Dual RNA-seq provides insight into the biology of the neglected intracellular human pathogen *Orientia tsutsugamushi*

Bozena Mika-Gospodorz[#a], Suparat Giengkam[#b], Alexander J. Westermann[a,c], Jantana Wongsantichon[b], Willow Kion-Crosby[d], Suthida Chuenklin[b], Loo Chien Wang[e], Piyanate Sunyakumthorn[f], Radoslaw Sobota[e], Selvakumar Subbian[g], Jörg Vogel[a,c], Lars Barquist[#*a,g] and Jeanne Salje[#*b,h,i]

Running title: RNA sequencing of *Orientia tsutsugamushi*

Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI), Würzburg, Germany[a]

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

Institute for Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany[c]

Rutgers, the State Univeristy of New Jersey, New Jersey, USA[d]

Functional Proteomics Laboratory, Institute of Molecular and Cell Biology, Agency for Science, Technology and Research (A*STAR), Singapore[e].

Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand[f];

Faculty of Medicine, University of Würzburg, Würzburg, Germany[g];

Public Health Research Institute, Rutgers University, New Jersey, USA[h];

Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom[i];

[#] these authors contributed equally

*address correspondence to js2522@njms.rutgers.edu; lars.barquist@helmholtz-hiri.de

Summary

Emerging and neglected pathogens pose challenges as their biology is frequently poorly understood, and genetic tools often do not exist to manipulate them. Organism agnostic sequencing technologies offer a promising approach to understand the molecular processes underlying these diseases. Here we apply dual RNA-seq to *Orientia tsutsugamushi* (Ot), the obligate intracellular causative agent of the vector-borne human disease scrub typhus. Half the Ot genome is composed of repetitive DNA, and there is minimal collinearity in gene order between strains. Integrating RNA-seq, comparative genomics, proteomics, and machine learning, we investigated the transcriptional architecture of Ot, including operon structure and non-coding RNAs, and found evidence for wide-spread post-transcriptional antisense regulation. We compared the host response to two clinical isolates and identified distinct immune response networks that are up-regulated in response to each strain, leading to predictions of relative virulence which were confirmed in a mouse infection model. Thus, dual RNA-seq can reveal the biology and host-pathogen interactions of a poorly characterized and genetically intractable organism such as Ot.

Keywords: neglected and emerging pathogens, intracellular bacteria, dual RNAseq, transcriptomics, host-pathogen cell biology, bacterial virulence

Main Text

Introduction

Improved surveillance and diagnostics have led to the recognition of previously neglected bacteria as serious pathogens, whilst human population growth, globalization and increased travel have contributed to outbreaks of new pathogens and changing patterns of infectious disease. The biology of neglected and emerging pathogens is often poorly characterized but is essential for understanding of their pathogenicity and developing therapeutic and preventative strategies. Bacterial pathogens that employ an obligate intracellular lifestyle present additional challenges, because many cause diseases that are difficult to detect and because they are difficult to manipulate experimentally. As a result, many obligate intracellular bacteria are emerging or neglected.

Obligate intracellular bacteria include the Rickettsiales, an order which includes the insect symbiont *Wolbachia* as well as a number of human and veterinary pathogens. *Orientia tsutsugamushi* (Ot, Order Rickettsiales Family Rickettsiaceae) is an obligate intracellular bacterium that causes the mite-borne human disease scrub typhus. Scrub typhus is a leading cause of severe febrile illness in the Asia Pacific [1], a region including approximately two thirds of the world's population. Locally acquired cases in the Middle East and Latin America suggest that this disease may be more widespread than previously appreciated [2,3]. Under-recognition and under-reporting are a major problem in scrub typhus because unambiguous diagnosis is difficult, and awareness is low amongst many clinicians. Scrub typhus disease causes headache, fever, rash, and lymphadenopathy beginning 7-14 days after inoculation via a feeding larval stage mite. If untreated this can progress to cause multiple organ failure and death.

Ot colonizes the ovaries and salivary glands in the mite vector. During acute infection of its mammalian host, the bacteria infect endothelial cells, dendritic cells and monocytes/macrophages at the mite bite site [4], and then disseminate via blood and lymphatic vessels to multiple organs including lung, liver, kidney, spleen and brain [5].

The genome of Ot is highly repetitive, with many mobile genetic elements [6,7]. Gene loss associated with adaptation to the intracellular niche means that obligate intracellular bacteria tend to have highly reduced genomes of ~1 Mbp. The genome of Ot is large by comparison, at 1.93-2.47 Mbp (depending on the strain), almost half of which is composed of repetitive DNA sequences. These repetitive elements include mobile genetic elements such as the integrative and conjugative element (ICE) RAGE (rickettsial amplified genetic element), palindromic repeats, and other transposable elements. Even closely related strains exhibit minimal genome synteny, suggesting an active process of genome recombination [8].

Ot strains are highly variable in terms of antigenicity and virulence. Hundreds of strains have been described based on differences in the sequence of the surface protein TSA56 [9–11]. These strains are frequently classified into one of 7 geographically diverse serotype groups, dominated by the 3 most prevalent groups Karp, Kato and Gilliam [12]. However, whole genome sequencing has revealed that serotype-based groupings do not necessarily reflect phylogenetic relationships [8]. Different strains of Ot exhibit different levels of pathogenicity in mouse models as measured by disease severity and bacterial levels [13–15]. This is dependent on both bacterial strain and host genotype. For example, strain Karp (serotype group Karp) causes lethal infection in BALBc and C3H/He mice at low doses, strain Gilliam (serotype group Gilliam) causes lethal infection in

C3H/He but not BALBc mice at similar doses, whilst strain TA716 (serotype group TA716) does not cause lethal infection in either strain at similar doses [14,16]. Ot infection stimulates a pro-inflammatory immune response [17–21] and levels of some proinflammatory cytokines, such as IL-8, correlate with disease severity in humans [22].

The transcriptome of an organism can reveal details of its physiology, and transcriptomics has been used to study bacterial pathogenesis [23]. Dual RNAseq quantifies RNA transcripts of intracellular pathogens and host cells in a single experiment [24,25]. For example, dual RNA-seq has been used to study obligate intracellular *Chlamydia trachomatis*[26] revealing the rewiring of *Chlamydia* metabolism during the onset of an infection of human epithelial cells, together with the corresponding host responses.

Here we apply dual RNA-seq to a much less studied obligate intracellular pathogen, Ot. We measured expression of genes encoded in the mobile genetic elements and in the core genome of Ot, and then we explored the basis of pathogenicity in scrub typhus by comparing the gene expression levels of bacterial and host genes after infection with two different strains of Ot. Dual RNAseq data was complemented with bacterial proteomics, which revealed insights into the regulation of gene expression as well as the basis of strain-specific differences in virulence in scrub typhus. The study serves as an example of using sequencing technologies to analyze the virulence mechanisms of neglected, genetically intractable bacterial pathogens.

Results

Distribution of RNA classes in Ot

We focused on two Ot clinical isolates: Karp, taken from a patient in New Guinea in 1943 [27], and UT176, a separate strain in the same serotype group as Karp, taken from a patient in northern Thailand in 2004 [28]. Human vein endothelial cells (HUVEC) were selected as host cells due to their similarity to cell types involved in both early and advanced infection. HUVEC cells were infected with bacteria at an MOI of 100:1 and grown for 5 days, by which point host cells were heavily loaded with bacteria (Fig. 1C, Supp. Figs 1, 2). Uninfected HUVEC cells were grown in parallel. After 5 days total RNA was isolated, depleted for rRNA, converted to cDNA and sequenced to ~35 million reads per library using Illumina technology. Reads were mapped to the completed genomes of Karp, UT176 [8] and, in parallel, the human genome. As the *Orientia* genome is repeat-rich, we additionally applied model-based quantification with Salmon [29] which uses uniquely mapping reads to assign multi-mapping reads, to these transcriptomes to improve our estimates of transcript abundance (see Methods).

We observed 2.8-4.9% bacterial reads in HUVECs infected with UT176 and 17.1-17.5% bacterial reads in HUVECs infected with Karp (Fig. 1D, Supp. Fig. 3). This difference likely reflects growth rate differences between Karp and UT176, which have doubling times of 19 and 27 hours in HUVEC, respectively (Fig. 1B). The distribution of RNA classes is shown in Fig. 1D. Depletion of ribosomal transcripts was efficient in the host transcriptome (<0.001% human rRNA reads). In contrast, we found an average of 21.03% and 52.06% rRNA reads in Karp and UT176, respectively. Most of the bacterial ribosomal reads were derived from 5S rRNA (Supp. Fig. 4), likely reflecting the divergence of 5S rRNA sequences between Ot and bacterial model

organisms used for optimization of the Ribo-Zero approach (https://emea.illumina.com/products/selection-tools/ribo-zero-kit-species-compatibility.html?langsel=/de/). Reads mapping to coding sequences (CDSs) were abundant in the HUVEC data subset (52% of all host-mapped reads), and also in the Ot-specific reads (41% of the Karp- and 33% of the UT176-mapped reads), allowing differential expression analysis. Dual RNA-seq also readily detected the various non-coding RNA classes from both host and bacteria (Fig. 1E).

Ot ncRNAs and evidence for tmRNA processing

Bacterial genomes encode many non-coding (nc)RNAs. Among the most conserved are several specialized, abundant housekeeping ncRNAs, including the RNA components of ribonuclease P (RNase P), the signal recognition particle (SRP), and transfer-messenger RNA (tmRNA), all of which were detected in the Karp transcriptome data (Fig. 1E, Supp. Table 1). To validate the RNA-seq data, we performed Northern blot analysis for the housekeeping ncRNAs (Fig. 2A). RNase P is a ribozyme. The catalytically active component is known as M1 RNA. Rfam annotations indicated Ot M1 RNA should be 385 nt in length and we detected a faint band at this position. However, a second band ran slightly higher, indicative of a length of ~450 nt (Fig. 2A), suggesting the existence of a precursor-M1 transcript.

SRP is an abundant ribonucleoprotein complex that shuttles *de-novo* synthesized membrane proteins to their subcellular destination. The RNA component of SRP, 4.5S RNA, is a transcript of ~100 nt. Our transcriptome data predicted Ot 4.5S RNA to be 98 nt in length, which matched the signal detected on the Northern blot (Fig. 2A).

We found evidence of tmRNA processing in Ot. tmRNA has both mRNA-like and tRNA-like features and rescues stalled ribosomes [30]. tmRNA generally consists of a tRNA-like (acceptor) domain encoded upstream of a short open reading frame (coding domain). However, most alpha-proteobacteria, two beta-proteobacteria, and some cyanobacteria, produce a two-piece tmRNA, with acceptor and coding domains split into two separate, base-pairing, RNA chains [31,32] (Fig. 2B). We detected three Ot tmRNA forms using Northern blot: (i) a long form, the precursor tmRNA (372 nt); (ii) a 5' fragment of ~80 nt, the acceptor domain; (iii) and the 3' coding domain of ~240 nt (Fig. 2B). Read coverage over the tmRNA locus in the Karp genome supported a cleavage event within the loop region that connects the tRNA- and mRNA-like domains in the full-length precursor (Fig. 2C).

In addition to these universally conserved housekeeping ncRNAs, bacterial genomes encode family-, genus-, species-, or strain-specific small ncRNAs (sRNAs) to adapt their gene expression to specific intrinsic and environmental cues [33,34]. Our RNA-seq data identified 57 intergenic sRNA candidates, between 77--803 nt, in the Karp transcriptome (Supp. Table 1). When normalized to the genome size of Ot, this is consistent with the number of sRNAs reported in bacterial model pathogens [35–40].

Conserved operons in a dynamic genome

The genome of Ot is highly dynamic [8] although the timescales and mechanisms of these genome rearrangements are unknown (Fig. 2D). Genes whose function is strictly dependent on co-expression should be under selective pressure to remain collinear in an otherwise dynamic

genome, and our previous analysis identified 657 core Ot genes (conserved in eight strains of Ot) located in 51 conserved operons, ranging in size from one to thirteen genes [8]. In that analysis singleton genes that were alone in all genomes were classified as operons with a size of 1. Here, we analyzed RNAseq data for evidence of co-expressed operons in Ot. Using Rockhopper [41] and manual curation, we identified adjacent genes expressed in a continuous transcript, classifying these as operons. We identified 132 operons fully conserved between Karp and UT176 (all genes expressed in both strains) and seven partly conserved (some genes expressed in both strains) (Fig. 2E). We compared these with the 51 previously identified operons based on conserved collinearity: 32 operons appeared in both the RNAseq and comparative genomics analyses. We also identified 212 transcribed operons present only in Karp, and 192 transcribed operons present only in UT176. The non-conserved operons were highly associated with the RAGE mobile element (73% in Karp and 93% in UT176) compared with conserved operons (14% of fully conserved operons). Operons conserved between Karp and UT176 likely encode essential genes whose collinearity is strongly selected for. In contrast, non-conserved operons are largely located in mobile genetic elements and may not perform useful functions within the cell.

The size of the conserved operons ranged from two to 30 genes. 84% consist of two or three genes (Supp. Fig. 6), consistent with the findings of conserved gene islands across multiple sequences and the dynamic nature of the Ot genome. The exception, an operon of 30 genes expressed in both Karp and UT176, encodes almost half of the 64 ribosomal genes in the Ot genome (Fig 2F).

Ot RAGE elements are regulated by antisense RNA

The RAGE element of Ot is an ICE that is present in at least 185 remnant copies [7]. It encodes an integrase gene (*int*), transposase *(tra)*, multiple genes from the VirB type IV secretion system (*vir*) and a number of potential effector genes including ankyrin-repeat containing proteins (*ank)*, tetratricopeptide repeat containing proteins (*TPR*), SpoT/RelA genes, DNA methyltransferases and replicative DNA helicases. Many of these genes are truncated and most RAGE element copies are highly degraded, containing only a subset of genes from the complete element. It is not known if this ICE is still active for transposition, nor whether Ot can express a functional Type IV secretion apparatus. In the RNAseq dataset ~50% of the most highly expressed genes were repetitive genes encoded by the RAGE element (defined throughout our analysis as integrase, transposase, conjugal transfer genes and hypothetical genes) in both strains. These same genes were also highly expressed in the antisense direction (Fig 3A, Supp. Fig. 7), leading us to hypothesize that the repetitive RAGE genes may be regulated by antisense gene expression.

Antisense transcription is widespread in bacteria [42], with between 5%--75% of coding sequences exhibiting antisense transcription. While functions for a number of specific antisense transcripts have been described, including regulation through occlusion of the ribosome binding-site or induction of RNase III mediated decay [43] their relevance as a general functional class remains unclear. Antisense promoters tend to be weakly conserved [44], arguing against specific functions, and mathematical modeling has suggested the majority of antisense transcripts are not expressed at sufficient levels to effect regulation of their cognate coding sequence [45].

In order to explore the relationship between sense and antisense expression of core Ot genes and repetitive RAGE genes we combined our RNAseq dataset with a proteomics dataset generated under the same experimental growth conditions. We observed substantially fewer RAGE genes detected by proteomics, compared with RNAseq (Fig. 3A). Genes with detected protein products

had higher transcript expression on average compared to those not detected by proteomics (Fig 3B). However, a many highly expressed transcripts appeared to produce no protein. Given our previous observations, we asked whether high levels of antisense transcription would correlate with low/no protein expression. Genes with detected proteins had an antisense-sense read count ratio of less than 1, in contrast to genes with no detected protein product, which had an antisense-sense read count spanning several orders of magnitude (Fig 3C) suggesting antisense RNA expression may be a factor in inhibiting translation.

To test this hypothesis more rigorously, we constructed three logistic regression models to predict protein detection from our transcriptomic data. The first used only transcripts per million (TPMs) derived from the sense strand as a predictor; the second used only the antisense-sense read count ratio as a predictor; the third used both features. Comparisons of the predictive power of these three models showed that antisense transcription is predictive of protein expression (Fig. 3D). Model 1, relying only on sense expression, did little better than chance at predicting protein detection. Models 2 and 3, which incorporate the antisense-sense ratio, led to large improvements in predictive power, suggesting that antisense transcription plays a wide-spread regulatory role in Ot. This was confirmed by 500-fold cross-validation showing that this was not an artifact of our training dataset (Methods, SI). We found significant enrichment for RAGE genes among those with high antisense-sense ratios (Fig. 3E), suggesting antisense transcription may work to control the expression of selfish genetic elements at the protein level. Twenty-six core genes also exhibited an antisense-sense ratio of >1 (Supp. Fig. 8). None of these models achieved greater than 67% balanced accuracy, suggesting the existence of other modes of post-transcriptional regulation.

Some of the genes whose translation is regulated by antisense gene expression encode putative secreted effector proteins, such as Ankyrin repeat containing proteins (Anks) and tetratricopeptide repeat containing proteins (TPR proteins). During sample preparation for proteomics analysis bacteria were first isolated from infected host cells in order to reduce the levels of host proteins and enrich the levels of bacterial proteins. Whilst cytoplasmic host cell proteins were not removed completely, as shown by the presence of numerous human proteins in our proteomics dataset, it is possible that secreted bacterial proteins such as Anks and TPR proteins were removed during sample preparation and not detected in the proteomics dataset. Removal of these genes from our analysis does not affect the outcomes of the prediction of antisense-sense ratio on protein expression.

Sets of genes exhibit different codon biases

A comparison of our RNAseq and proteomics datasets revealed a further 324 core genes detected by RNAseq but not proteomics. Some of this discrepancy is likely due to differences in sample preparation and sensitivities between the two methods, loss of secreted proteins, and differences in protein stability. However, some of this difference might also be explained by differences in codon usage resulting in differences translation efficiency. To address this question, we analyzed global codon usage in Ot and compared it with two other alpha-proteobacterial species, *Caulobacter crescentus* and *Rickettsia typhi* (Fig. 4A), and the gamma-proteobacterium *E. coli*. Genome-wide biases in Ot and *R. typhi* were very similar, but they differed from those in *C. crescentus* and *E. coli*, supporting previous analyses suggesting that codon bias in obligate intracellular bacteria is not strong, which may reflect slow growth rates not limited by the speed of translation [46–48]. We observed no clear correlation between codon usage and the presence or

absence of a cognate tRNA in Ot. Ot encodes 34 out of 61 possible amino-acid coding tRNAs, leaving 27 potential codons without a cognate tRNA molecule. Of the 20 amino acids, 7 have a strong bias against the cognate Ot tRNA, 10 have biases for both cognate and non-cognate Ot tRNAs and only 3 have biases for the cognate Ot tRNA (Fig. 4B-C, Supp. Fig. 10), unlike codon biases in other organisms, which correlate with tRNA copies [49–52]. There may be mechanisms by which bias towards non-cognate tRNA is preferential, or perhaps fast translation rates are not strongly selected for in Ot.

To determine whether Ot uses codon bias to regulate gene expression we compared codon usage between different groups of Ot genes (Fig. 4B, C). When comparing core and non-core gene groups, seventeen out of twenty amino acids showed differences in codon usage (Fig. 4B, Supp. Fig. 10), likely reflecting the horizontal acquisition of RAGE genes and possible differences in their subsequent optimization of translation.

We compared codon usage between core genes that were either detected or not detected in our proteomics dataset and found nine amino acids that were differentially encoded between these groups (Fig. 4B and C). The differences were statistically significant but smaller than those seen in other organisms, such as *E. coli*[53,54]. We compared these codon usage biases to those of a group of seven genes whose expression was high based on relative peptide levels in the proteomics dataset (RpoB1, RpoB2, GroEL, DnaK, GdhA, HtpG, TSA56) and observed the same trend as seen between those not detected and detected in the proteomics dataset. Thus, these differences likely reflect a true relationship between codon usage and translation, although the contribution of this bias to protein expression remains unknown.

Differential expression of genes in Karp and UT176

Of 657 predicted core Ot genes [8] 549 were expressed and 281 were highly expressed (see Methods for definitions). Ot grows in multiple mammalian cell types, and in vector cells, suggesting that different core genes may be required for the various stages of the Ot ecological cycle or at other stages of cellular infection.

Having identified transcriptional patterns in one strain of Ot we used a comparative approach to identify shared and strain-specific features. Karp and UT176 are closely related strains with a sequence identity of 95% in their TSA56 gene (used to classify strains). Consistent with a closed pan genome for Ot, the gene content of Karp and UT176 are similar, with differences primarily in gene copy number, pseudogenes, and gene order along the genome.

Using pathway and gene ontology (GO) analyses (Fig. 5A, Supp. Table 1) we found that most pathways were up-regulated in Karp compared with UT176, including those involved in DNA replication and metabolism, consistent with Karp's higher growth rate (Fig. 1B). We analyzed differential expression between Karp and UT176 at the gene level (Supp. Table 1, Figs 5A, B) and found several surface and effector proteins (Anks) were differentially regulated between the two strains. Ot encodes five autotransporter domain-containing proteins (ScaA-ScaE) and three immunogenic-type surface antigens (TSA22, TSA47, TSA56). All these surface proteins are immunogenic, based on their reactivity to patient sera [55], with TSA56 being the most abundant Ot surface protein. TSA56 has four variable domains and these lead to strain-specific antibody responses in patients. TSA47, TSA56, and ScaA have been evaluated as possible vaccine

candidates [56,57]. Of the core Ot genes, those most differentially expressed between Karp and UT176 were *scaE*, *tsa56*, and *tsa22* (1.39, 3.06 and 3.94 logFC in Karp over UT176, respectively). In contrast, *scaD* levels were increased in UT176 but to a lesser degree (0.80 logFC in Karp over UT176). Given their highly immunogenicity it is likely that their differential expression results in differential host responses.

Genes for ankyrin repeat-containing proteins (Ank) and tetratricopeptide repeat-containing proteins (TPR) are present in 33 (Ank)/29 (TPR) and 21 (Ank)/22 (TPR) copies in Karp and UT176, respectively [8]. Some *ank* genes function as effectors in eukaryotic cells whilst others are uncharacterized. We compared the expression of Ank and TPR genes in Karp and UT176. *ank2*, *ank3*, *ank12*, *tpr6*, and *tpr8* were up-regulated with a logFC >1.5 in UT176, and *ank1*, *ank16*, *ank17*, *ak20*, and *tpr1* were up-regulated with a logFC >1.5 in Karp. Most of these proteins were not detected in the Karp proteomics dataset suggesting that either the RNAs were not translated, or that the proteins were secreted and lost during purification. The protein products of all of these *ank* genes localise to the ER or cytoplasm [58]. Ank1 interferes with NFkB translocation to the nucleus and inhibit its transcriptional activation, similarly Ank6 [59]. The activity of the other differentially expressed Anks is not known. Given that these effector proteins interact directly with host cell proteins we expect that this differential expression will lead to downstream differences in host response.

Karp and UT176 induce a type-I interferon proinflammatory response

The transcriptional profile of HUVEC cells infected with Karp or UT176, compared with uninfected cells, showed up-regulation of specific genes in response to both strains (Fig. 6A) and

other genes in response to one or other of the strains (Fig. 6A blue and teal spots). The response to both strains was dominated by genes involved in the type-I interferon proinflammatory response (Supp. Table 1), shown previously in cultured endothelial cells and monocytes, and patient-derived macrophages response [17–21]. Genes up- or down-regulated in response to infection with Ot strain Karp in the context of canonical interferon signaling pathway is shown in Fig. 6B. The overall response in this pathway was similar in cells infected with UT176 (Fig. 6C).

Up-regulated genes include IFNB1 (interferon beta) and genes involved in regulating the type-I interferon response: IRF9 (interferon-regulated factor 9) and STAT1/2. Interferon-stimulated genes were also up-regulated upon Ot infection, including various interferon induced proteins with tetratricopeptide repeats (IFIT) genes and 2'-5'-oligoadenylate synthase 1 (OAS1). In addition to the type-I interferon pathway, the joint Ot response led to up-regulation of proinflammatory chemokines including CCL20, CXCL10, CXCL11, and cytokine receptors IL13RA2, IL7R, IL15RA, and IL3RA (Supp. Table 1).

The upstream signals leading to activation of these signaling pathways are unknown but Ot has been shown to activate host cells by signalling through the NOD1-IL32 [60] and TLR2 [61]pathways. Our data showed that TLR3 is up-regulated in cultured HUVEC cells in response to both Karp and UT176 (Supp. Table 1). TLR3 recognizes viral double-stranded (ds)RNA in the cytoplasm [62] and it is possible that it responds to Ot dsRNA. The up-regulation of transcription factor IRF7, which is known to respond to stimulation from membrane-bound TLRs, further supports a role for TLR2 and TLR3 in the detection of Ot.

Differential host responses to Karp and UT176

Although Karp and UT176 both induced a type-I interferon proinflammatory response compared to uninfected HUVEC cells (Fig. 6B,C), the magnitude of induction of certain pathways was distinct. The levels of multiple cytokines, chemokines, and cytokine receptors were higher in HUVEC cells infected with UT176 compared with Karp (Fig. 7A, Supp Fig 14, 15, 17). A network map of proinflammatory chemokines and cytokines, and their differential induction in response to UT176 and Karp is shown in Fig 7A. Most of the cytokines, chemokines, and cytokine receptors were differentially up-regulated by infection with UT176 compared with Karp, including CXCL8, CXCL1, CXCL2, CXCL10, IL6, IL1RL1, IL18R1. The levels of surface adhesion molecules associated with activation of the endothelium, VCAM1 and ICAM1, were also up-regulated in UT176 infected cells compared with Karp (Supp. Table 1). Whilst TLR3 was up-regulated in both strains, TLR3 activation in UT176 infected cells was 1.5 logFC higher than in response to Karp. Comparison of NFkB pathway genes and genes associated with NOS2 production revealed that genes in both pathways were up-regulated in UT176 but they were not up-regulated or were significantly less up-regulated in response to Karp infection (Supp. Fig. 16). Expression of host genes associated with leukocyte proliferation and mononuclear leukocyte differentiation was strongly induced in HUVECs infected with UT176 but were significantly less up-regulated or not up-regulated when infected with Karp (Supp. Fig. 17). Thus, UT176 seem to induce a stronger proinflammatory response and this may lead to more effective pathogen clearance.

In contrast to the multiple chemokines and cytokines up-regulated in UT176-infected HUVEC cells, only IL33 was specifically up-regulated in Karp-infected HUVEC cells (5 logFC

difference; Supp Table 1). IL33 is a proinflammatory cytokine that is involved in pathogenicity in a mouse model of scrub typhus [63]. To investigate Karp-mediated activation of IL33 we analyzed gene induction in the IL33-FAS network (Fig. 7B). Most genes in the network were differentially induced in Karp-infected HUVEC cells compared to in UT176-infected HUVECs. Up-regulation of IL33-NOS-mediated signaling contributes to tissue inflammation. We analyzed networks of genes involved in (i) organismal growth failure (ii) organismal morbidity and mortality and (iii) organismal death. In all cases Karp induced these networks whilst UT176 dampened them (Supp. Fig. 18).

Two Ot strains differ in virulence in a mouse model

To assay the effects of the differential inflammatory responses induced by Karp and UT176 we tested the relative virulence of the two strains in an intravenous mouse infection model. 1.25×10^6 bacteria were inoculated into outbred ICR mice and monitored for disease symptoms for 12 days prior to euthanasia. Blood and tissue from lung, liver, spleen, and kidneys were isolated and the bacterial load measured by qPCR. We found that Karp was significantly more virulent than UT176 as determined by disease outcome, weight loss, bacterial load, and histopathological analysis (Fig. 8, Supp. Fig. 19-21) This difference is likely to result from a combination of differential bacterial growth rate, differential expression of bacterial virulence related genes, and differential induction of host immune networks.

Discussion

Here, we have used dual RNAseq combined with proteomics and bioinformatics to analyse a neglected and genetically intractable obligate intracellular human pathogen, *Orientia tsutsugamushi*. We have revealed details of the transcriptome structure, gene expression mechanisms, and drivers of virulence in infected hosts.

Ot has a genome of 1.9-2.5 Mbp, almost half of which is composed of repetitive regions of >1,000 bp in length [8]. This is in contrast to the most closely related rickettsial species, whose genomes are typically around 1.1-1.3 Mbp [64]. Ot exhibits minimal synteny in gene order between strains, something not observed in other rickettsial genomes, implying dynamic or unstable behavior of Ot genomes over evolutionary timescales. RNAseq enabled identification of co-transcribed groups of genes. Comparision with a previous analysis of gene groups whose adjacent positions were conserved across 8 non-syntenic Ot genomes allowed us to identify genes whose position within an operon was strongly selected, which should help generate hypotheses about the functions of some of the 94 hypothetical proteins in the 657 core Ot genes.

Ot (Karp) encodes only 12 predicted transcription factors and 3 sigma factors, in contrast to 300 and 7, respectively, in *E. coli.* This limits the ability of Ot to regulate gene expression at the level of transcription initiation and suggests that other mechanisms are likely involved. The lack of synteny between strains indicates that positional co-regulation, seen in other bacteria [65] may be rare in Ot. However, the conservation of certain genes within operons indicates that some degree of local regulation is employed. The most prominent mechanism of regulation involved inhibition of translation of RAGE genes by antisense transcription. RAGE in Ot is a transposable

element of the integrative and conjugative element group. Transposable elements are regulated by antisense transcription, a mode of regulation preserved in Ot RAGE despite the fact that these elements contain mostly non-functional transposase and integrase genes and are not thought to be capable of mediating transposition [66]. It is unclear whether these untranslated RAGE transcripts play an as yet uncharacterized role in Ot, or whether they are purely selfish DNA elements that Ot has been unable to dispose of due to the small population size and evolutionary bottleneck of the obligate intracellular lifestyle, in which case such wasteful transcription would impose a high metabolic burden on Ot cells.

In addition to regulation by antisense transcription, we identified 57 (Karp) and 84 (UT176) sRNA transcripts in intergenic regions (IGRs). sRNAs from IGRs often regulate mRNA stability or translation of target genes in diverse bacterial organisms, and in Gram-negative species these sRNA-mRNA interactions are often mediated by the RNA chaperone Hfq. Ot does not encode an identifiable Hfq homolog, however it is possible that Ot sRNAs may regulate gene expression in an Hfq-independent manner. Due to the lack of genetic tools for Ot we were unable to test the functions of the sRNAs through deletion or over-expression. However a number of sRNAs have been identified in *Rickettsia prowazekiii* and *Rickettsia conorii* and their binding to candidate target genes demonstrated [67–69].

Degeneracy in the genetic code means that many amino acids can be encoded by more than one codon, and biases in codon usage are seen at the genome, and individual gene levels. At the genome level, different organisms exhibit different relative preferences for certain codons, typically reflecting the abundance of cognate tRNAs. At the gene level, common codons are

often present in highly expressed genes, with rare codons occasionally used to slow down translation and enhance protein folding. At the genome level in Ot, we observed that codon preferences do not reflect the presence of cognate tRNAs, indicating a high dependence on wobble-pairing tRNA binding. Similar codon preferences are seen in *R. typhi* but not in the related free-living alpha-proteobacteria, *Caulobacter crescentus*, indicating a relationship between codon bias and bacterial lifestyle. At the gene level we found a slight but significant difference in codon usage between Ot core and non-core genes. Since most non-core genes are encoded by RAGE this likely reflects a difference in origin of these horizontally acquired mobile genetic elements. When we compared codon usage between genes not detected by proteomics, detected by proteomics, or very highly expressed based on number of peptide hits, we observed differential codon biases between these groups, reflecting a relationship between codon usage and protein level, and implying the existance of another level of gene regulation in Ot. Unless relative tRNA abundance is directly regulated by Ot, it is difficult to imagine how this level of regulation would be responsive to different cell types and between mite vectors and mammalian infection.

Our data provide evidence that gene expression in Ot is regulated by antisense transcription and codon bias, and that Ot expresses a number of IGR sRNAs that may also play an regulatory roles. One limitation of the current study is that our data was collected at a single timepoint and in a single cell type. Therefore, differential gene expression in response to environmental changes could not be assayed.

The inflammatory response triggered by Ot infection is a key driver of pathogenesis in scrub typhus. An insufficient response results in pathogen proliferation and ensuing host damage, whilst an over-active immune response leads to excessive inflammation and ensuing host damage. Pathogen clearance and patient health is dependent on the balance of pathogen and host factors. We compared the response of endothelial cells to two strains of Ot and found that differential activation of the immune response correlated with differential outcomes in a scrub typhus mouse model. Whilst both Karp and UT176 induced an antiviral proinflammatory response, as shown previously, UT176 strongly induced an IL6-mediated pro-inflammatory response whilst Karp induced an IL33-NOS3-FAS response, differences likely to influence the relative pathogenicity of these strains.

IL33 was one of the most strongly differentially regulated genes in UT176 and Karp infections (5.1 logFC higher in Karp-infected HUVECs). IL33 has previously been shown to play a role in pathogenesis in a scrub typhus murine model, using the Karp strain, where it was shown that IL33 levels were increased during Ot infection, that IL33-/- mice showed less severe disease symptoms, and that addition of rIL33 increased severity and mortality [63]. Our observations of reduced induction of IL33 by the less virulent UT176 strain further supports a role for this molecule in the pathogenesis of scrub typhus.

In the absence of genetic tools it is difficult to identify specific genes that drive virulence differences between UT176 and Karp. However, comparative genomics has revealed that whilst the pan genome of Ot is open, it is largely composed of gene duplications rather than newly acquired genes. This lack of gene novelty likely reflects the environmental isolation associated

with an obligate intracellular lifestyle. Consequently, strain-specific differences in virulence are likely to be driven largely by differences in relative gene expression rather than the presence or absence of virulence genes. Consistent with this, we observed an up-regulation of virulence-associated surface proteins in Karp compared with UT176. Other genes differentially expressed between these strains may also play an unknown role.

In summary, we have used dual RNA-seq to gain insights into the transcriptome structure and mechanisms of gene regulation in Ot during infection. We provide evidence for widespread antisense regulation to regulate expression of RAGE genes. We identified a relationship between the relative induction of IL33- and IL6-based gene networks in the host and disease pathogenesis. These findings will lay the groundwork for subsequent studies on the regulation of gene expression in Ot and mechanisms of pathogenesis.

Figure legends

Figure 1. Overview. A. Schematic experimental overview. HUVEC = human umbilical vein endothelial cell. **B.** Growth curve showing replication of Ot in cultured HUVEC cells. Bacteria were grown in 24 well plates and the total bacteria per well is shown. Mean and SD from three independent replicates are shown. **C.** Confocal microscopy images of Ot in HUVEC cells 5 days post infection. Additional images and time points are shown in Supp. Fig. 1 and 2. Blue = DAPI (DNA), Red = Evans blue (host cells), green = Ot labelled with FITC-click-methionine. **D.** RNA mapping statistics showing the fraction of host and Ot RNA for each condition. The mean of three independent experiments is shown. Individual results for each replicate are shown in Supp. Fig. 3. **E.** Percentage of RNA-seq reads assigned to different classes of RNA in Karp, UT176 and HUVEC.

Figure 2. RNA biology in Ot. A. Northern blot analysis of core non-coding RNAs in Ot. **B.** Structure of the two-piece tmRNA observed in the Ot transcriptome **C.** RNA-seq read covereage over the tmRNA gene mirrors cleavage observed by Northern blot. **D.** A comparison of genomic synteny of two species within the enterobacteriaceae (*Escherichia coli* MG1655 and *Salmonella enterica* serovar Typhimurium SL1344, top), with synteny between the two Orientia strains from this study (bottom). **E.** Pie charts illustrating the relative abundance of RAGE genes in conserved (right) and strain-specific (left) operons. **F.** Visualization of the largest conserved operon in Ot, encoding multiple ribosomal genes, showing RNA-seq coverage in both strains.

Figure 3. Antisense transcription is enriched on mobile genetic elements and is predictive of an absence of protein expression. A. B. Plot showing the relationship between protein

expression, defined by LFQs, and transcript expression, defined by TPMs. Genes cluster into two groups based on their protein expression. The red line indicates the threshold for expressed genes (TPM value equal to 10). **C.** Sense transcription and the ratio of reads assigned to the antisense and sense strands, showing classification based on proteomics detection. The red line indicates the sense-antisense ratio (1.059182) above which translation was not detected by mass spectrometry. **D.** ROC curves evaluating the performance of logistic regression models to predict protein expression from RNA-seq read counts. Model 1 strictly uses sense expression, Model 2 the antisense-sense ration, and Model 3 uses both. Incorporating antisense expression clearly improves model performance. **E.** Fraction of core genes and RAGE genes in the set of genes with high antisense-sense ratios, compared to all expressed genes.

Figure 4. Codon bias in Ot. A. Genome-wide codon biases in *O. tsutsugamushi, R. typhi, C. crescentus* and *E. coli*. Genome-wide RSCU values are indicated by the green boxes, while the red boxes are the values for 50 randomly chosen individual genes. tRNA gene copy numbers are given in parentheses. B. Relative use of different codons by subsets of genes in Ot (Karp) that are absent in proteomics, detected in proteomics, highly expressed in proteomics, encoded by non-core gene set or encoded by core gene set. Three illustrative examples are shown (Lys, Ile, Ser) with the full set of 20 amino sets given in Supp. Fig. 12. The error bars represent 99.7% confidence intervals and were generated by computing the normalized codon frequencies from 10000 sets of randomly selected genes, with each set being equal in size to the corresponding gene set. C. Overview of relative use of codons to encode each amino acid in the different proteomics gene sets. The bias between groups (yes in green, no in red) indicates whether the different groups use the different possible codons differently. This is defined by a

consistent trend between the not detected, detected and highly expressed groups. As an example, Lys in Fig. 4B=no, whilst Ile and Ser=yes. The direction of genome-wide bias indicates whether the genome-wide relative use of the different codon correlates with the presence or absence of cognate tRNA, or not. (positive correlation in green, negative correlation in red, split correlation in yellow). As an example Lys=positive, Ile=negative, Ser=split.

Figure 5. Differential bacterial expression. A. Heatmap illustrating pathways enriched in differentially expressed genes. All illustrated categories are more highly expressed in Karp. B. Volcano plot showing differential expression of bacterial genes in Karp and UT176. Bacterial surface genes (red) and ankyrin repeat containing effector proteins (blue) with log fold change ≥1 are highlighted.

Figure 6. Ot induces an antiviral interferon response in HUVEC cells. A. Summary of the host response showing joint and strain specific responses. **B.** Activation of multiple genes in the canonical interferon signaling pathway in Karp infected HUVEC cells compared with uninfected HUVEC cells. **C.** Heat map showing up-regulation of genes in the interferon signaling pathway in HUVEC cells infected with Karp and UT176 compared with uninfected cells.

Figure 7. Karp and UT176 lead to up-regulation of distinct networks in HUVEC cells. A. Up-regulation of multiple proinflammatory chemokines and cytokines in HUVEC cells infected with UT176. **B.** Induction of the IL33-FAS-mediated anoikis network in Karp infected HUVEC cells.

Figure 8. Karp is more virulent than UT176 in a mouse infection model. A. Weight change over 12 days infection. **B.** Clinical observation score of mice 12 days post infection. This number is a composite score based on appetite, activity, and hair coat with higher numbers representing low appetite, low activity, and ruffled fur. Details provided in Supp. Fig. 20. **C.** Bacterial copy number in 100 µl blood taken from euthanized mice 12 days post infection, measured by qPCR. **D.** Ratio of bacterial DNA copy number to mouse DNA copy number in lung, liver, spleen, and kidney of euthanized mice 12 days post infection, measured by qPCR. **E.** Lesion scores of Hematoxylin and Eosin stained lung, liver, spleen, and kidneys of euthanized mice 12 days post infection. Scores range from 0 to 5 with 0 representing normal tissue and 5 representing severe lesion damage. All graphs show mean and standard deviation. Statistical significance is calculated using unpaired t test using GraphPad Prism software. **$p \leq 0.01$ ***$p \leq 0.001$ ****$p \leq 0.0001$ F. Images of Hematoxylin and Eosin stained lung tissue of mice infected with buffer, UT176 or Karp. Scale bars = 50 µm. * indicates airway and ** indicates blood vessel. Uninfected control: airway, blood vessel and alveoli all appear normal. UT176-infected lungs: There is diffuse thickening and infiltration of alveolar septa with a mixed population of macrophages and lymphocytes (arrows). There is also mild perivascular lymphohistiocytic inflammation (open arrow). Karp-infected lungs: There is diffuse moderate thickening and infiltration of alveolar septa with a mixed population of macrophages and lymphocytes. The airway (*) is unaffected and normal. Additional figures are shown in Supp. Fig. 20.

Materials and Methods

Growth of Ot and isolation of RNA

The clinical isolate strains (Karp and UT176) of *Orientia tsutsugamushi* were propagated in a confluent monolayer of host cells (HUVEC, Human Umbilical Vein Endothelial Cells) for 5 days at MOI 100:1. Cells were cultured using Media200 (Thermo Fisher, Catalog no. M200-500) supplemented with LVES media (Thermo Fisher, Catalog no. A14608-01) at 35 °C and 5% CO_2. The infectivity was determined by qPCR of the single copy Ot gene 47kDa at day 5-7. Both uninfected cells and infected cells were harvested by incubating the cells on ice and quickly resuspending in RNAprotect Bacteria Reagent (Qiagen, catalog no. 76506), then storing at -80 °C until use. RNA extraction was performed using the Qiagen RNeasy Plus kit (Qiagen, catalog number 74136) according to manufacturer's instructions and as described previously (Atwal, 2017).

RNA processing and sequencing

The integrity of the DNase-treated RNA samples was assessed in a Bioanalyzer (Agilent). All samples had RIN (RNA integrity number) values ≥8.0. Ribosomal transcripts were removed using the Ribo-Zero Gold (epidemiology) kit (Illumina). Following the manufacturer's instructions, 500 ng of total, DNase-treated RNA was used as an input to the ribo-depletion procedure. rRNA-depleted RNA was precipitated in ethanol for 3 h at -20°C.

cDNA libraries for Illumina sequencing were generated by Vertis Biotechnologie AG, Freising-Weihenstephan, Germany. rRNA-free RNA samples were first sheared via ultrasound sonication (four 30-s pulses at 4°C) to generate on average 200- to 400-nt fragments. Fragments of 20 nt

were removed using the Agencourt RNAClean XP kit (Beckman Coulter Genomics) and the Illumina TruSeq adapter was ligated to the 3' ends of the remaining fragments. First-strand cDNA synthesis was performed using M-MLV reverse transcriptase (NEB) wherein the 3' adapter served as a primer. The first-strand cDNA was purified, and the 5' Illumina TruSeq sequencing adapter was ligated to the 3' end of the antisense cDNA. The resulting cDNA was PCR-amplified to about 10 to 20 ng/µl using a high fidelity DNA polymerase. The TruSeq barcode sequences were part of the 5' and 3' TruSeq sequencing adapters. The cDNA library was purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics) and analyzed by capillary electrophoresis (Shimadzu MultiNA microchip).

For sequencing, cDNA libraries were pooled in approximately equimolar amounts. The cDNA pool was size fractionated in the size range of 200 to 600 bp using a differential cleanup with the Agencourt AMPure kit (Beckman Coulter Genomics). Aliquots of the cDNA pools were analyzed by capillary electrophoresis (Shimadzu MultiNA microchip). Sequencing was performed on a NextSeq 500 platform (Illumina) at Vertis Biotechnologie AG, Freising-Weihenstephan, Germany (single-end mode; 75 cycles).

Northern blots

Each 15 µg of total RNA (i.e. a mixture of human and Ot RNA) prepared as above were loaded per lane and separated in 6% (vol/vol) polyacrylamide–7 M urea gels. Blotting was performed as previously described [24]. After the transfer onto Hybond XL membranes (Amersham), RNA was cross-linked with UV light and hybridized at 42°C with gene-specific 32P-end-labeled DNA oligonucleotides (Supp. Fig. 13) in Hybri-Quick buffer (Carl Roth AG). After exposure, the screens were read out on a Typhoon FLA 7000 phosphorimager (GE Healthcare).

qRT-PCR

qRT-PCR was performed with the Power SYBR Green RNA-to-CT1-Step kit (Applied Biosystems) according to the manufacturer's instructions and a CFX96 Touch real-time PCR detection system (Bio-Rad). Human U6 snRNA served as reference transcripts. Fold changes in expression were determined using the 2(-ΔΔCt) method [70]. Primer sequences are given in Supp. Fig. 13, and their specificity had been confirmed using Primer-BLAST (NCBI).

RNA seq read processing and quantification

The raw reads were initially processed according to our established dual RNA-seq pipeline[24]. Briefly, raw reads were trimmed for adaptor sequences and a minimum read quality of 20 using cutadapt[71]. Reads were then mapped against the human (GRCh38) and Ot (UT176 accession: LS398547.1; Karp accession: LS398548.1) reference sequences using the READemption pipeline (v0.4.3, [72]) and segemehl with the lack remapper (v0.2.0 [73]), removing reads that mapped equally well to the bacterial and host genomes. For downstream analysis of human gene expression, only uniquely mapping reads were retained for quantification.

To improve quantification of repetitive sequences, reads mapped to the Ot genomes were used for quantification of bacterial transcript expression using Salmon (v0.9.1) [29]. Salmon is a quasi-mapping based gene expression quantification tool that consists of two steps, indexing and quantification.

Transcript fasta files were created from the Genbank annotations using the gene coordinates. The indexing step was performed in quasi-mapping mode (--type quasi). Expression of the transcripts was quantified using both stranded forward library type (-lSF) and removing incompatible

mappings (--incompatPrior 0.0). Salmon identified identical gene repeats that are collected in 218 groups (see supplementary table, Karp groups of duplicates). For quantification purposes we retained a single gene from each group.

For the purposes of summarizing gene expression, we calculated mean TPM values from three replicates for each strain. Genes with a mean TPM greater than 10 were classified as expressed, and those with a mean TPM value greater than 50 highly expressed.

Gene annotation

For each gene we retrieved the gene name, gene product, and amino acid sequence from the Genbank annotation. In addition, using eggNOG-mapper[10] we predicted gene names and both KEGG pathways [11] and GO terms. We manually identified surface antigen encoding proteins using BLAST. The KEGGREST (Tenenbaum, D (2019) KEGGREST: Client-side REST access to KEGG. R package version 1.18.1.) and GO.db (Carlson M (2019). GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.5.0.). R packages were used to retrieve KEGG and GO terms, respectively.

Non-coding RNA prediction

Noncoding RNAs were annotated using Rockhopper[76], ANNOgesic [77] (v0.7.17) and Infernal[78] (v1.1.2) searching sequences against the Rfam database[79]. These provided inconsistent predictions of intergenic sRNAs. Intergenic sRNAs were manually curated by visual comparison of the predicted sRNA coordinates with the read coverage in the Integrative Genomics Viewer [80]

(v2.5.2). Infernal predicted the core housekeeping ncRNAs tmRNA, RNaseP, SRP and 5s rRNA. The quantification of the bacterial transcriptomes complemented with predicted sRNAs was performed using Salmon.

Genomic alignment

Genomic comparisons in Figure panels D and F were performed using Easyfig[81]. Escherichia coli K-12 MG1655 (Accession number U00096) and Salmonella enterica serovar Typhimurium SL1344 (Accession number FQ312003) were used as comparators for synteny analysis.

Orthology and conserved operon prediction

We predicted orthologous genes between the two Orientia strains using Poff[82] with default parameters in synteny mode. To identify conserved operons we used operon structures predicted in each strain by Rockhopper [41]. Based on visual analysis of read coverage in the Integrative Genomics Viewer, some of the operons were manually extended by addition of genes or merging two operons into one. We also identified partially conserved operons missing some genes in one strain.

Differential gene expression

For the bacteria, differential gene expression analysis was performed between orthologous genes identified by Poff. Genes that were predicted as an orthologous group (more than two genes) were removed from the analysis. Additionally, we removed duplicates (transcripts with perfectly identical sequence) that were identified by Salmon in either strain. For both human and bacterial

RNA-seq data, we performed differential gene expression analysis with the edgeR package [83] (v3.20.9) using robust quasi-likelihood estimation [84]. To identify biological processes that differ between two Orientia strains, we have performed gene set analysis using KEGG and GO terms that contain at least 4 expressed genes using the fry test in the edgeR package.

Proteomic sample preparation

Bacteria were propagated in HUVEC cell line at MOI 100:1 and harvested at 5 dpi. Ot was isolated, washed with 0.3 M sucrose, and lysed with 1% triton-X prior to acetone precipitation of protein. Total protein was then alkylated, reduced, and subsequently treated with Lys-C/Trypsin. Digested peptides were desalted using Oasis® HLB reversed-phase cartridges, vacuum dried, and stored for MS runs.

Mass spectrometry

The dried samples were resuspended with 10 µl of 2% (v/v) acetonitrile solution containing 0.06% (v/v) trifluoroacetic acid and 0.5% (v/v) acetic acid and loaded onto an autosampler plate. Online chromatography was performed using EASY-nLC 1000 (Thermo Scientific) in single-column setup using 0.1% formic acid in water and 0.1% formic acid in acetonitrile as mobile phases. Pre-run column equilibration was carried out by flowing 5 µl of 0.1% formic acid in water through the reversed-phase C18 column (EASY-Spray LC Column, 75 µm inner diameter x 50 cm, 2 µm particle size) (Thermo Scientific). The samples were injected and separated on the analytical column maintained at 50 °C using a 2–23% (v/v) acetonitrile gradient over 60 min, then ramped to 50% over the next 20 min, and finally to 90% within 5 min. The final mixture was maintained for 5 min to elute all remaining peptides. Total run duration for each sample was 90 min at a constant flow rate of 300 nl/min.

Data was acquired using an Orbitrap Fusion mass spectrometer (Thermo Scientific) in data-dependent mode. Samples were nebulized using 2.5 kV and 300 °C at the nanospray source and positively-charged precursor MS1 signals were detected using an Orbitrap analyzer set to 60,000 resolution, automatic gain control (AGC) target of 400,000 ions, and maximum injection time (IT) of 50 ms. Precursors with charges 2–7 and having the highest ion counts in each MS1 scan were further fragmented using collision-induced dissociation (CID) at 35% normalized collision energy and their MS2 signals were analyzed by ion trap at an AGC of 10,000 and maximum IT of 35 ms. Precursors used for MS2 scans were excluded for 90 s in order to avoid re-sampling of high abundance peptides. The MS1-MS2 cycles were repeated every 3 s until completion of the run.

Identification of proteins within each sample was performed using MaxQuant (v1.5.5.1). Raw mass spectra were searched against Orientia tsutsugamushi primary protein sequences derived from complete genome data. Human whole proteome sequences were obtained from Uniprot and included as background. Carbamidomethylation on Cys was set as the fixed modification and acetylation on protein N-terminus and oxidation of Met were set as dynamic modifications for the search. Trypsin was set as the digestion enzyme and was allowed up to 3 missed cleavage sites. Precursors and fragments were accepted if they had a mass error within 20 ppm. Peptides were matched to spectra at a false discovery rate (FDR) of 1% against the decoy database.

Proteomic data analysis

Protein expression was measured by label free quantification values (LFQs). A protein was classified as detected if at least two peptides were detected in at least 2 biological replicates, and the mean LFQ across the three replicates was used for further analysis. Otherwise the protein

was classified as undetected, and the LFQ value was set to zero. The proteomic data includes 19 protein groups that couldn't be resolved, consisting of 93 proteins. In our analysis we discarded these proteins to simplify the analysis.

Transcript classification

Sense transcript expression was defined by mean TPM value across replicates. The antisense/sense ratio was calculated as the ratio of mean read counts assigned to the antisense and sense strand of coding annotations. The duplicates predicted by Salmon (Supp. Table 1. Karp groups of duplicates) and non-coding RNAs were removed from the analysis.

We divided the data set into two classes, detected and undetected in proteomics. Within our analyzed data set, 322 genes were detected, whereas 1608 genes were not detected by mass spectrometry. We found a weak positive correlation between TPMs and LFQs for genes with detected proteins (Spearman’s correlation coefficient equal to 0.3235764), but it was not a linear association (Pearson's correlation coefficient equal to 0.03722184). For the further analysis we selected transcripts with sense expression greater than 10 TPMs, previously defined as our expression threshold.

Logistic Regression model

To test whether antisense-sense ratios are predictive of protein expression, we have applied logistic regression, which models the probability of a binary response, that is, whether a protein is expressed or not. We have built 3 competing models. Model 1 makes predictions of the protein expression based solely on sense transcription:

$\beta_0 + \beta_1$* (TPM sense)

Model 2 makes predictions solely on the antisense-sense ratio:

$\beta_0 + \beta_1$* (number of antisense reads) / (number of sense reads)

Model 3 uses both sense transcription and the antisense-sense ratio to make predictions:

$\beta_0 + \beta_1$* (TPM sense) + β_2* (number of antisense reads) / (number of sense reads)

Since the data is highly imbalanced, 295 transcripts with detected proteins, and 814 without, we have used a downsampling procedure (downSample function) implemented in the caret R package[85] to create a balanced data set for model training purposes. Next, the function glm() with a logit link function from the caret package was used to fit models to the reduced data set. For a first indication as to whether any of these models are predictive, we trained all three models on a downsampled data set consisting of 590 genes, then tested them on the complete data set. To more rigorously assess this result, we have applied 500-fold cross validation. For each fold the data was split randomly into 2 data sets, training and testing which included 1055 and 54 genes, respectively. Each time the new training data set was reduced to 562 genes, which were used to estimate the model parameters, and then the model was evaluated on the testing data set. The model performance was evaluated using a variety of measures, i.e. sensitivity, specificity,

accuracy (caret R package) as well as with ROC curves [86] (pROC 1.14.0) and the area under the ROC curve (AUC).

Immunofluorescence microscopy

The protocol for L-Homopropargylglycine (HPG) incorporation, click chemistry and fluorescence detection were based on recommendations from Click-iT® HPG Alexa Fluor® Protein Synthesis Assay Kits (Molecular probe by Life Technologies). HUVEC cells were grown on chambered coverslip slides (Ibidi, USA), for 2 days before infection with bacteria at MOI 100:1. To incorporate HPG at times indicated, medium was removed and replaced with L-methionine-free medium (Dulbecco's Modified Eagle Medium, DMEM, Cat. no. 21013) containing 25 µM HPG for 30 min at 37 ºC. Labeled bacteria were washed twice in 1X PBS+ 1mg/ml BSA, pH 7.4 before fixing with 4% formaldehyde and subsequently being permeabilized with 0.5% TritonX for 20 min on ice. After washing with PBS + 1 mg/ml BSA, the Click-iT® reaction cocktail (Click-iT® HPG Alexa Fluor® Protein Synthesis Assay Kits cat. C10428) was incubated with cells for 30 min at room temperature in the dark. The Azide dye (Alexa Fluor®488) was used at a final concentration of 5 µM. After the click reaction, cells were labeled with the actin probe Alexa Fluor® 594 phalloidin at a dilution of 1:40 and the nuclear stain Hoechst diluted to 1:1000 for 30 min at 37 °C. Cells were washed 3X with PBS which was replaced with mounting media after the final wash. Imaging was performed using a Zeiss LSM 7000 equipped with a 63 × 1.4 NA objective lens (Carl Zeiss, USA) and also a Leica SP8 laser scanning confocal microscope.

Analysis of codon bias

We calculated the RSCU (relative synonymous codon usage) for each codon to quantify genome-wide or gene-specific codon usage bias following [52]. To determine the genomic codon counts for each species and gene set, we parsed nucleotide sequence data and annotation in the GenBank file format, downloaded from the NCBI database. We also obtained tRNA gene copy numbers from the GtRNAdb database [87,88], and integrated protein abundance for E. coli K-12 MG1655 data from PaxDB [89].

Mice and Ethic statement

All animal research was performed strictly under approved Institutional Animal Care and Use Committee (IACUC) protocol by the IACUC and Biosafety Review Committee at the Armed Forces Research Institute of Medical Sciences (AFRIMS) Bangkok, Thailand, an AAALAC International-accredited facility. The protocol number was PN16-05. The animal research was conducted in compliance with Thai laws, the Animal Welfare Act, and all applicable U.S. Department of Agriculture, Office of Laboratory Animal Welfare and U.S. Department of Defense guidelines. All animal research adhered to the Guide for the Care and Use of Laboratory Animals, NRC Publication (8th Edition).

C57BL/6NJcl mice were purchased from Nomura Siam International, Bangkok, Thailand. Mice were housed in an animal biosafety level 2 facility and moved to an animal biosafety level 3 containment 2 days before the inoculation. Mice at 6- 8 weeks of age were used in these experiments. Two group of mice (n =8 per group) were intravenously injected in the tail vein with 1.25×10^6 genome copies of *O. tsutsugamushi* of either Karp strain or UT176 strain. The *O. tsutsugamushi* inoculum was derived from *O. tsutsugamushi*-infected L929 cells. Clinical signs and body weight were evaluated daily. After 12 days post inoculation, all mice were euthanized.

Blood and tissue samples including lungs, liver, spleen, and kidneys were collected for bacteria quantification and histopathology.

Acknowledgements

JS is funded by a Royal Society Dorothy Hodgkin Research Fellowship (DH140154) and this project was additionally funded by a grant from the University of Oxford Medical Sciences Division Medical Research Fund. The authors are grateful to Guy Riddihough from Life Science Editors for editorial support on this manuscript.

Author contributions

Conceptualization, J.V., L.B., J.S.; Investigation, B.M.G., S.G., A.J.W., J.W., S. C., L.C.W., P.S., L.B.; Formal Analysis, B.M.G., W.K.C., S.S., R.S., L.B.; Resources, R.S., J.V., L.B., J.S.; Writing – Original Draft, L.B., J.S.; Writing – Review and Editing, A.J.W., L.B., J.S.; Supervision, L.B., J.S.; Funding Acquisition, J.V., J.S.

Competing interests

The authors declare no competing interests.

References

1. Luce-Fedrow, A., Lehman, M., Kelly, D., Mullins, K., Maina, A., Stewart, R., Ge, H., John, H., Jiang, J. & Richards, A. A Review of Scrub Typhus (Orientia tsutsugamushi and Related Organisms): Then, Now, and Tomorrow. *Tropical Medicine and Infectious Disease* **3**, 8 (2018).
2. Izzard, L., Fuller, A., Blacksell, S. D., Paris, D. H., Richards, A. L., Aukkanit, N., Nguyen, C., Jiang, J., Fenwick, S., Day, N. P., Graves, S. & Stenos, J. Isolation of a novel Orientia species (O. chuto sp. nov.) from a patient infected in Dubai. *J Clin Microbiol* **48**, 4404-4409 (2010).
3. Weitzel, T., Dittrich, S., López, J., Phuklia, W., Martinez-Valdebenito, C., Velásquez, K., Blacksell, S. D., Paris, D. H. & Abarca, K. Endemic Scrub Typhus in South America. *N Engl J Med* **375**, 954-961 (2016).
4. Paris, D. H., Phetsouvanh, R., Tanganuchitcharnchai, A., Jones, M., Jenjaroen, K., Vongsouvath, M., Ferguson, D. P., Blacksell, S. D., Newton, P. N., Day, N. P. & Turner, G. D. Orientia tsutsugamushi in human scrub typhus eschars shows tropism for dendritic cells and monocytes rather than endothelium. *PLoS Negl Trop Dis* **6**, e1466 (2012).
5. Moron, C. G., Popov, V. L., Feng, H. M., Wear, D. & Walker, D. H. Identification of the target cells of Orientia tsutsugamushi in human cases of scrub typhus. *Mod Pathol* **14**, 752-759 (2001).
6. Cho, N. H., Kim, H. R., Lee, J. H. & Kim, I. S. The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host– cell interaction genes. *PNAS* **104**, 7981-7986 (2007).
7. Nakayama, K., Yamashita, A., Kurokawa, K., Morimoto, T., Ogawa, M., Fukuhara, M., Urakami, H., Ohnishi, M., Uchiyama, I., Ogura, Y., Ooka, T., Oshima, K., Tamura, A., Hattori, M. & Hayashi, T. The Whole-genome sequencing of the obligate intracellular bacterium Orientia tsutsugamushi revealed massive gene amplification during reductive genome evolution. *DNA Res* **15**, 185-199 (2008).
8. 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).
9. Kelly, D. J., Fuerst, P. A., Ching, W. M. & Richards, A. L. Scrub typhus: the geographic distribution of phenotypic and genotypic variants of Orientia tsutsugamushi. *Clin Infect Dis* **48 Suppl 3**, S203-30 (2009).
10. Premaratna, R., Blanton, L. S., Samaraweera, D. N., de Silva, G. N., Chandrasena, N. T., Walker, D. H. & de Silva, H. J. Genotypic characterization of Orientia tsutsugamushi from patients in two geographical locations in Sri Lanka. *BMC Infect Dis* **17**, 67 (2017).
11. Varghese, G. M., Janardhanan, J., Mahajan, S. K., Tariang, D., Trowbridge, P., Prakash, J. A., David, T., Sathendra, S. & Abraham, O. C. Molecular Epidemiology and Genetic Diversity of Orientia tsutsugamushi from Patients with Scrub Typhus in 3 Regions of India. *Emerg Infect Dis* **21**, 64-69 (2015).
12. James, S. L., Blacksell, S. D., Nawtaisong, P., Tanganuchitcharnchai, A., Smith, D. J., Day, N. P. & Paris, D. H. Antigenic Relationships among Human Pathogenic Orientia tsutsugamushi Isolates from Thailand. *PLoS Negl Trop Dis* **10**, e0004723 (2016).

13. Fukuhara, M., Fukazawa, M., Tamura, A., Nakamura, T. & Urakami, H. Survival of two Orientia tsutsugamushi bacterial strains that infect mouse macrophages with varying degrees of virulence. *Microb Pathog* **39**, 177-187 (2005).
14. Hanson, B. Comparative susceptibility to mouse interferons of Rickettsia tsutsugamushi strains with different virulence in mice and of Rickettsia rickettsii. *Infect Immun* **59**, 4134-4141 (1991).
15. Sunyakumthorn, P., Paris, D. H., Chan, T. C., Jones, M., Luce-Fedrow, A., Chattopadhyay, S., Jiang, J., Anantatat, T., Turner, G. D., Day, N. P. & Richards, A. L. An intradermal inoculation model of scrub typhus in Swiss CD-1 mice demonstrates more rapid dissemination of virulent strains of Orientia tsutsugamushi. *PLoS One* **8**, e54570 (2013).
16. Groves, M. G. & Kelly, D. J. Characterization of factors determining Rickettsia tsutsugamushi pathogenicity for mice. *Infect Immun* **57**, 1476-1482 (1989).
17. Cho, N., others & Kim, I. S. Expression of Chemokine Genes in Murine Macrophages Infected with Orientia tsutsugamushi. *infection and Immunity* (2000).
18. Cho, N. H., Seong, S. Y., Choi, M. S. & Kim, I. S. Expression of chemokine genes in human dermal microvascular endothelial cell lines infected with Orientia tsutsugamushi. *Infect Immun* **69**, 1265-1272 (2001).
19. Koh, Y. S., Yun, J. H., Seong, S. Y., Choi, M. S. & Kim, I. S. Chemokine and cytokine production during Orientia tsutsugamushi infection in mice. *Microbial pathogenesis* (2004).
20. Tantibhedhyangkul, W., Prachason, T., Waywa, D., El Filali, A., Ghigo, E., Thongnoppakhun, W., Raoult, D., Suputtamongkol, Y., Capo, C., Limwongse, C. & Mege, J. L. Orientia tsutsugamushi stimulates an original gene expression program in monocytes: relationship with gene expression in patients with scrub typhus. *PLoS Negl Trop Dis* **5**, e1028 (2011).
21. Tantibhedhyangkul, W., Ben Amara, A., Textoris, J., Gorvel, L., Ghigo, E., Capo, C. & Mege, J. L. Orientia tsutsugamushi, the causative agent of scrub typhus, induces an inflammatory program in human macrophages. *Microb Pathog* **55**, 55-63 (2013).
22. Astrup, E., Janardhanan, J., Otterdal, K., Ueland, T., Prakash, J. A., Lekva, T., Strand, O. A., Abraham, O. C., Thomas, K., Damas, J. K., Mathews, P., Mathai, D., Aukrust, P. & Varghese, G. M. Cytokine network in scrub typhus: high levels of interleukin-8 are associated with disease severity and mortality. *PLoS Negl Trop Dis* **8**, e2648 (2014).
23. Saliba, A. E., C Santos, S. & Vogel, J. New RNA-seq approaches for the study of bacterial pathogens. *Curr Opin Microbiol* **35**, 78-87 (2017).
24. Westermann, A. J., Förstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., Müller, L., Reinhardt, R., Stadler, P. F. & Vogel, J. Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature* (2016).
25. Westermann, A. J., Barquist, L. & Vogel, J. Resolving host-pathogen interactions by dual RNA-seq. *PLOS Pathog* **13**, (2017).
26. Humphrys, M. S., Creasy, T., Sun, Y., Shetty, A. C., Chibucos, M. C., Drabek, E. F., Fraser, C. M., Farooq, U., Sengamalay, N., Ott, S., Shou, H., Bavoil, P. M., Mahurkar, A. & Myers, G. S. Simultaneous transcriptional profiling of bacteria and their host cells. *PLoS One* **8**, e80597 (2013).
27. Enatsu, T., Urakami, H. & Tamura, A. Phylogenetic analysis of Orientia tsutsugamushi strains based on the sequence homologies of 56-kDa type-specific antigen genes. *FEMS Microbiol Lett* **180**, 163-169 (1999).

28. Paris, D. H., Aukkanit, N., Jenjaroen, K., Blacksell, S. D. & Day, N. P. A highly sensitive quantitative real-time PCR assay based on the groEL gene of contemporary Thai strains of Orientia tsutsugamushi. *Clin Microbiol Infect* **15**, 488-495 (2009).
29. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
30. Withey, J. H. & Friedman, D. I. A salvage pathway for protein structures: tmRNA and trans-translation. *Annu Rev Microbiol* **57**, 101-123 (2003).
31. Gaudin, C., Zhou, X., Williams, K. P. & Felden, B. Two-piece tmRNA in cyanobacteria and its structural analysis. *Nucleic Acids Res* **30**, 2018-2024 (2002).
32. Sharkady, S. M. & Williams, K. P. A third lineage with two-piece tmRNA. *Nucleic Acids Res* **32**, 4531-4538 (2004).
33. Storz, G., Vogel, J. & Wassarman, K. M. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* **43**, 880-891 (2011).
34. Wagner, E. G. H. & Romby, P. Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Adv Genet* **90**, 133-208 (2015).
35. Albrecht, M., Sharma, C. M., Dittrich, M. T., Müller, T., Reinhardt, R., Vogel, J. & Rudel, T. The transcriptional landscape of Chlamydia pneumoniae. *Genome Biol* **12**, R98 (2011).
36. Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K. & Sharma, C. M. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. *PLoS Genet* **9**, e1003495 (2013).
37. Kröger, C., Dillon, S. C., Cameron, A. D., Papenfort, K., Sivasankaran, S. K., Hokamp, K., Chao, Y., Sittka, A., Hébrard, M., Händler, K., Colgan, A., Leekitcharoenphon, P., Langridge, G. C., Lohan, A. J., Loftus, B., Lucchini, S., Ussery, D. W., Dorman, C. J., Thomson, N. R., Vogel, J. & Hinton, J. C. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc Natl Acad Sci U S A* **109**, E1277-86 (2012).
38. Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F. & Vogel, J. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**, 250-255 (2010).
39. Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tiensuu, T., Vaitkevicius, K., Barthelemy, M., Vergassola, M., Nahori, M. A., Soubigou, G., Régnault, B., Coppée, J. Y., Lecuit, M., Johansson, J. & Cossart, P. The Listeria transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950-956 (2009).
40. Vogel, J., Bartels, V., Tang, T. H., Churakov, G., Slagter-Jäger, J. G., Hüttenhofer, A. & Wagner, E. G. RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* **31**, 6435-6443 (2003).
41. Tjaden, B. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol* **16**, 1 (2015).
42. Georg, J. & Hess, W. R. Widespread Antisense Transcription in Prokaryotes. *Microbiol Spectr* **6**, (2018).
43. Wade, J. T. & Grainger, D. C. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Microbiol* **12**, 647-653 (2014).
44. Raghavan, R., Sloan, D. B. & Ochman, H. Antisense transcription is pervasive but rarely conserved in enteric bacteria. *MBio* **3**, (2012).
45. Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W. H., Bork, P., Glass, J. I., Serrano, L. & Lluch-Senar, M. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv* **2**, e1501363 (2016).

46. Andersson, S. G. & Sharp, P. M. Codon usage and base composition in Rickettsia prowazekii. *J Mol Evol* **42**, 525-536 (1996).
47. Botzman, M. & Margalit, H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol* **12**, R109 (2011).
48. Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* **33**, 1141-1153 (2005).
49. Duret, L. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**, 287-289 (2000).
50. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**, 389-409 (1981).
51. Ikemura, T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**, 573-597 (1982).
52. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**, 32-42 (2011).
53. Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M. & Pilpel, Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci U S A* **115**, E4940-E4949 (2018).
54. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475-479 (2013).
55. Ha, N. Y., Kim, Y., Choi, J. H., Choi, M. S., Kim, I. S., Kim, Y. S. & Cho, N. H. Detection of antibodies against Orientia tsutsugamushi Sca proteins in scrub typhus patients and genetic variation of sca genes of different strains. *Clin Vaccine Immunol* **19**, 1442-1451 (2012).
56. Ha, N. Y., Sharma, P., Kim, G., Kim, Y., Min, C. K., Choi, M. S., Kim, I. S. & Cho, N. H. Immunization with an Autotransporter Protein of Orientia tsutsugamushi Provides Protective Immunity against Scrub Typhus. *PLoS Negl Trop Dis* **9**, e0003585 (2015).
57. Ha, N. Y., Shin, H. M., Sharma, P., Cho, H. A., Min, C. K., Kim, H. I., Yen, N. T., Kang, J. S., Kim, I. S., Choi, M. S., Kim, Y. K. & Cho, N. H. Generation of protective immunity against Orientia tsutsugamushi infection by immunization with a zinc oxide nanoparticle combined with ScaA antigen. *J Nanobiotechnology* **14**, 76 (2016).
58. VieBrock, L., Evans, S. M., Beyer, A. R., Larson, C. L., Beare, P. A., Ge, H., Singh, S., Rodino, K. G., Heinzen, R. A., Richards, A. L. & Carlyon, J. A. 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 (2014).
59. Evans, S. M., Rodino, K. G., Adcox, H. E. & Carlyon, J. A. Orientia tsutsugamushi uses two Ank effectors to modulate NF-κB p65 nuclear transport and inhibit NF-κB transcriptional activation. *PLoS Pathog* **14**, e1007023 (2018).
60. Cho, K. A., Jun, Y. H., Suh, J. W., Kang, J. S., Choi, H. J. & Woo, S. Y. Orientia tsutsugamushi induced endothelial cell activation via the NOD1-IL-32 pathway. *Microb Pathog* **49**, 95-104 (2010).
61. Gharaibeh, M., Hagedorn, M., Lilla, S., Hauptmann, M., Heine, H., Fleischer, B. & Keller, C. Toll-Like Receptor 2 Recognizes Orientia tsutsugamushi and Increases Susceptibility to Murine Experimental Scrub Typhus. *Infect Immun* **84**, 3379-3387 (2016).

62. Kawai, T. & Akira, S. Signaling to NF-kappaB by Toll-like receptors. *Trends Mol Med* **13**, 460-469 (2007).
63. Shelite, T. R., Liang, Y., Wang, H., Mendell, N. L., Trent, B. J., Sun, J., Gong, B., Xu, G., Hu, H., Bouyer, D. H. & Soong, L. IL-33-Dependent Endothelial Activation Contributes to Apoptosis and Renal Injury in Orientia tsutsugamushi-Infected Mice. *PLoS Negl Trop Dis* **10**, e0004467 (2016).
64. McLeod, M. P., Qin, X., Karpathy, S. E., Gioia, J., Highlander, S. K., Fox, G. E., McNeill, T. Z., Jiang, H., Muzny, D., Jacob, L. S., Hawes, A. C., Sodergren, E., Gill, R., Hume, J., Morgan, M., Fan, G., Amin, A. G., Gibbs, R. A., Hong, C., Yu, X. J., Walker, D. H. & Weinstock, G. M. Complete genome sequence of Rickettsia typhi and comparison with sequences of other rickettsiae. *J Bacteriol* **186**, 5842-5855 (2004).
65. Yildirim, A. & Feig, M. High-resolution 3D models of Caulobacter crescentus chromosome reveal genome structural variability and organization. *Nucleic Acids Res* **46**, 3937-3952 (2018).
66. Ellis, M. J. & Haniford, D. B. Riboregulation of bacterial and archaeal transposition. *Wiley Interdiscip Rev RNA* **7**, 382-398 (2016).
67. Narra, H. P., Schroeder, C. L., Sahni, A., Rojas, M., Khanipov, K., Fofanov, Y. & Sahni, S. K. Small Regulatory RNAs of Rickettsia conorii. *Sci Rep* **6**, 36728 (2016).
68. Schroeder, C. L., Narra, H. P., Rojas, M., Sahni, A., Patel, J., Khanipov, K., Wood, T. G., Fofanov, Y. & Sahni, S. K. Bacterial small RNAs in the Genus Rickettsia. *BMC Genomics* **16**, 1075 (2015).
69. Schroeder, C. L., Narra, H. P., Sahni, A., Rojas, M., Khanipov, K., Patel, J., Shah, R., Fofanov, Y. & Sahni, S. K. Identification and Characterization of Novel Small RNAs in Rickettsia prowazekii. *Front Microbiol* **7**, 859 (2016).
70. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408 (2001).
71. Martin, M. Cutadapt Removes Adapter Sequences From High-throughput Sequencing Reads. *EMBnet.journal* **17**, 10-12 (2011).
72. Förstner, K. U., Vogel, J. & Sharma, C. M. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics* **30**, 3421-3423 (2014).
73. Otto, C., Stadler, P. F. & Hoffmann, S. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics* **30**, 1837-1843 (2014).
74. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. & Bork, P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**, D286-93 (2016).
75. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017).
76. McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., Vanderpool, C. K. & Tjaden, B. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res* **41**, e140 (2013).
77. Yu, S. H., Vogel, J. & Förstner, K. U. ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *Gigascience* **7**, (2018).
78. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

79. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D. & Petrov, A. I. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**, D335-D342 (2018).
80. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).
81. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009-1010 (2011).
82. Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., Hartmann, R. K., Prohaska, S. J. & Stadler, P. F. Orthology detection combining clustering and synteny for very large datasets. *PLoS One* **9**, e105015 (2014).
83. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
84. Chen, Y., Lun, A. T. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438 (2016).
85. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **8**, (2008).
86. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C. & Müller, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
87. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37**, D93-7 (2009).
88. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* **44**, D184-9 (2016).
89. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163-3168 (2015).

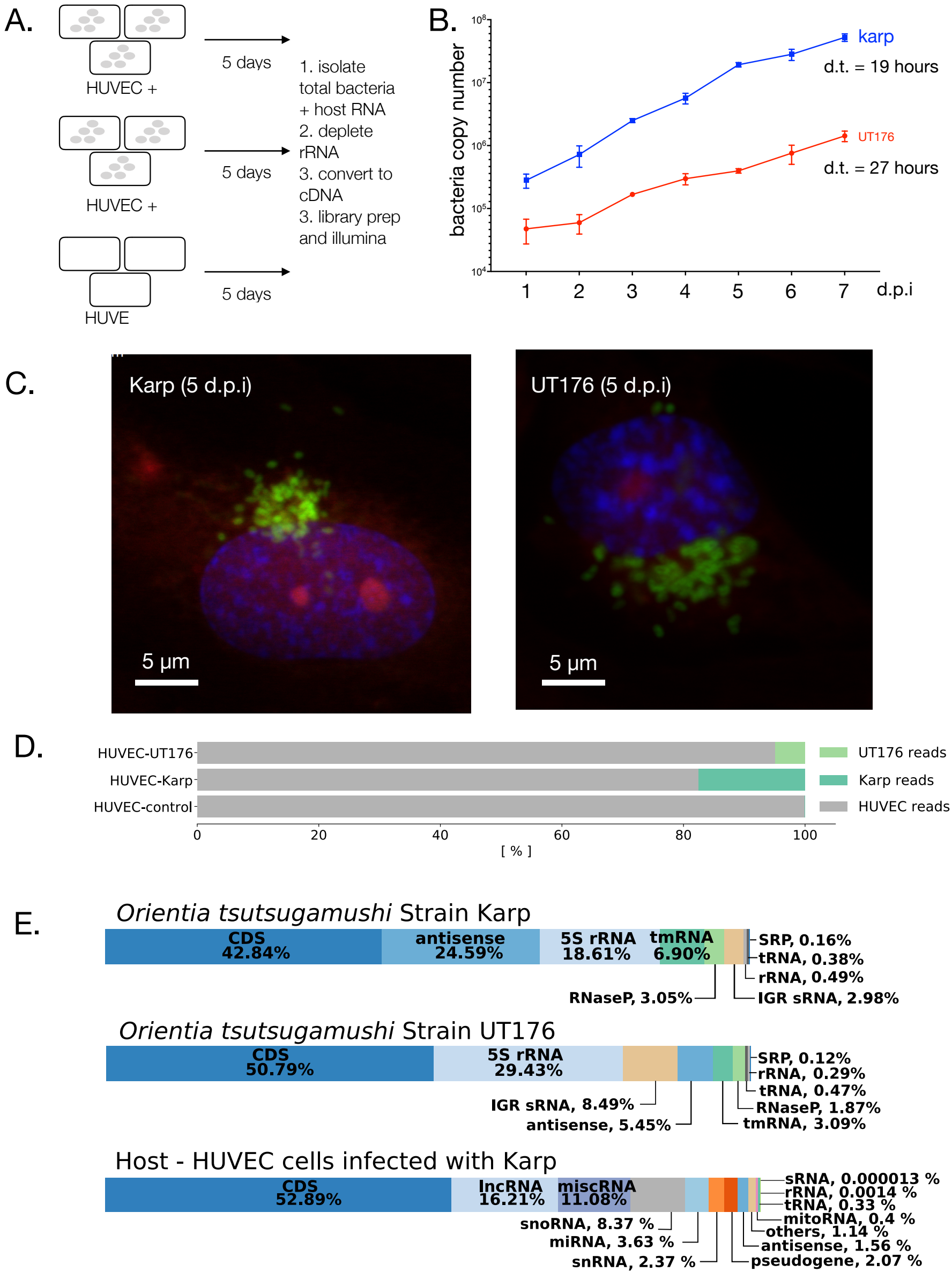


Figure 1. Overview. A. Schematic experimental overview. HUVEC = human umbilical vein endothelial cell. **B.** Growth curve showing replication of Ot in cultured HUVEC cells. Bacteria were grown in 24 well plates and the total bacteria per well is shown. Mean and SD from three independent replicates are shown. **C.** Confocal microscopy images of Ot in HUVEC cells 5 days post infection. Additional images and time points are shown in Supp. Fig. 1 and 2. Blue = DAPI (DNA), Red = Evans blue (host cells), green = Ot labelled with FITC-click-methionine. **D.** RNA mapping statistics showing the fraction of host and Ot RNA for each condition. The mean of three independent experiments is shown. Individual results for each replicate are shown in Supp. Fig. 3. **E.** Percentage of RNA-seq reads assigned to different classes of RNA in Karp, UT176 and HUVEC.

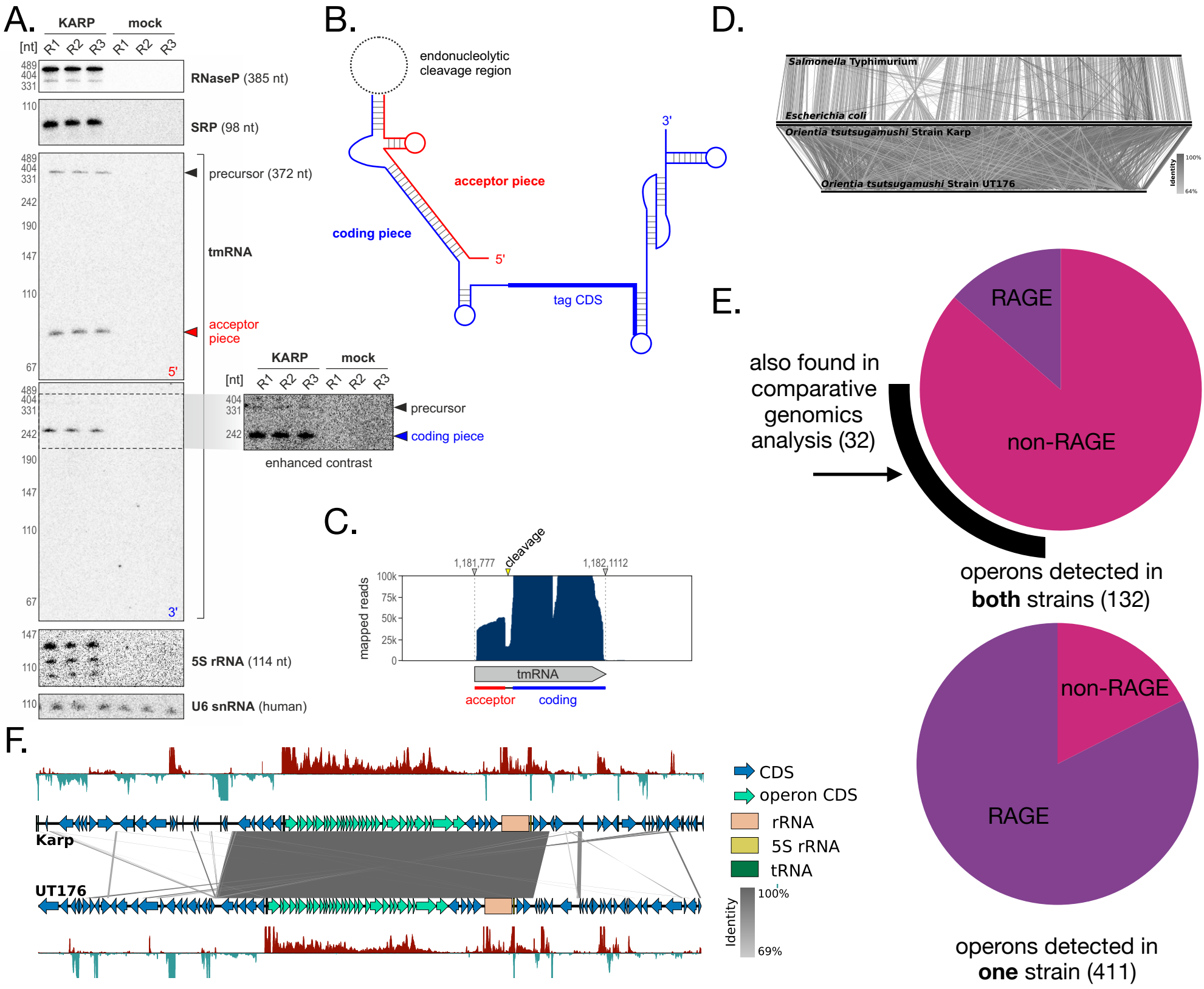


Figure 2. RNA biology in Ot. A. Northern blot analysis of core non-coding RNAs in Ot. **B.** Structure of the two-piece tmRNA observed in the Ot transcriptome **C.** RNA-seq read covereage over the tmRNA gene mirrors cleavage observed by Northern blot. **D.** A comparison of genomic synteny of two species within the enterobacteriaceae (*Escherichia coli* MG1655 and *Salmonella enterica* serovar Typhimurium SL1344, top), with synteny between the two Orientia strains from this study (bottom). **E.** Pie charts illustrating the relative abundance of RAGE genes in conserved (right) and strain-specific (left) operons. **F.** Visualization of the largest conserved operon in Ot, encoding multiple ribosomal genes, showing RNA-seq coverage in both strains.

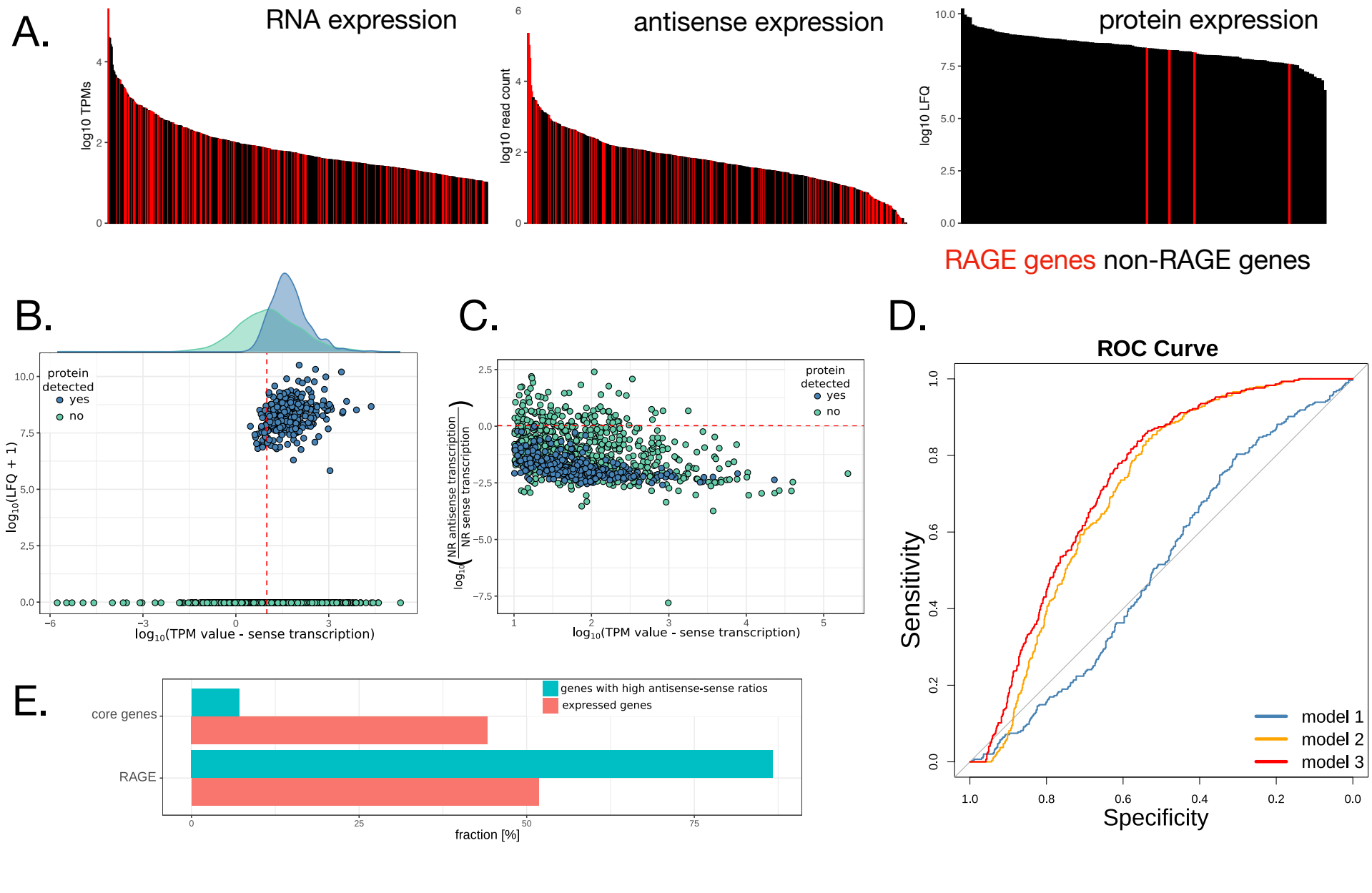
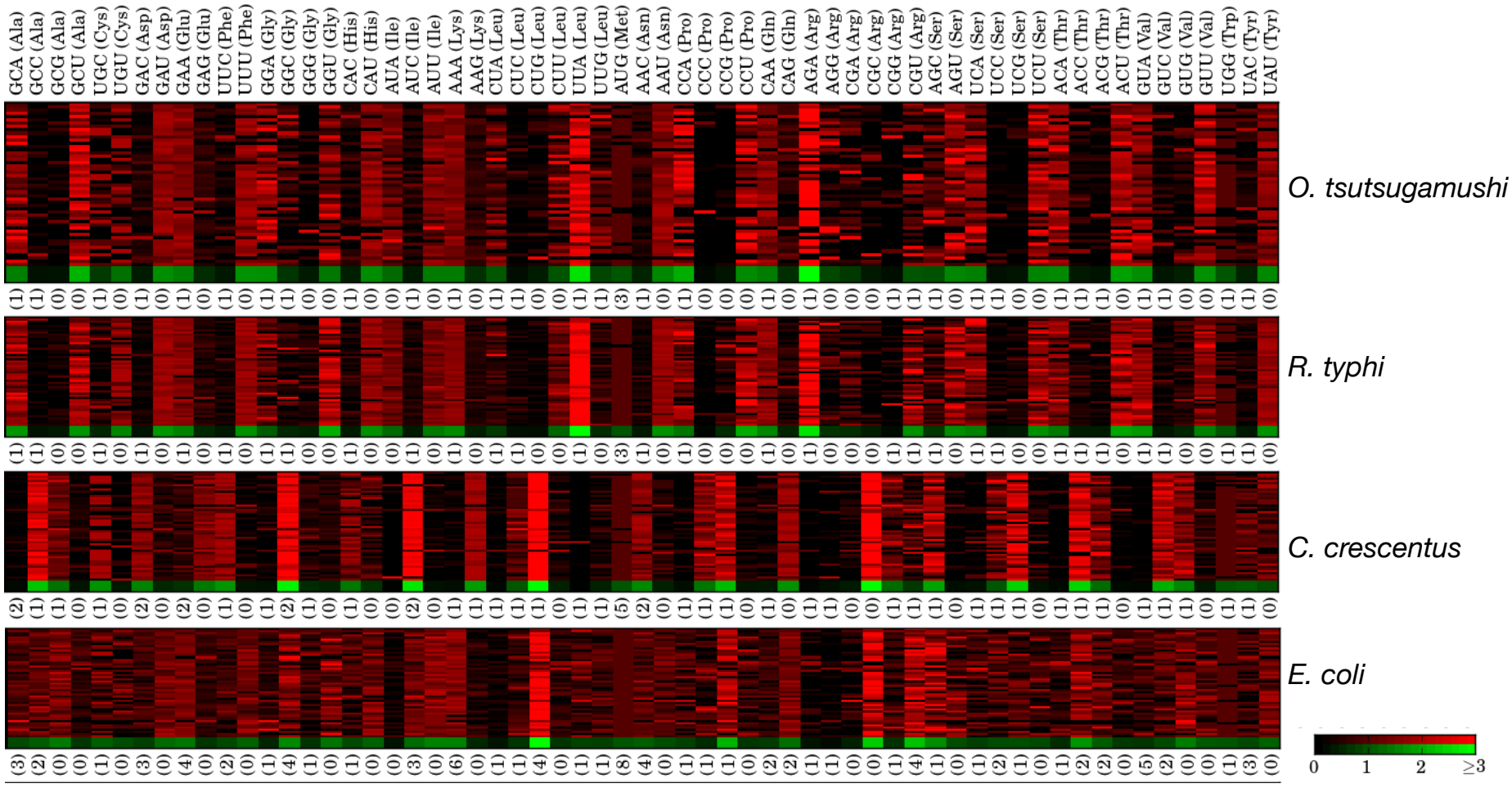
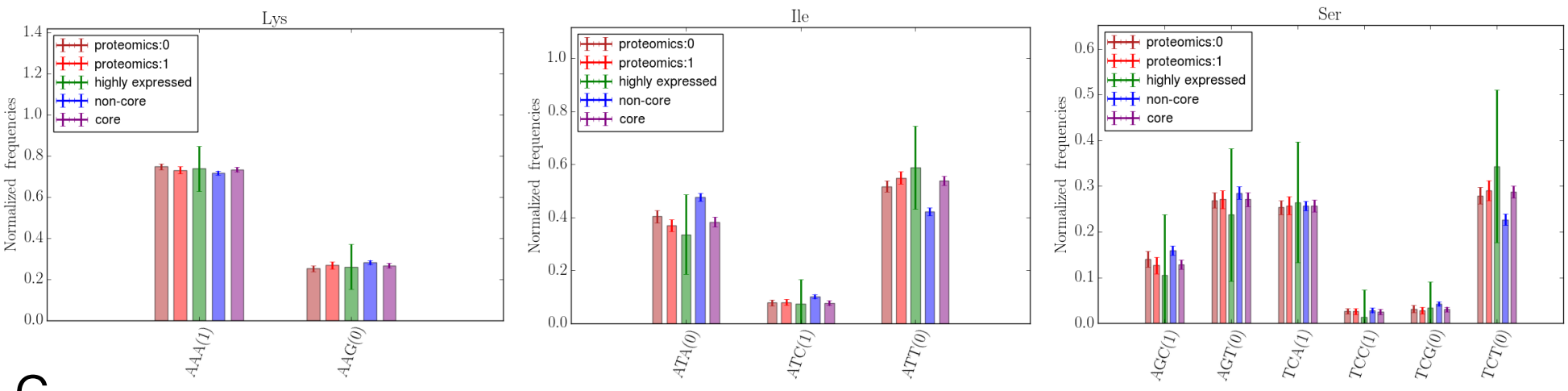


Figure 3. Antisense transcription is enriched on mobile genetic elements and is predictive of an absence of protein expression. A. B. Plot showing the relationship between protein expression, defined by LFQs, and transcript expression, defined by TPMs. Genes cluster into two groups based on their protein expression. The red line indicates the threshold for expressed genes (TPM value equal to 10). **C.** Sense transcription and the ratio of reads assigned to the antisense and sense strands, showing classification based on proteomics detection. The red line indicates the sense-antisense ratio (1.059182) above which translation was not detected by mass spectrometry. **D.** ROC curves evaluating the performance of logistic regression models to predict protein expression from RNA-seq read counts. Model 1 strictly uses sense expression, Model 2 the antisense-sense ration, and Model 3 uses both. Incorporating antisense expression clearly improves model performance. **E.** Fraction of core genes and RAGE genes in the set of genes with high antisense-sense ratios, compared to all expressed genes.

A.



B.



C.

amino acid	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
number of codons	4	2	2	2	2	4	2	3	2	6	1	2	4	2	6	6	4	4	1	2
bias between groups	Y	Y	N	N	N	Y	N	Y	N	N	-	N	Y	N	Y	Y	Y	Y	-	N
direction of genome-wide bias	S	N	N	P	N	S	N	N	P	S	-	N	S	P	S	S	S	S	-	N

Figure 4. Codon bias in Ot. A. Genome-wide codon biases in *O. tsutsugamushi, R. typhi, C. crescentus* and *E. coli.* Genome-wide RSCU values are indicated by the green boxes, while the red boxes are the values for 50 randomly chosen individual genes. tRNA gene copy numbers are given in parentheses. B. Relative use of different codons by subsets of genes in Ot (Karp) that are absent in proteomics, detected in proteomics, highly expressed in proteomics, encoded by non-core gene set or encoded by core gene set. Three illustrative examples are shown (Lys, Ile, Ser) with the full set of 20 amino sets given in Supp. Fig. 12. The error bars represent 99.7% confidence intervals and were generated by computing the normalized codon frequencies from 10000 sets of randomly selected genes, with each set being equal in size to the corresponding gene set. C. Overview of relative use of codons to encode each amino acid in the different proteomics gene sets. The bias between groups (yes in green, no in red) indicates whether the different groups use the different possible codons differently. This is defined by a consistent trend between the not detected, detected and highly expressed groups. As an example, Lys in Fig. 4B=no, whilst Ile and Ser=yes. The direction of genome-wide bias indicates whether the genome-wide relative use of the different codon correlates with the presence or absence of cognate tRNA, or not. (positive correlation in green, negative correlation in red, split correlation in yellow). As an example Lys=positive, Ile=negative, Ser=split.

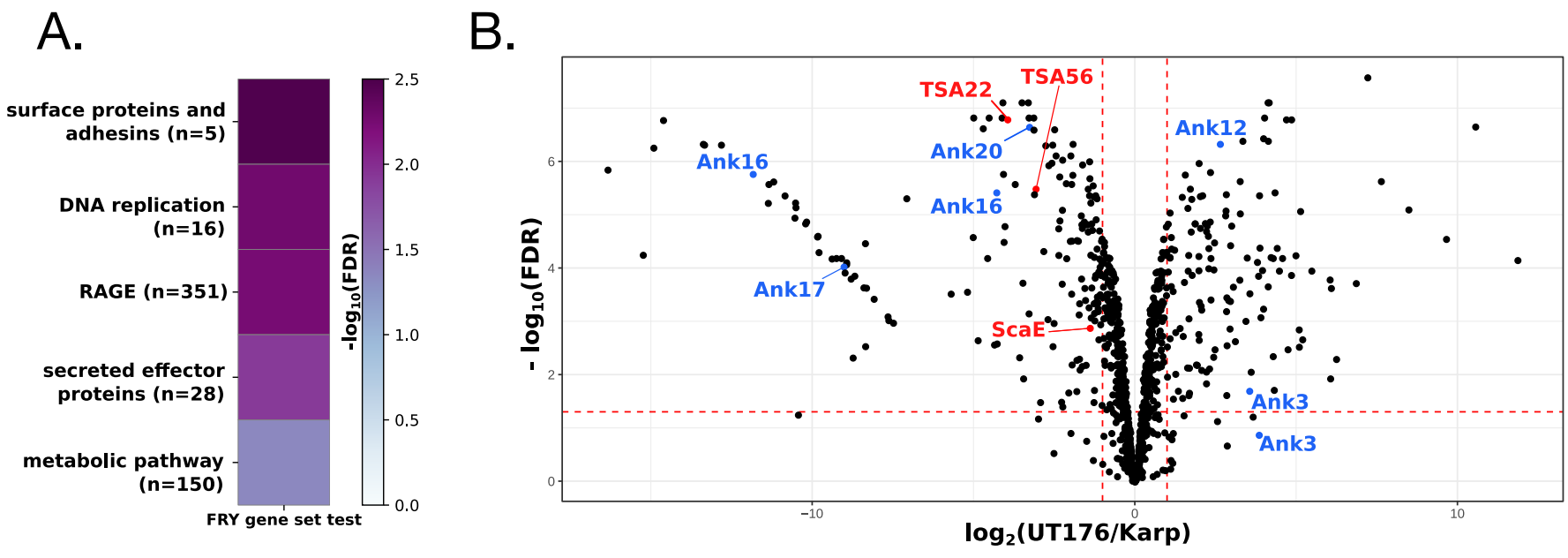


Figure 5. Differential bacterial expression. A. Heatmap illustrating pathways enriched in differentially expressed genes. All illustrated categories are more highly expressed in Karp. B. Volcano plot showing differential expression of bacterial genes in Karp and UT176. Bacterial surface genes (red) and ankyrin repeat containing effector proteins (blue) with log fold change ≥1 are highlighted.

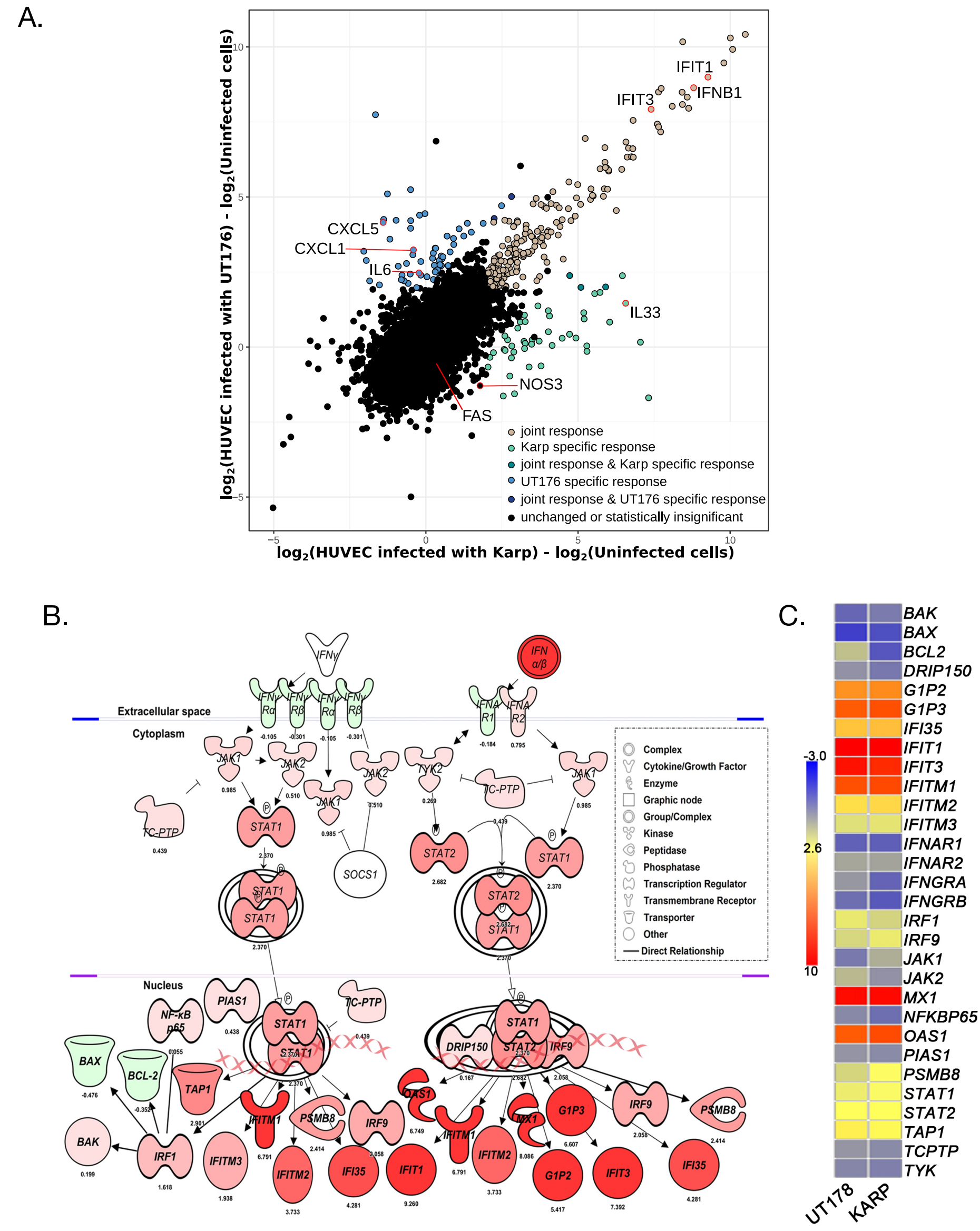


Figure 6. Ot induces an antiviral interferon response in HUVEC cells. A. Summary of the host response showing joint and strain specific responses. **B.** Activation of multiple genes in the canonical interferon signaling pathway in Karp infected HUVEC cells compared with uninfected HUVEC cells. **C.** Heat map showing up-regulation of genes in the interferon signaling pathway in HUVEC cells infected with Karp and UT176 compared with uninfected cells.

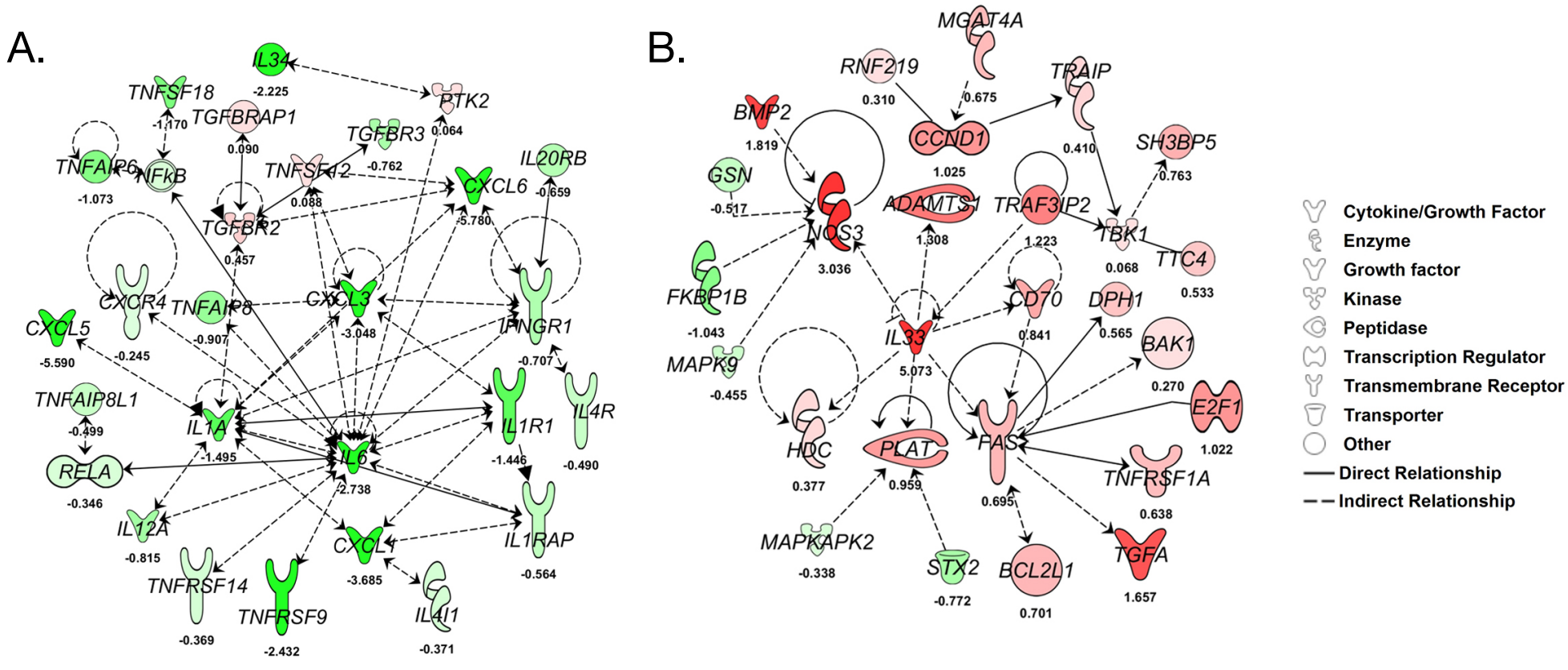


Figure 7. Karp and UT176 lead to up-regulation of distinct networks in HUVEC cells. A. Up-regulation of multiple proinflammatory chemokines and cytokines in HUVEC cells infected with UT176. **B.** Induction of the IL33-FAS-mediated anoikis network in Karp infected HUVEC cells.

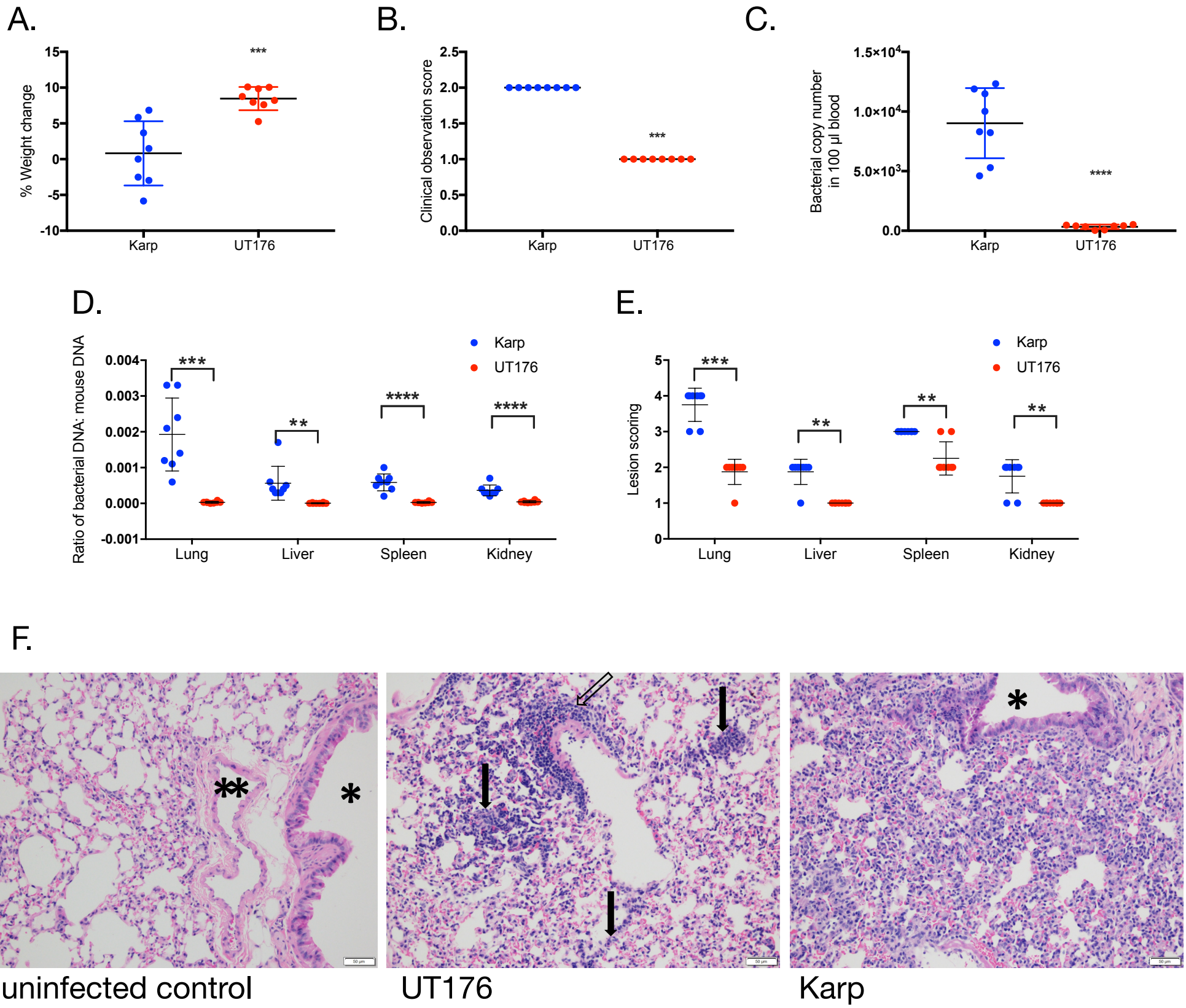


Figure 8. Karp is more virulent than UT176 in a mouse infection model. A. Weight change over 12 days infection. **B.** Clinical observation score of mice 12 days post infection. This number is a composite score based on appetite, activity, and hair coat with higher numbers representing low appetite, low activity, and ruffled fur. Details provided in Supp. Fig. 20. **C.** Bacterial copy number in 100 μl blood taken from euthanized mice 12 days post infection, measured by qPCR. **D.** Ratio of bacterial DNA copy number to mouse DNA copy number in lung, liver, spleen, and kidney of euthanized mice 12 days post infection, measured by qPCR. **E.** Lesion scores of Hematoxylin and Eosin stained lung, liver, spleen, and kidneys of euthanized mice 12 days post infection. Scores range from 0 to 5 with 0 representing normal tissue and 5 representing severe lesion damage. All graphs show mean and standard deviation. Statistical significance is calculated using unpaired t test using GraphPad Prism software. **$p \leq 0.01$ ***$p \leq 0.001$ ****$p \leq 0.0001$ F. Images of Hematoxylin and Eosin stained lung tissue of mice infected with buffer, UT176 or Karp. Scale bars = 50 μm. * indicates airway and ** indicates blood vessel. Uninfected control: airway, blood vessel and alveoli all appear normal. UT176-infected lungs: There is diffuse thickening and infiltration of alveolar septa with a mixed population of macrophages and lymphocytes (arrows). There is also mild perivascular lymphohistiocytic inflammation (open arrow). Karp-infected lungs: There is diffuse moderate thickening and infiltration of alveolar septa with a mixed population of macrophages and lymphocytes. The airway (*) is unaffected and normal. Additional figures are shown in Supp. Fig. 20.