly abundant	Membrane proteins unique to RAGE	32	56	24	22	41	30	26	27
	Dam, D12 DNA methyltransferases	33	34	29	18	30	26	18	22
	DNA replicative helicases	40	54	46	24	52	50	38	53
	Multidrug Resistance Proteins/Histidine Kinases	103 (5)	57 (10)	103 (7)	50 (7)	119 (7)	95 (7)	82 (8)	129 (7)
	SpoT stringent response regulators	64	54	64	36	78	62	62	49
	Hypothetical Protein-encoding genes	462	547	419	324	297	464	417	473
CARGO (high copy): put. effectors	Tetratricopeptide epeat containing proteins	48	29	34	31	39	45	21	31
	Ankyrin repeat containing proteins	62	66	63	53	75	53	47	56
CARGO (high copy): less abundant	Peroxiredoxin	2	2	2	3	5	7	4	5
	XthA2 (exodeoxyribonuclease III)	2	3 (1)	4 (1)	2 (1)	4 (1)	2 (1)	4 (1)	2
	Repeat-containing protein D	5 (1)	9	3	4 (1)	3 (1)	2 (1)	6 (1)	2 (1)
Mobile Genetic Elements	Transposase and inactive derivatives (numerous types)	634	373	275	531	361	279	350	480
	Reverse transcriptase	23	64	12	7	20	38	36	38
	HNH endonucleases	1	2	2	0	22	9	1	18

FIG. 2

FIG. 3

A Membrane proteins

Ot strain	YccA, modulator of FtsH protease	Vut1, vitamin uptake transporter	RhaT	Ot RAGE membrane protein
Gilliam	1	1	1	30
Boryong	1	1	1	18
UT76	1	1	1	23
UT176	1	1	1	20
Karp	1	1	1	38
Kato	1	1	1	28
Ikeda	1	1	1	23
TA686	1	1	1	23

B Dam, D12 DNA methyltransferases

Ot strain	Full length	Truncated	Degraded	Total
Gilliam	26	7	-	33
Boryong	8	26	-	34
UT76	11	16	2	29
UT176	11	7	-	18
Karp	18	11	1	30
Kato	13	12	1	26
Ikeda	10	8	-	18
TA686	13	2	7	22

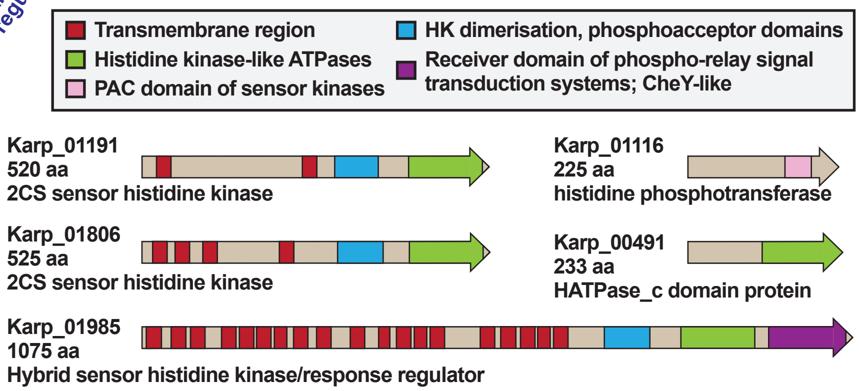
C DNA helicases

Ot strain	UvrD	DnaB, full length	DnaB, degraded	DnaB, full length	Total
Gilliam	2	21	15	2	40
Boryong	2	1	40	11	54
UT76	2	12	18	14	46
UT176	2	6	12	4	24
Karp	2	23	19	8	52
Kato	2	15	31	2	50
Ikeda	2	13	19	4	38
TA686	2	21	10	20	53

D

Ot strain	MRP/NBP35 family ATP binding protein	ABC membrane and AAA-ATPases	PutP Na/Pro symporter	PanF Na/panto-thenate symporter	HATPase_c domain protein	HATPase_c dom. protein (degraded)	His phospho-transferase	Two component sensor His kinase	Hybrid His kinase response regulator
Gilliam	1	1	4	2	79	18	1	2	1
Boryong	1	1	4	2	38	18	1	2	1
UT76	1	1	5	2	68	30	1	2	1
UT176	1	1	5	2	35	10	1	2	1
Karp	1	1	5	2	90	23	1	2	1
Kato	1	1	8	2	83	30	1	2	1
Ikeda	1	1	7	2	56	20	1	2	1
TA686	1	1	5	2	78	45	1	2	1

Multidrug Resistance Protein (MRP) and Histidine Kinases


🟥 Transmembrane region
🟩 Histidine kinase-like ATPases
▨ PAC domain of sensor kinases
🟦 HK dimerisation, phosphoacceptor domains
🟪 Receiver domain of phospho-relay signal transduction systems; CheY-like

Karp_01191
520 aa
2CS sensor histidine kinase

Karp_01116
225 aa
histidine phosphotransferase

Karp_01806
525 aa
2CS sensor histidine kinase

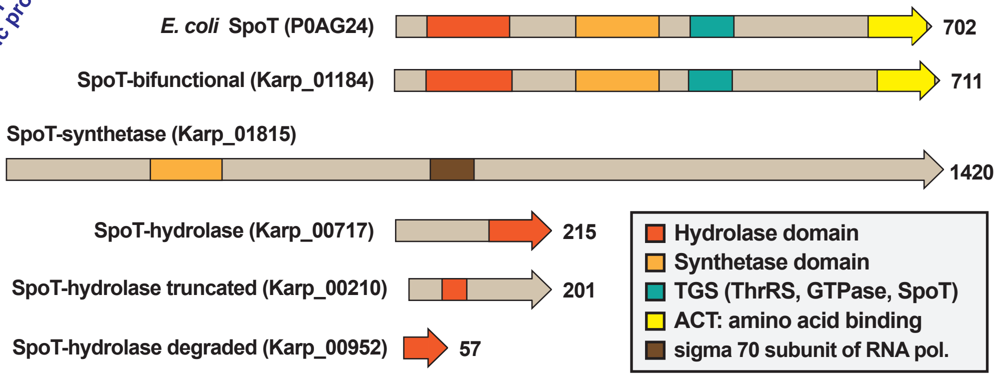
Karp_00491
233 aa
HATPase_c domain protein

Karp_01985
1075 aa
Hybrid sensor histidine kinase/response regulator

E

Ot strain	SpoT-bifunctional (full-length)	SpoT-synthetase (full-length)	SpoT-hydrolase (full-length)	SpoT-hydrolase (truncated)	SpoT-hydrolase (degraded)	HATPase-SpoT chimeric protein	MRC-SpoT chimeric protein
Gilliam	1	1	42	8	12	-	-
Boryong	1	1	21	5	26	-	-
UT76	1	1	23	9	27	1	2
UT176	1	1	18	5	11	-	-
Karp	1	1	46	11	19	-	-
Kato	1	1	25	15	19	1	-
Ikeda	1	1	16	6	38	-	-
TA686	1	1	15	5	23	2	2

SpoT stringent response regulators


E. coli SpoT (P0AG24) 702
SpoT-bifunctional (Karp_01184) 711
SpoT-synthetase (Karp_01815) 1420
SpoT-hydrolase (Karp_00717) 215
SpoT-hydrolase truncated (Karp_00210) 201
SpoT-hydrolase degraded (Karp_00952) 57

🟧 Hydrolase domain
🟨 Synthetase domain
🟩 TGS (ThrRS, GTPase, SpoT)
🟨 ACT: amino acid binding
🟫 sigma 70 subunit of RNA pol.

F HP, Ot RAGE hypothetical proteins

Ot strain	DnaA_N	Phage portal protein	Peptidase C48	Peptidase M61	Predicted lipase	Zinc ribbon 4	RHOD	AHH	FISH1 Ser hydrolase	MagZ-like	HNHc nuclease	na/nt deaminase	BrkB-like	CdAMP_rec	PDu(A)C	Recombinase	Rvt_1(PF00078)	Rvt_N 19	RAGE_hypo_Gr1 DUF167	RAGE_hypo_Gr2 DUF155	RAGE_hypo_Gr3 DUF3857	RAGE_hypo_Gr4 DUF2659	RAGE_hypo_Gr5 DUF4065	RAGE_hypo_Gr6 DUF2610	RAGE_hypo_Gr7 unknown domains	Total
Gilliam	23	1	-	-	-	1	1	-	1	-	-	1	-	1	3	-	-	-	1	1	2	-	8	-	418	462
Boryong	12	1	1	-	-	1	1	-	1	-	-	1	1	-	-	-	-	1	-	1	-	-	2	-	525	547
UT76	15	1	1	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	1	2	1	5	-	391	419
UT176	10	1	1	1	1	1	1	-	-	1	-	3	-	-	-	-	-	-	-	1	2	1	7	-	293	324
Karp	12	1	1	-	-	1	1	1	-	-	-	-	-		-	1	-	-	-	1	2	1	6	1	280	308
Kato	20	1	1	-	-	1	1	-	1	-	-	-	1	1	1	-	-	-	-	1	-	-	-	-	436	464
Ikeda	17	1	1	-	-	1	1	-	-	-	-	3	1	1	-	-	-	-	-	1	2	1	2	-	385	417
TA686	16	1	-	-	1	1	1	2	-	-	1	-	-	1	-	-	1	-	-	1	2	-	8	-	437	473

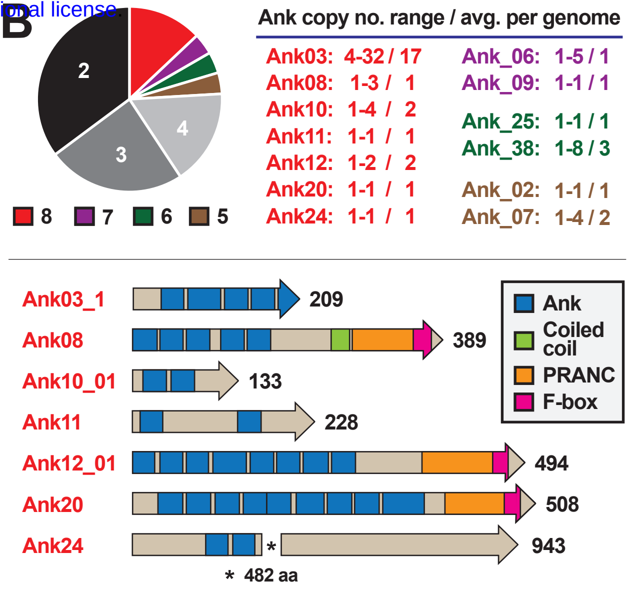
FIG. 4

A

Ankyrin repeat containing proteins (*n* = 129)

Ot strain	orthologous groups (*n* = 54)						singletons (*n* = 75)					
	No. orthologous groups (OGs)	Tot. proteins (% in OGs)	CC, PRANC, F-box	CC or PRANC, F-box	F-box only	w/o F-box	No. singletons	Tot. proteins	CC, PRANC, F-box	CC or PRANC, F-box	F-box only	w/o F-box
Gilliam	25	48, 79%	5	5	24	14	12	15	-	2	5	8
Boryong	22	36, 67%	4	3	7	22	21	28	3	2	2	21
UT76	36	53, 84%	6	8	20	19	8	11	2	2	5	2
UT176	29	51, 94%	2	9	22	18	3	3	-	-	-	3
Karp	24	66, 88%	3	12	35	16	7	9	-	1	2	6
Kato	26	51, 93%	6	5	24	16	4	4	1	-	-	3
Ikeda	22	49, 88%	10	10	20	9	7	7	-	-	-	7
TA686	20	40, 70%	5	5	23	7	14	17	4	2	4	7

B



C

TPR repeat containing proteins

Ot strain	TPR Group									Total proteins
	1	2	3	4	5	6	7	8	9	
Gilliam	14	17	2	6	-	1	3	5	-	48
Boryong	7	12	6	1	-	1	1	1	-	29
UT76	7	15	2	3	2	1	-	2	2	34
UT176	9	15	4	1	-	1	-	1	-	31
Karp	16	15	3	4	-	1	-	-	-	39
Kato	11	18	2	5	2	1	1	3	2	45
Ikeda	5	5	3	2	2	-	-	2	2	21
TA686	11	15	-	2	-	1	-	2	-	31
Repeat range	1-18	1-16	1-9	3-11	2	5-11	4-9	1-10	1-2	

D

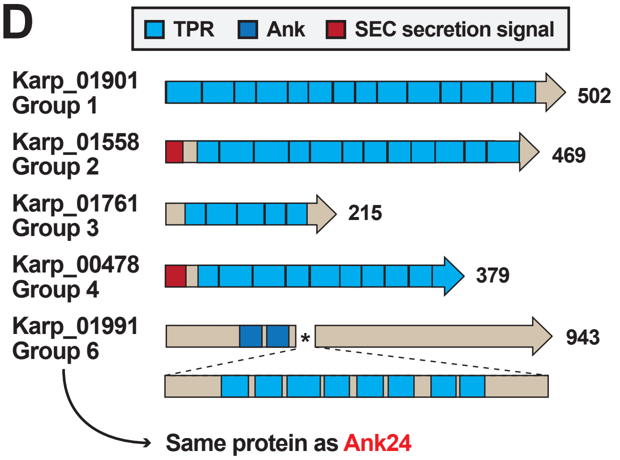


FIG. 5

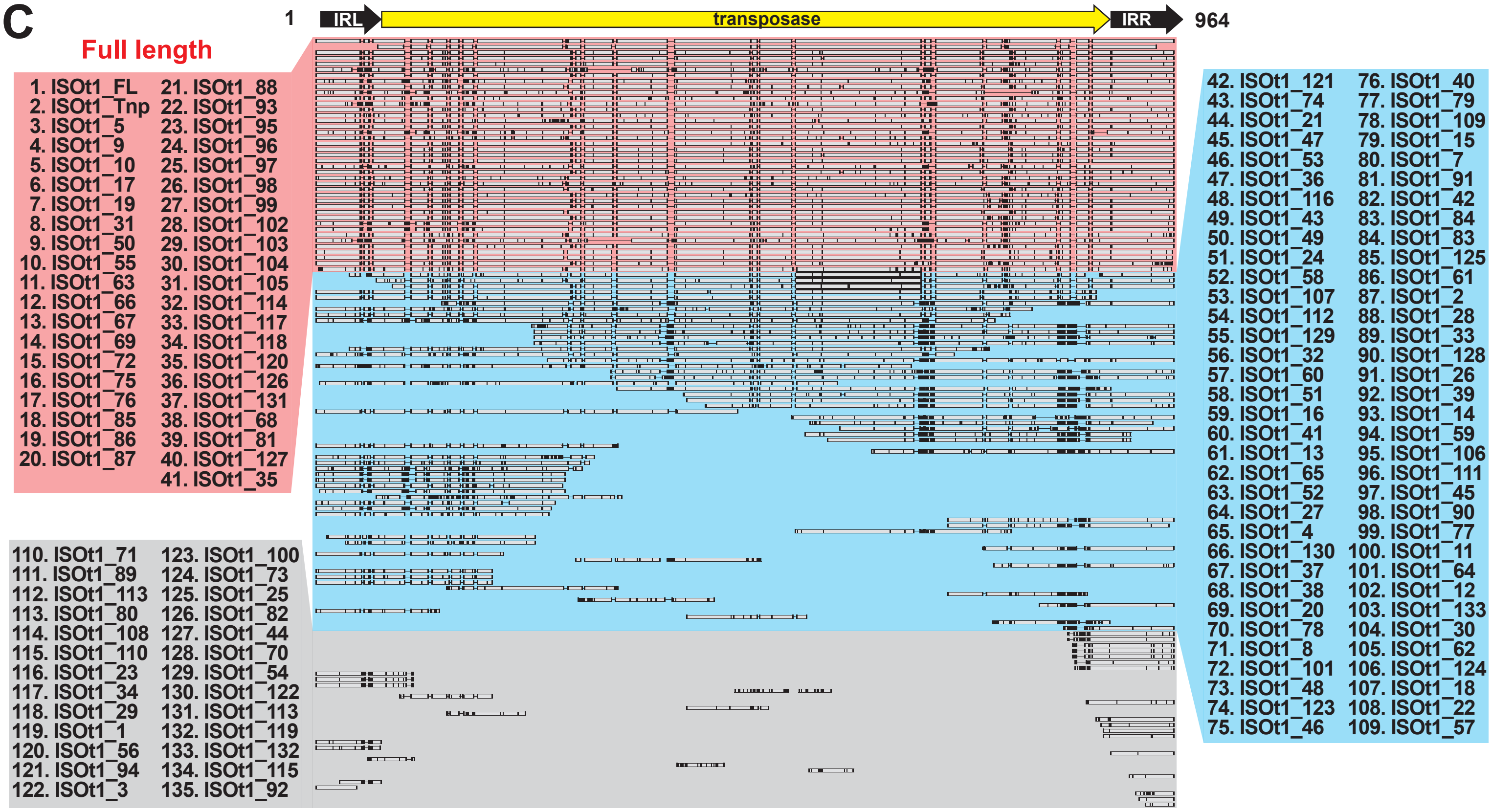
A

Ot strain	Integrases: Full lenth	Integrases: Truncated	Integrases: No domain	Integrases: Total	RTs: Full lenth	RTs: Truncated	RTs: No domain	RTs: Total	ISOt3: Full lenth	ISOt5: Full lenth
Gilliam	10	60	25	95	0	4	18	23	115	3
Boryong	13	30	34	77	0	47	17	64	3	9
UT76	5	54	40	99	1	4	7	12	7	0
UT176	4	30	24	58	0	2	5	7	120	13
Karp	6	50	33	89	9	6	5	20	0	14
Kato	10	54	38	102	0	31	7	38	7	7
Ikeda	3	46	34	83	0	29	7	36	33	70
TA686	8	61	29	98	1	28	9	38	9	10

B

	IS in Ikeda and Karp									IS unique to Karp						
										Karp Tnp group...						
Ot strain	IS5 fam. tnp ISOt1	mISOt1	IS630 fam. tnp ISOt2	mISOt2a	mISOt2b	IS630 fam. tnp ISOt3	IS982 fam. tnp ISOt4	mISOt4	IS110 fam. tnp ISOt5	1	2	3	4	5	6	7
Ikeda	113	61	17	30	65	33	124	55	70	-	-	-	-	-	-	-
Karp:																
Total	133	71	27	37	28	17	154	65	59	35	31	4	4	1	1	1
Full len.	41	70	6	14	35	-	14	16	14							

C



D

Ot strain	Region	Region length	Completeness	Specific Keywords	Tot. proteins	Breakdown: Phage	Breakdown: HP	Breakdown: Bacteria	Phage, HP	*att*	Genome coordinates
Gilliam	1	22.4	Incomplete	Int, Tnp	15	8	3	4	73.3%	Y	934067-956536
	2	18.3	Questionable	Int, Tnp, Head	19	9	6	4	78.9%	Y	1286362-1304702
Boryong	none	-	-	-	-	-	-	-	-	-	-
UT76	1	24.4	Incomplete	Int, Tnp, Head	15	7	1	7	53.3%	Y	1758287-1782760
UT176	1	5.5	Incomplete	Cap, Head	8	6	1	1	87.5%	N	1197907-1203463
	2	6.6	Incomplete	Tnp, Pro	7	6	0	1	85.7%	N	1713761-1720444
Karp	1	35.6	Intact	Int, Tnp	26	10	7	9	65.3%	Y	1645670-1681300
Kato	1	22.5	Incomplete	Int, Tnp, Head	9	5	0	4	55.5%	Y	967021-989612
Ikeda	1	42.1	Questionable	Int, Tnp, Env	21	10	4	7	66.6%	Y	1232656-1274803
	2	26.9	Questionable	Int, Tnp	24	9	4	11	54.1%	Y	1543584-1570567
TA686	1	36.1	Intact	Int, Tnp, Head, Cap	30	15	0	15	50.0%	Y	474458-510588
	2	43.5	Intact	Int, Tnp	38	14	0	24	36.8%	Y	1131201-1157620
	3	26.4	Questionable	Int, Tnp, Plate	14	6	0	8	42.8%	Y	934067-956536

E

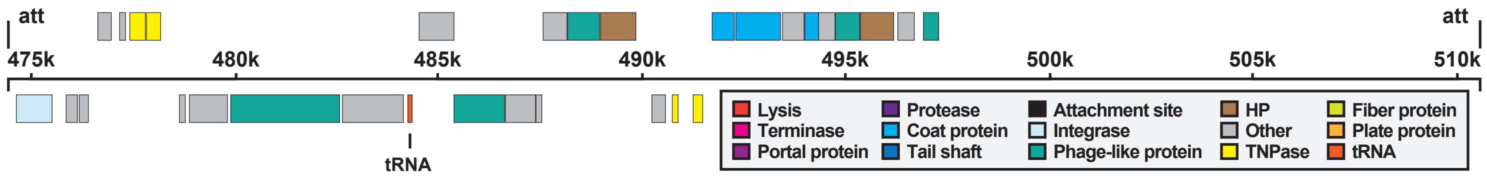
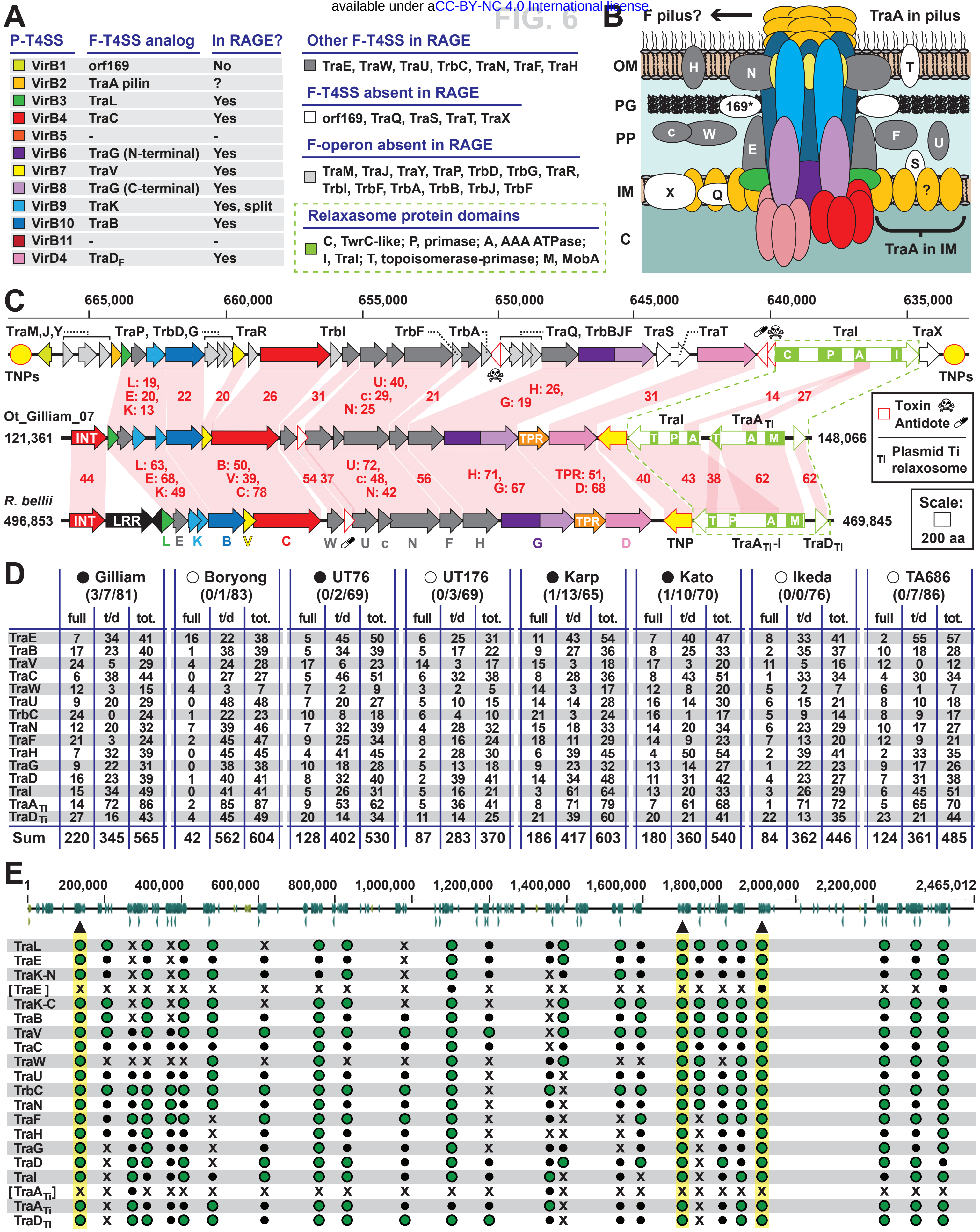


FIG. 6



A

P-T4SS	F-T4SS analog	In RAGE?
🟨 VirB1	orf169	No
🟧 VirB2	TraA pilin	?
🟩 VirB3	TraL	Yes
🟥 VirB4	TraC	Yes
🟧 VirB5	-	-
🟪 VirB6	TraG (N-terminal)	Yes
🟨 VirB7	TraV	Yes
🟪 VirB8	TraG (C-terminal)	Yes
🟦 VirB9	TraK	Yes, split
🟦 VirB10	TraB	Yes
🟥 VirB11	-	-
🟪 VirD4	TraD_F	Yes

Other F-T4SS in RAGE

■ TraE, TraW, TraU, TrbC, TraN, TraF, TraH

F-T4SS absent in RAGE

□ orf169, TraQ, TraS, TraT, TraX

F-operon absent in RAGE

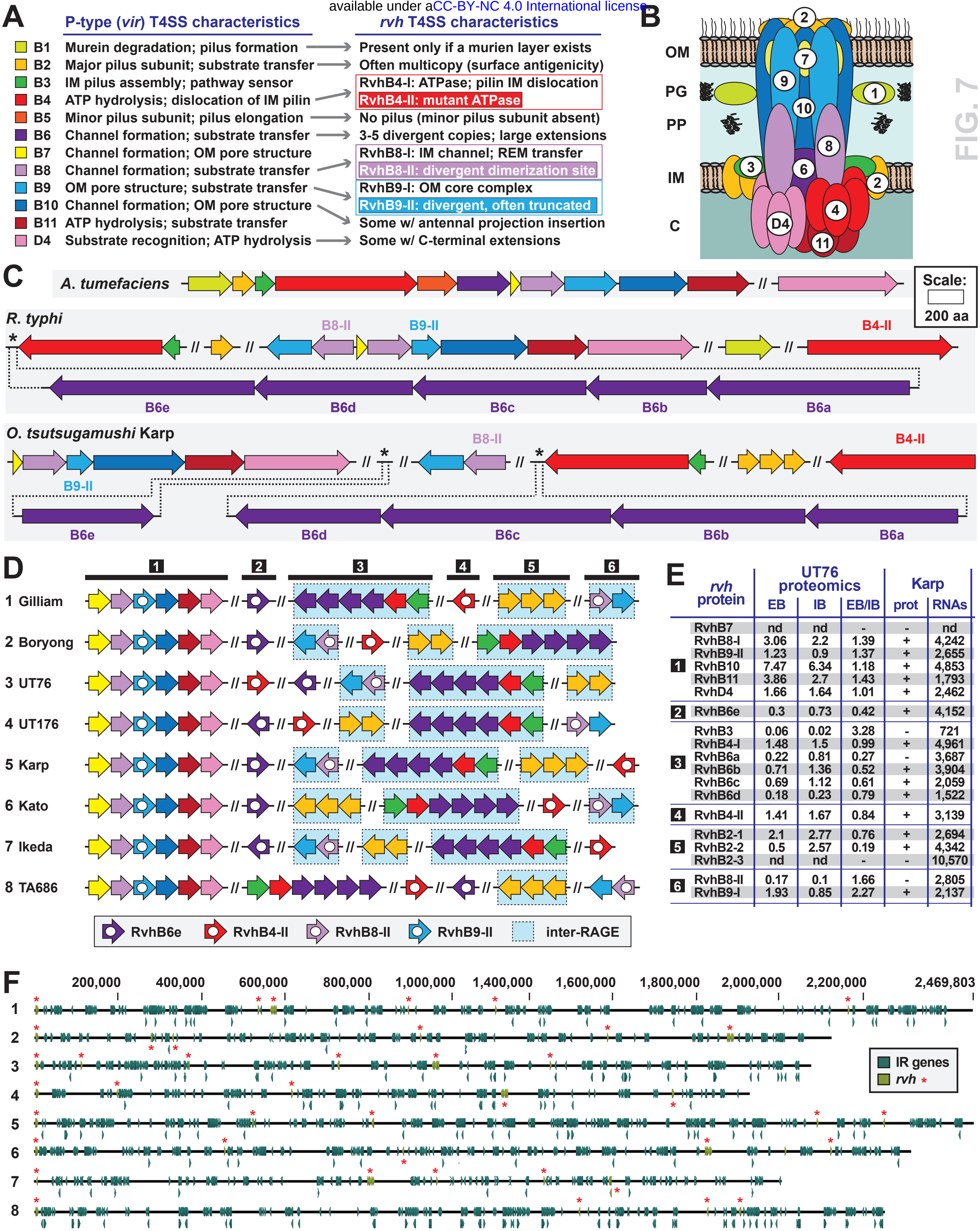
▨ TraM, TraJ, TraY, TraP, TrbD, TrbG, TraR, TrbI, TrbF, TrbA, TrbB, TrbJ, TrbF

Relaxasome protein domains

🟩 C, TwrC-like; P, primase; A, AAA ATPase; I, TraI; T, topoisomerase-primase; M, MobA

	● Gilliam (3/7/81)			○ Boryong (0/1/83)			● UT76 (0/2/69)			○ UT176 (0/3/69)			● Karp (1/13/65)			● Kato (1/10/70)			○ Ikeda (0/0/76)			○ TA686 (0/7/86)		
	full	t/d	tot.	full	t/d	tot.	full	t/d	tot.	full	t/d	tot.	full	t/d	tot.	full	t/d	tot.	full	t/d	tot.	full	t/d	tot.
TraE	7	34	41	16	22	38	5	45	50	6	25	31	11	43	54	7	40	47	8	33	41	2	55	57
TraB	17	23	40	1	38	39	5	34	39	5	17	22	9	27	36	8	25	33	2	35	37	10	18	28
TraV	24	5	29	4	24	28	17	6	23	14	3	17	15	3	18	17	3	20	11	5	16	12	0	12
TraC	6	38	44	0	27	27	5	46	51	6	32	38	8	28	36	8	43	51	1	33	34	4	30	34
TraW	12	3	15	4	3	7	7	2	9	3	2	5	14	3	17	12	8	20	5	2	7	6	1	7
TraU	9	20	29	0	48	48	7	20	27	5	10	15	14	14	28	16	14	30	6	15	21	8	10	18
TrbC	24	0	24	1	22	23	10	8	18	6	4	10	21	3	24	16	1	17	5	9	14	8	9	17
TraN	12	20	32	7	39	46	7	32	39	4	28	32	15	18	33	14	20	34	6	23	29	10	17	27
TraF	21	3	24	2	45	47	9	25	34	8	16	24	18	11	29	14	9	23	7	13	20	12	9	21
TraH	7	32	39	0	45	45	4	41	45	2	28	30	6	39	45	4	50	54	2	39	41	2	33	35
TraG	9	22	31	0	38	38	10	18	28	5	13	18	9	23	32	13	14	27	1	22	23	9	17	26
TraD	16	23	39	1	40	41	8	32	40	2	39	41	14	34	48	11	31	42	4	23	27	7	31	38
TraI	15	34	49	0	41	41	5	26	31	5	16	21	3	61	64	13	20	33	3	26	29	6	45	51
TraA_{Ti}	14	72	86	2	85	87	9	53	62	5	36	41	8	71	79	7	61	68	1	71	72	5	65	70
TraD_{Ti}	27	16	43	4	45	49	20	14	34	11	14	25	21	39	60	20	21	41	22	13	35	23	21	44
Sum	220	345	565	42	562	604	128	402	530	87	283	370	186	417	603	180	360	540	84	362	446	124	361	485

FIG. 7



A

	P-type (*vir*) T4SS characteristics	*rvh* T4SS characteristics
B1	Murein degradation; pilus formation	Present only if a murien layer exists
B2	Major pilus subunit; substrate transfer	Often multicopy (surface antigenicity)
B3	IM pilus assembly; pathway sensor	RvhB4-I: ATPase; pilin IM dislocation
B4	ATP hydrolysis; dislocation of IM pilin	RvhB4-II: mutant ATPase
B5	Minor pilus subunit; pilus elongation	No pilus (minor pilus subunit absent)
B6	Channel formation; substrate transfer	3-5 divergent copies; large extensions
B7	Channel formation; OM pore structure	RvhB8-I: IM channel; REM transfer
B8	Channel formation; substrate transfer	RvhB8-II: divergent dimerization site
B9	OM pore structure; substrate transfer	RvhB9-I: OM core complex
B10	Channel formation; OM pore structure	RvhB9-II: divergent, often truncated
B11	ATP hydrolysis; substrate transfer	Some w/ antennal projection insertion
D4	Substrate recognition; ATP hydrolysis	Some w/ C-terminal extensions

E

	rvh protein	UT76 proteomics EB	UT76 proteomics IB	UT76 proteomics EB/IB	Karp prot	Karp RNAs
1	RvhB7	nd	nd	-	-	nd
	RvhB8-I	3.06	2.2	1.39	+	4,242
	RvhB9-II	1.23	0.9	1.37	+	2,655
	RvhB10	7.47	6.34	1.18	+	4,853
	RvhB11	3.86	2.7	1.43	+	1,793
	RvhD4	1.66	1.64	1.01	+	2,462
2	RvhB6e	0.3	0.73	0.42	+	4,152
3	RvhB3	0.06	0.02	3.28	-	721
	RvhB4-I	1.48	1.5	0.99	+	4,961
	RvhB6a	0.22	0.81	0.27	-	3,687
	RvhB6b	0.71	1.36	0.52	+	3,904
	RvhB6c	0.69	1.12	0.61	+	2,059
	RvhB6d	0.18	0.23	0.79	+	1,522
4	RvhB4-II	1.41	1.67	0.84	+	3,139
5	RvhB2-1	2.1	2.77	0.76	+	2,694
	RvhB2-2	0.5	2.57	0.19	+	4,342
	RvhB2-3	nd	nd	-	-	10,570
6	RvhB8-II	0.17	0.1	1.66	-	2,805
	RvhB9-I	1.93	0.85	2.27	+	2,137